

IA et énergie

Transformer une interdépendance
en capacité de décision



Image réalisée par une IA

Ana Semedo

RAPPORT · 2026

HUB
FRANCE
IA

G9+



IL Expansions

Table des matières

<i>Executive Summary</i>	5
IA et énergie : transformer une interdépendance en capacité de décision.....	5
Remerciements.....	8
Préambule - Comment lire ce rapport.....	9
Introduction générale - Mise en perspective stratégique.....	11
Chapitre 1 - Derrière l'IA, l'infrastructure.....	13
1.1 Le data center comme système à six dimensions en interaction.....	13
1.2 Six pièges pour lire les chiffres.....	15
1.3 Le paradoxe français.....	18
1.4 La grille analytique : substitution, relocalisation, addition, effet d'appel.....	19
Chapitre 2 - Le raccordement : test de réalité des projets IA.....	21
2.1 La file de raccordement : un symptôme, pas un problème technique.....	21
2.2 RTE et Enedis : deux opérateurs, deux univers.....	22
2.3 Les effets collatéraux : quand le raccordement d'un projet bloque d'autres transitions.....	22
2.4 La localisation : ce que le réseau révèle des territoires.....	23
2.5 Du « premier arrivé » au « premier prêt » : une réforme de doctrine.....	24
2.6 La flexibilité : levier central, partage des coûts non réglé.....	26
2.7 Vers une doctrine de raccordement ?.....	27
Chapitre 3 - Le territoire : là où l'attractivité devient négociation.....	29
3.1 Une demande hypercentralisée : logiques de concentration et boucles de rétroaction.....	29
3.2 L'asymétrie structurelle et la « gentrification de l'électricité ».....	30
3.3 Trois approches territoriales.....	32
3.4 Foncier, eau, chaleur : tests matériels d'implantation.....	35
3.5 Le vide réglementaire : aggravant structurel.....	39
3.6 Le coût de sortie : instruire l'après dès l'entrée.....	41
3.7 Ce que ce chapitre implique pour la doctrine.....	42
Chapitre 4 - Localiser ne suffit pas.....	44
4.1 La localisation sans maîtrise : le paradoxe du bâtiment supposé souverain.....	44
4.2 La chaîne de production : une chaîne de dépendances.....	46
4.3 Raccorder ici, capter ailleurs : l'économie territoriale du data center.....	49
4.4 Héberger sans savoir : l'opacité des clients, des usages et des données.....	50

4.5 Investir sans construire de capacité stratégique.....	51
4.6 Ne pas singer les hyperscalers : choisir les couches où l'Europe peut être critique.....	57
4.7 Ce que ce chapitre implique : de l'attractivité à la maîtrise.....	60
Interlude stratégique - Ne pas singer les <i>hyperscalers</i> : vers le data center gouvernable.....	62
IS.1 Outiller les territoires : passer du consentement à l'instruction.....	62
IS.2 Genesis : quand le calcul devient distribuable.....	63
IS.3 Distribuer le calcul, mais aussi la décision.....	63
IS.4 DCflex : tester le comportement système des data centers.....	64
IS.5 La souveraineté se mesure aussi dans le cycle de vie.....	64
IS.6 Faire le bon calcul au bon moment.....	65
IS.7 SOCloud : rendre l'utilisateur partie prenante de l'arbitrage.....	65
IS.8 Une géographie fonctionnelle du calcul IA.....	66
Chapitre 5 - L'IA frugale : consommer juste.....	70
5.1 Ce que la frugalité n'est pas.....	70
5.2 L'IA qui consomme juste.....	71
5.3 La distinction fondatrice : IA consommatrice versus IA utile.....	71
5.4 Les déclencheurs réels de la frugalité.....	73
5.5 Quatre profils d'organisations face à l'IA.....	75
5.6 Le triptyque de l'architecture frugale émergente : SLM + RAG + routeur.....	76
5.7 Les budgets tokens et les agents.....	79
5.8 Dix niveaux d'arbitrage de la frugalité.....	79
5.9 La robustesse comme condition : l'IA frugale dans les systèmes critiques.....	81
5.10 La tension structurelle : <i>business models</i> vs frugalité.....	81
5.11 La frugalité comme critère d'achat : ce qui ne se diffusera pas autrement.....	82
5.12 L'architecture agnostique.....	83
Chapitre 6 - Rendre l'inférence visible : condition de toute frugalité.....	85
6.1 L'IA cachée : le vecteur de demande invisible.....	85
6.2 La chaîne d'intermédiation : pourquoi l'opacité est structurelle.....	86
6.3 Ce que la mesure actuelle ne dit pas.....	87
6.4 <i>Location-based, market-based</i> : la comptabilité qui ne dit pas toute la réalité physique.....	88
6.5 Le <i>carbon-aware computing</i> : intérêt réel, limites françaises.....	89
6.6 Le Scope 3 de l'IA : la chaîne matérielle oubliée.....	90
6.7 Ce que la traçabilité d'exécution devrait couvrir.....	93

6.8 <i>AI for Green vs. Green AI</i> : une distinction nécessaire.....	94
6.9 <i>AI for Brown</i> : ce que la mesure doit aussi couvrir.....	95
6.10 L'inférence gouvernable : voir, décider, contractualiser.....	97
Chapitre 7 - Industrialiser la frugalité : ce qui ne se fera pas par bonne volonté.....	99
7.1 Les trois niveaux de la frugalité : usage, architecture, infrastructure.....	99
7.2 Les leviers logiciels et architecturaux.....	99
7.3 Les leviers infrastructurels.....	101
7.4 La contrainte structurelle du « <i>build on top</i> » : le développeur comme assembleur.....	103
7.5 La tension structurelle : généralité versus spécificité.....	103
7.6 Les quatre formes d'effets rebond.....	104
7.7 La frugalité institutionnalisée : mécanismes organisationnels.....	105
7.8 La frugalité comme avantage concurrentiel émergent.....	106
Chapitre 8 - Quand l'IA devient utile : intégration métier, robustesse et résistances.....	108
8.1 L'IA utile est souvent une IA qui disparaît dans l'outillage métier.....	108
8.2 Réseaux électriques : optimiser sous contraintes fortes.....	108
8.3 Bâtiments, chauffage, climatisation, recharge, <i>microgrids</i> : l'IA face aux usages réels.....	109
8.4 IA générative : utile, mais pas toujours centrale dans l'énergie.....	109
8.5 L'intégration industrielle : un verrou plus critique que le modèle seul.....	110
8.6 Ce qui résiste : données, confiance, responsabilité, temps organisationnel.....	112
8.7 Conclusion - L'IA industrielle demande une double robustesse.....	114
Chapitre 9 - Là où l'IA produit déjà des gains mesurables.....	115
9.1 La méthode : partir des usages, pas des promesses.....	115
9.2 Des usages inégalement matures.....	115
9.3 L'optimisation énergétique du bâtiment.....	119
9.4 Les freins réels à l'adoption.....	120
9.5 L'IA dans la maîtrise thermique des bâtiments : une illustration ciblée.....	121
9.6 Ce que ce panorama dit de la Partie III.....	122
Chapitre 10 - Passer de la promesse à la preuve.....	123
10.1 Le passage du POC au déploiement : la première rupture.....	123
10.2 La qualité de la donnée : condition préalable à toute preuve.....	124
10.3 La mesure du gain net : le test de vérité.....	124
10.4 L'explicabilité : une condition de maîtrise, pas seulement d'ergonomie.....	125
10.5 L'acceptabilité des métiers cœur : la condition la plus sous-estimée.....	126

10.6 Le verrou du jugement métier : l'effet sur les compétences critiques.....	127
10.7 La robustesse dans le temps : ce que l'IA doit tenir.....	127
10.8 Frugalité par conception : mesurer le coût de l'IA pour crédibiliser le gain.....	128
10.9 Le maintien des collectifs humains : la condition la plus négligée.....	128
10.10 Synthèse : sept conditions, un système.....	129
Chapitre 11 - Les angles morts qui peuvent transformer l'avantage français en passif.....	132
11.1 Premier angle mort : la gouvernance des données hébergées.....	132
11.2 Deuxième angle mort : le cycle de vie et la fin d'exploitation.....	133
11.3 Troisième angle mort : l'impact sur les prix du foncier et l'effet d'éviction.....	134
11.4 Quatrième angle mort : la mesure de l'empreinte réelle des data centers.....	135
11.5 Cinquième angle mort : la souveraineté par couches ou ce qu'il devient critique de ne pas maîtriser.....	137
11.6 Sixième angle mort : l'expertise publique indépendante.....	144
11.7 Septième angle mort transversal : l'absence d'apprentissage collectif.....	144
11.8 Ce que ces angles morts ont en commun.....	145
Chapitre 12 - Accueillir sans subir : instrumenter la décision IA-énergie.....	147
Sept recommandations pour mettre en œuvre la doctrine.....	148
Conclusion générale.....	152
Glossaire.....	155

Executive Summary

IA et énergie : transformer une interdépendance en capacité de décision

L'intelligence artificielle (IA) et l'énergie ne peuvent plus être pensées séparément. Ce rapport part de leur interdépendance : l'IA mobilise de l'électricité, du foncier, de l'eau, des équipements, des réseaux, des compétences et des capacités industrielles ; l'énergie utilise à son tour l'IA pour prévoir, optimiser, maintenir et piloter des systèmes de plus en plus complexes. Le véritable enjeu est donc de savoir quelles IA nous voulons rendre possibles, avec quelles ressources, sous quelles dépendances et pour quelle valeur collective.

1. Transformer l'avantage français en capacité de choix

La France dispose d'atouts majeurs : mix électrique largement décarboné, réseau robuste et écosystème d'ingénierie des systèmes complexes au cœur des flux européens. Ces atouts peuvent soutenir une trajectoire différente de la simple course mondiale à la puissance de calcul. Ils peuvent aussi transformer la France en territoire d'accueil d'une demande numérique décidée ailleurs, mobilisant ici électricité, foncier et infrastructures pour une valeur principalement captée ailleurs.

Les data centers rendent ce paradoxe tangible : l'IA y devient puissance électrique, eau, chaleur, câbles, serveurs, GPU¹, équipements de secours et dépendances industrielles. Une électricité bas-carbone constitue un avantage réel, mais elle peut aussi produire son propre effet d'appel, alors même que les mêmes ressources sont attendues par l'industrie, la mobilité, le chauffage, les services publics et la transition écologique. L'enjeu n'est donc pas de maximiser les capacités accueillies, mais de déterminer quels usages justifient quelles ressources rares et quelles capacités stratégiques ils contribuent à construire.

Cette lecture oblige à dépasser les milliards d'euros annoncés, les gigawatts demandés ou les puissances contractualisées. Ces indicateurs ne disent ni ce qui sera effectivement construit, ni ce qui sera réellement consommé, ni ce qui tournera dans l'infrastructure. Le rapport distingue ainsi substitution, relocalisation, addition et effet d'appel : quatre situations qui peuvent afficher une puissance comparable mais ne produisent ni le même bilan environnemental, ni la même valeur stratégique, ni la même légitimité collective.

2. Faire du raccordement un instrument d'instruction

La file de raccordement ne mesure pas seulement une consommation électrique future : elle révèle une compétition pour sécuriser une option sur une capacité collective rare. Projet industriel engagé, couverture face à l'incertitude, positionnement foncier et réservation stratégique peuvent aujourd'hui s'y côtoyer. Dimensionner trop tôt le réseau, le foncier ou les capacités industrielles pour des projets insuffisamment matures peut alors transformer une incertitude privée en coût collectif.

L'instruction doit donc porter conjointement sur la maturité réelle des projets, la nature des usages, la pertinence de la localisation, ainsi que les contributions et contreparties contractualisées. Pour un opérateur mondial, un site est une option dans un portefeuille ; pour une collectivité, il engage pendant plusieurs décennies du foncier, des réseaux, de l'eau, des infrastructures et des choix d'aménagement. L'accueil d'un data center devient ainsi une négociation sur des ressources collectives et doit être traité comme un contrat stratégique : puissance contre maturité, raccordement

¹ Voir glossaire pour GPU, latence, robustesse, inférence, fine-tuning, edge...

contre transparence, foncier contre valeur territoriale, investissement contre réversibilité, capacité de calcul contre usages qualifiés.

3. Faire de la frugalité une convergence organisée de contraintes

La frugalité n'entre pas nécessairement dans les organisations au nom de l'environnement. Elle se déclenche souvent sous l'effet de contraintes déjà actives : maîtriser les coûts, réduire la latence, protéger les données, préserver la robustesse, sécuriser les usages, limiter la dépendance fournisseur ou composer avec la disponibilité du calcul. Ces contraintes conduisent à des choix plus proportionnés qui peuvent, en retour, réduire les coûts, les dépendances ou les délais et améliorer maîtrise, résilience ou performance.

L'enjeu est donc de rendre cette convergence explicite et de l'organiser. La frugalité devient une discipline systémique d'action : qualifier l'usage, proportionner le calcul et l'architecture, mesurer le coût complet et le bénéfice net, borner les chaînes d'action, contractualiser les exigences fournisseurs et réviser les choix dans le temps. Elle ne signifie pas systématiquement moins d'IA, mais une IA proportionnée à ce qu'elle rend possible, à ce qu'elle mobilise et à ce qu'elle transforme.

4. Instruire la souveraineté comme une capacité maintenue

La localisation d'un bâtiment, la nationalité d'un fournisseur ou l'hébergement des données ne suffisent pas à établir la souveraineté. Celle-ci se joue dans une chaîne de dépendances : capitaux, opérateurs, clients, modèles, données, puces, câbles, cloud, logiciels, plateformes, interfaces, compétences, contrats et conditions d'accès. Elle se mesure aux droits de décision, aux dépendances critiques et aux capacités de substitution et de sortie conservées sur cette chaîne.

La question est alors : souverain pour quel usage, face à quel risque et à quelle échelle ? Une dépendance acceptable pour un usage ordinaire peut devenir critique pour une infrastructure essentielle. La souveraineté n'est donc pas un état acquis mais une capacité à maintenir dans le temps : auditer, décider, substituer, reconstruire ou sortir lorsque les conditions changent.

Cette lecture ouvre une voie industrielle européenne distincte de l'imitation des hyperscalers. La France et l'Europe peuvent chercher des positions clés là où elles disposent d'avantages distinctifs ou de capacités à renforcer : intégration de l'IA dans les systèmes énergétiques et industriels, logiciels opérationnels, robotique et IA physique, systèmes distribués, flexibilité réseau, architectures frugales, cloud de confiance, orchestration. Ces positions passent aussi par certaines briques de la chaîne matérielle : des semi-conducteurs et composants embarqués au packaging, aux capteurs, à l'électronique de puissance ou aux infrastructures de connectivité. Il ne s'agit pas nécessairement de maîtriser toute la chaîne, mais d'occuper des positions critiques, difficiles à substituer, qui donnent une capacité de négociation et d'action sur les couches déterminantes de l'écosystème.

5. Faire de l'énergie un test de vérité pour l'utilité de l'IA

L'IA appliquée à l'énergie montre ce que signifie réellement une IA utile. Les usages les plus utiles sont souvent peu spectaculaires : prévision, optimisation, détection d'anomalies, contrôle qualité, assistance métier ou pilotage de systèmes physiques. Leur valeur ne vient pas de la taille du modèle mais de leur intégration dans des données fiables, des règles métier, des équipements, des procédures de sûreté et une supervision humaine.

Ces usages imposent aussi un standard de preuve plus exigeant : bénéfice net, robustesse dans le temps, explicabilité, capacité de reprise et maintien du jugement métier. L'énergie devient ainsi un terrain de vérité pour l'IA : la performance statistique d'un modèle ne suffit plus lorsqu'il doit fonctionner dans un système physique, critique et placé sous responsabilité humaine.

6. Passer des promesses à une capacité collective de décision

Les principaux angles morts identifiés, données hébergées, cycle de vie, coûts d'opportunité fonciers, empreinte réelle, souveraineté par couches, expertise publique indépendante et apprentissage collectif, ont un point commun : aucun acteur ni outil sectoriel ne peut les traiter séparément. Un data center est simultanément un actif immobilier, une infrastructure électrique et numérique, un investissement financier, un objet territorial, environnemental et souverain. La réponse doit donc organiser les interactions entre ces dimensions plutôt qu'ajouter des procédures indépendantes.

Gouverner la relation IA-énergie suppose dès lors une doctrine cohérente : définir et qualifier avant d'autoriser ; conditionner l'accès aux ressources rares ; contractualiser les contreparties ; renforcer l'ingénierie publique ; rendre les dépendances et les impacts visibles ; organiser l'apprentissage collectif ; conserver la capacité de réviser ou d'arrêter. Il s'agit de faire passer l'IA et ses infrastructures du registre de la promesse à celui de la preuve, puis de la preuve à celui de la décision gouvernable.

La France peut choisir l'IA qu'elle veut rendre possible, les usages qu'elle souhaite servir, les infrastructures qu'elle accepte de mobiliser et les capacités qu'elle veut conserver ou construire. Frugalité, souveraineté par couches, intégration des systèmes complexes, ingénierie territoriale, preuve du bénéfice net et apprentissage collectif deviennent alors les composantes d'une même capacité stratégique de décision.

Remerciements

Ce rapport prolonge les travaux engagés en 2024 avec la publication de Green AI & AI for Green². Fruit d'une coopération entre le Hub France IA, l'Institut G9+ et ILExpansions, il a été conçu et rédigé par Ana Semedo, cofondatrice d'ILExpansions.

Il s'appuie sur plus de cinquante entretiens menés avec des professionnels issus d'horizons variés : énergie, numérique, industrie, infrastructures, data centers, recherche, conseil, régulation, territoires, ingénierie, architecture, innovation et numérique responsable.

Nous remercions vivement l'ensemble des personnes qui ont accepté de consacrer du temps à ces échanges, de partager leurs analyses, leurs retours d'expérience, leurs interrogations et parfois leurs alertes. Leur contribution a été déterminante pour saisir la complexité réelle des liens entre intelligence artificielle, énergie, infrastructures, territoires, souveraineté et frugalité.

Nos remerciements s'adressent également aux organisations dont sont issus plusieurs interlocuteurs, parmi lesquelles : ADEME, AFNUM, Agence internationale de l'énergie, ARCEP, Bpifrance, Bull, CEA, CEME, Cigref, Deloitte, Docaposte, EDF, Enedis, Engie, Enia, France Urbaine, Gimelec, Hager, Hub France IA, Institut G9+, Institut Paris Région, John Cockerill, Mines Paris - PSL, Mines Nancy, Ministère des Armées, Orange, OVHcloud, Schneider Electric, SERCE, Studio Degel et Vinci Group.

Nous remercions également le Cigref, pour plusieurs mises en relation ayant contribué aux entretiens conduits dans le cadre de ce travail.

Ces organisations sont citées pour situer la diversité des environnements professionnels représentés dans les entretiens. Sur un sujet aussi vaste et transversal, les échanges recueillis apportent des éclairages d'experts sans prétendre résumer, à eux seuls, la position de chacune d'elles.

Les personnes interviewées se sont exprimées en leur nom. Leurs contributions ont nourri la profondeur des constats, la diversité des angles d'analyse et la formulation des propositions portées dans ce rapport.

² <https://www.g9plus.org/fr/publications/ai-for-green-green-for-ai> et <https://www.g9plus.org/fr/publications/ai-for-green-green-for-ai---complementent-rse-ia>

Préambule - Comment lire ce rapport

Méthode et convention de lecture

Ce rapport doit être lu comme une enquête stratégique. Il s'appuie sur des entretiens d'experts complétés par l'analyse de publications récentes, notamment celles de l'ADEME, de l'AIE, de RTE, du Shift Project, de France Urbaine et de Ville de Demain³. Le guide territorial France Urbaine / Ville de Demain a nourri plusieurs critères d'instruction mobilisés dans ce rapport.

Beaucoup de personnes interviewées ont demandé que leurs apports soient anonymisés. Cet engagement a été pris afin de préserver la liberté des échanges et la qualité des analyses recueillies. En conséquence, les verbatims sont anonymisés sans distinction. Ils sont mobilisés pour éclairer des tendances, formuler des tensions ou donner à voir des raisonnements d'acteurs, sans être rattachés publiquement à une personne ou à une organisation.

Les données chiffrées issues de sources publiques sont référencées à leur source. Les ordres de grandeur recueillis lors des entretiens sont signalés comme tels et ne sont pas assimilés à des données statistiques établies. Les signaux émergents, non stabilisés ou encore peu documentés sont présentés comme tels. Cette distinction est importante, car les relations entre IA et énergie forment un objet en construction, où se croisent annonces d'investissement, projections de consommation, intérêts industriels, données partielles, contraintes territoriales et arbitrages publics. Ce rapport ne prétend donc pas à l'exhaustivité : il vise à rendre visibles des interdépendances, à qualifier des tensions et à outiller la décision.

Des outils d'intelligence artificielle ont été utilisés au cours du travail pour l'assistance à des tâches de recherche, de confrontation, de structuration ou de reformulation. L'architecture du rapport, les analyses, la sélection et la vérification des sources, les arbitrages, les interprétations et la rédaction finale relèvent de l'autrice. Le rapport a également bénéficié de relectures expertes couvrant ses différents champs : IA, énergie et réseaux électriques, data centers et infrastructures numériques, souveraineté, frugalité, enjeux environnementaux, territoires, usages et organisations.

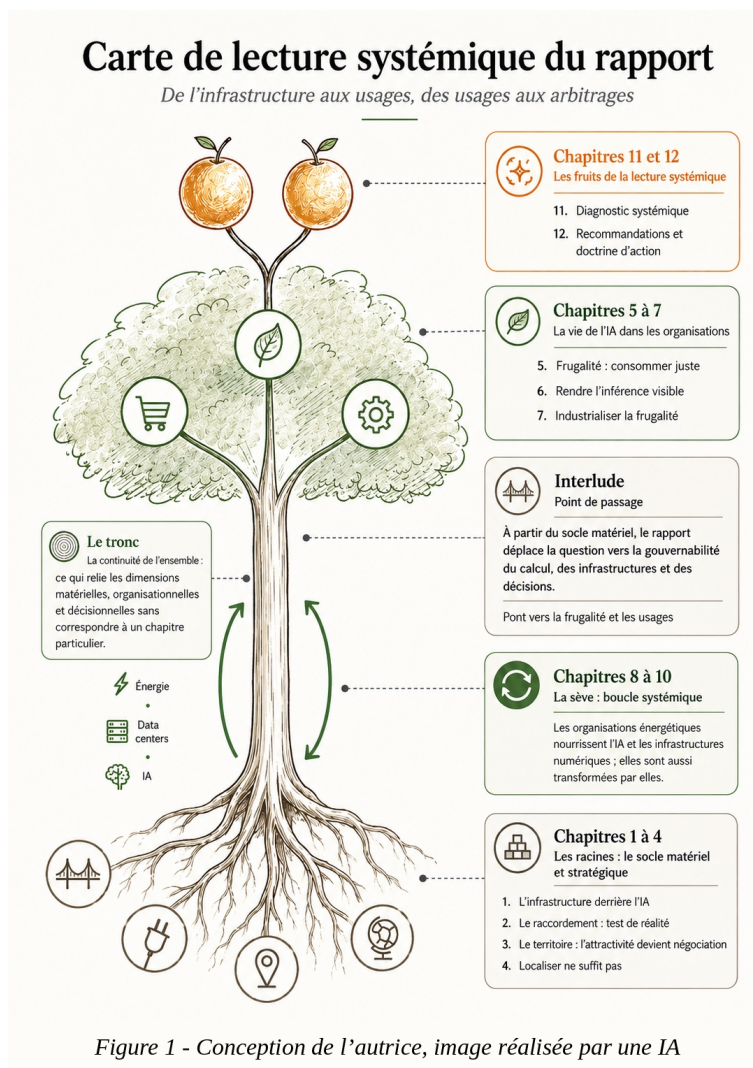
Une interdépendance à gouverner

Ce rapport part d'un constat : l'énergie et l'IA sont deux systèmes qui se nourrissent l'un l'autre en permanence. Le data center est l'un des lieux les plus visibles, où cette réciprocité prend une forme physique. La question du rapport va plus loin : où l'IA consomme-t-elle de l'énergie, où permet-elle d'en économiser et comment piloter cet équilibre plutôt que le subir ?

Une lecture systémique

La logique systémique du rapport concerne l'ensemble de la relation IA-énergie. Imaginez un arbre. Ses racines - les chapitres 1 à 4 - plongent dans le réel matériel des data centers : infrastructure physique, raccordement électrique, négociation territoriale et souveraineté.

³ Afin de ne pas alourdir le préambule, les références détaillées des sources mobilisées dans ces pages sont indiquées dans le corps du rapport lors de leur première utilisation sur le fond.



Le tronc représente la continuité de l'ensemble : ce qui relie les dimensions matérielles, organisationnelles et décisionnelles sans correspondre à un chapitre particulier.

L'interlude marque un point de passage : à partir de ce socle matériel, le rapport déplace la question vers celle de la gouvernabilité du calcul, des infrastructures et des décisions.

La sève circule ensuite dans les deux sens : les chapitres 8 à 10 montrent comment les organisations énergétiques nourrissent l'intelligence artificielle et les infrastructures numériques, tout en étant elles-mêmes transformées par elles. Cette circulation irrigue aussi la canopée - les chapitres 5 à 7 -, qui correspond à la vie de l'IA dans les organisations : frugalité, inférence visible, achats, architectures, usages et gouvernance.

Les deux derniers chapitres sont les fruits de cette lecture systémique. Le chapitre 11 rassemble les angles morts

du diagnostic. Le chapitre 12 propose une doctrine d'action : instruire les data centers comme des projets stratégiques multi-acteurs et multi-échelles, qualifier les usages, contractualiser les contreparties et construire une capacité collective d'action et de décision.

Pourquoi cette lecture compte

La question n'est pas de savoir si la France doit développer l'intelligence artificielle ou transformer son système énergétique. Ces deux mouvements sont engagés. Il s'agit plutôt de savoir si le pays et les organisations choisissent de piloter leur relation ou de la subir en traitant séparément l'énergie, l'IA, les infrastructures, les usages et les territoires.

Ce rapport ne prétend pas trancher à la place du lecteur. Il donne des instruments pour comprendre, réfléchir et instruire des décisions en connaissance de cause : une grille de lecture systémique, un vocabulaire adapté, des critères d'instruction et des propositions de doctrine d'accueil et d'usage.

L'enjeu est d'accueillir sans subir, dimensionner sans renoncer, récolter sans l'illusion de promesses.

Introduction générale - Mise en perspective stratégique

Le rapport pose une question simple, mais qui engage beaucoup : avons-nous vocation à devenir un simple point de passage de l'intelligence artificielle mondiale, un territoire qui héberge et consomme ce que d'autres pilotent, ou pouvons-nous construire une stratégie qui nous est propre et s'appuie sur nos atouts ?

Cette question traverse l'ensemble du rapport. Elle suppose de ne pas traiter séparément les infrastructures, les usages, l'énergie, la souveraineté, la frugalité et les organisations. Elle suppose aussi de se demander quelle IA nous voulons rendre possible : une IA utile, proportionnée, maîtrisable, tournée vers l'industrie, la transition écologique et les usages à forte valeur collective ; une IA qui ne soit pas seulement tirée par le divertissement génératif de masse, les copilotes généralistes ou les usages conversationnels de confort.

Le data center est le lieu où cette question devient concrète et matérielle : puissance électrique, foncier, eau, chaleur, câbles, métaux, serveurs, GPU, raccordements, compétences, contrats et dépendances industrielles. Il doit être lu comme une chaîne d'arbitrages : où, pour quels usages, avec quelles conditions, au bénéfice de qui, et avec quelles ressources mobilisées alors qu'elles sont aussi attendues par l'industrie, la mobilité, le logement, la chaleur, les services publics et la transition écologique. Il est aussi le lieu où se rencontrent stratégies publiques et privées, regards concordants ou contrastés sur le devenir de l'IA.

La frugalité rend cette question opérationnelle avant même que la puissance soit dimensionnée : qualifier un usage avant de choisir une architecture, préférer lorsque c'est pertinent des modèles plus petits, spécialisés, ouverts ou maîtrisés localement, rendre l'inférence visible plutôt que de la laisser se diffuser sans décision. Elle n'est donc pas seulement une contrainte environnementale, mais une discipline d'arbitrage entre usages, coûts, souveraineté, compétences, robustesse, dépendances et impacts.

Gouverner la boucle entre énergie et IA ne suffit pas : il faut aussi gouverner ce qui la rend possible, couche par couche. C'est dans ce sens que le rapport propose de lire l'accueil des data centers comme un **contrat stratégique**. Bien conçu, il peut permettre d'associer les ressources collectives mobilisées à des preuves de maturité, de transparence, de valeur territoriale, de sobriété, de réversibilité et de maîtrise. À défaut, l'avantage français reste exposé à un risque central : une consommation ici, pour une valeur captée ailleurs.

PARTIE I

L'IA cliente de l'énergie : anatomie d'une demande

Chapitre 1 - Derrière l'IA, l'infrastructure

Le chapitre en une phrase

Un data center ne s'instruit pas comme un bâtiment mais comme une chaîne d'arbitrages : puissance, foncier, eau, chaleur, câbles, compétences. Le chapitre introduit une grille analytique en quatre situations - substitution, relocalisation, addition, effet d'appel - pour dépasser la seule question du bilan carbone et interroger ce que chaque projet fait réellement au système énergétique et territorial.

À travers les data centers, notre représentation de l'IA devient aussi puissance électrique, foncier, eau, chaleur, câbles, métaux, raccordements, compétences et dépendances industrielles.

Concrètement, un data center est un bâtiment industriel qui héberge des serveurs informatiques (des ordinateurs spécialisés, alignés par centaines ou milliers), qui stockent des données et exécutent des calculs. Autour d'eux, tout un système les fait fonctionner : alimentation électrique, refroidissement, câblage réseau, dispositifs de sécurité... C'est là que sont hébergés, stockés ou exécutés des données, des sites web, des messageries, des services cloud, des calculs, des applications métiers, des modèles d'IA.

Depuis le AI Action Summit de février 2025, où 109 milliards d'euros d'investissements avaient été annoncés en France, le rythme des annonces ne s'est pas ralenti : Choose France 2026 a annoncé 93 milliards d'euros de promesses d'investissement et RTE avait réservé en mai 2026⁴ près de 18 GW de capacité pour environ 80 projets, contre environ 5 GW fin 2024.

Les annonces mêlent souvent milliards d'euros investis et gigawatts à raccorder, comme si ces deux indicateurs suffisaient à qualifier un projet. Cette simplification peut conduire à de mauvaises décisions, car un data center ne vaut pas que par sa puissance affichée, mais par ce qui y tourne, pour qui, à quelle intensité, avec quelle maturité industrielle, quelles contreparties territoriales et quelle valeur réellement partagée. Un data center ne s'instruit pas comme un bâtiment, mais comme une **chaîne d'arbitrages**.

1.1 Le data center comme système à six dimensions en interaction

Un data center est plutôt un nœud de convergence entre plusieurs dimensions qui interagissent en permanence. Nous en retenons six :

Dimension	Composantes	Ce qu'elle révèle
Enjeux	Souveraineté, politique, économique, écologique, réglementaire, temporel	Les tensions qui traversent toute décision concernant un data center et qui ne peuvent pas être réduites à des indicateurs techniques. Deux dimensions clés : souveraineté (qui décide, qui capte la valeur) ; écologie (foncier, eau, énergie, chaleur, effets rebond).
Infrastructures⁵	Foncier, eau, énergie, chaleur, refroidissement, puces/GPU, systèmes de secours, métaux critiques, bâtiment	Ce qui consomme, chauffe, s'use et devient obsolète. La couche physique qui détermine l'empreinte réelle et les exigences techniques des bâtiments et des raccordements.

⁴ <https://www.entreprises.gouv.fr/la-dge/actualites/choose-france-2026-un-nouveau-record-dinvestissements-en-faveur-de-lindustrie-de> ; <https://www.rte-france.com/bases-electricite/consommation-electricite/essor-data-centers-france>

⁵ Voir notamment l'ACV du data center DC19 publiée par Data4 <https://www.data4group.com/ressources/data-center-mesurer-limpact-pour-mieux-agir/> et, pour l'analyse des cycles de vie des GPU, de celles de l'ADEME <https://bibliothèque.ademe.fr/economie-circulaire-et->

Dimension	Composantes	Ce qu'elle révèle
Réseaux	Raccordement HTB/HTA ⁶ , flexibilité, câbles sous-marins, géographie	C'est un point de tension. C'est ici que la demande de calcul rencontre les postes électriques, les délais de raccordement, les règles d'accès et les conflits d'usage.
Acteurs	Opérateurs data center, collectivités, État/régulateurs, fonds/promoteurs	Des intérêts très différents. Forte asymétrie entre un opérateur mondial (un site représente une option dans un portefeuille) et une collectivité locale (un site est un enjeu économique, écologique, technique, politique, social structurant à 30 ans). Forte asymétrie aussi entre des acteurs privés très mobiles, financiarisés et internationalisés, et des acteurs publics ou territoriaux plus contraints par les procédures, les silos administratifs, les temporalités politiques et les règles d'instruction. Cette asymétrie peut <u>renforcer le poids des stratégies d'influence et compliquer la capacité publique à instruire les projets au bon niveau de granularité.</u>
Clients	Big Tech, organisations (mutualisé/non mutualisé), pays/États, individus	Le type de client détermine l'usage, la puissance requise et les exigences de souveraineté ; notamment la distinction entre acteurs soumis au Cloud Act américain et acteurs sous juridiction européenne .
Usages	Stockage, entraînement IA, calcul/inférence, SaaS, edge	Ce qui tourne dans le data center. C'est l'usage, pas le bâtiment, qui détermine la densité énergétique réelle, les exigences de latence, la localisation optimale et les exigences de raccordement.

Ces six dimensions ne sont jamais isolées. Une localisation peut créer une opportunité de chaleur fatale ou une externalité locale. Un choix d'usage transforme les exigences de latence, de puissance et de souveraineté. Une densification GPU modifie le refroidissement, le raccordement et la consommation d'eau. La suite du rapport aidera à instruire cette chaîne d'arbitrages.

Il faut d'abord donner au lecteur quelques repères essentiels et l'aider à déjouer les principaux pièges de communication.

dechets/9103-analyse-de-cycle-de-vie-de-gpu-cartes-graphiques-pour-l-intelligence-artificielle.html

⁶ HTA signifie **Haute Tension A**, souvent assimilée à la moyenne tension - entre **1 kV et 50 kV** (Enedis), avec un niveau courant de **20 kV** en France. HTB signifie **Haute Tension B** : Enedis la définit comme les tensions de **50 kV à 400 kV**, réservées au réseau de transport exploité par RTE.

1.2 Six pièges pour lire les chiffres

Parce que le data center est un système, ses chiffres sont complexes à lire. Dans le matériau des entretiens, nous avons relevé plusieurs pièges et en retenons ici six qui reviennent régulièrement :

Piège	Ce qui est annoncé	Ce qu'il faut lire
Puissance ≠ énergie	18 GW ⁷ de demandes de raccordement chez RTE	Un GW demandé ne signifie ni un GW construit, ni un GW consommé en permanence ⁸ . RTE retient un taux moyen d'utilisation de la puissance raccordée de 20% à l'horizon 2030 dans sa trajectoire de référence.
Annonce ≠ projet	109 Md€ d'investissements annoncés en France (dont 16 Md€ fléchés vers les infrastructures physiques selon une association énergétique).	Une annonce d'investissement ou une demande de raccordement ne vaut pas projet réalisé. Une part importante des demandes se situe encore en phase très amont ; certaines relèvent d'intentions, d'options stratégiques ou de positionnements fonciers. En même temps on assiste à une mise en mouvement que nous illustrons au travers des propos d'un interviewé « <i>L'un de nos gros clients annonce 3 Md€ dans les data centers et l'après-midi, passe quatre tranches de construction, alors qu'on attendait la première depuis longtemps. Face à l'emballement, chacun veut être le premier. Les promoteurs immobiliers s'y mettent aussi.</i> » Voir le chapitre 2 sur la file de raccordement.
Data center ≠ objet homogène	Les data centers consommeront X TWh	Un data center peut héberger du cloud généraliste, de l'entraînement de modèles IA, de l'inférence IA, du stockage, des usages critiques. Il est important de distinguer ce qui tourne dedans, à quelle intensité et avec quel taux d'occupation.
Montée en puissance ≠ saturation nationale	Les data centers vont saturer le réseau français.	« <i>Aujourd'hui, aucun projet prêt de data center n'attend son raccordement au réseau. C'est plutôt le réseau qui attend son projet</i> », un interviewé. De plus, la France a exporté 92,3 TWh en 2025.
Demande ≠ capacité industrielle	Des projets IA représentent plusieurs GW de puissance future	La contrainte est électrique, mais aussi industrielle. Elle dépend de la capacité mondiale à produire les briques matérielles qui rendent cette demande exécutable. Les entretiens soulignent que les goulots successifs - packaging avancé, mémoires, serveurs, raccordements, équipements - peuvent ralentir, déplacer ou renchérir la

⁷ Lors des entretiens, le chiffre de demandes avancé était de 14 GW ; RTE a ensuite annoncé une mise à jour à près de 18 GW de capacité réservée en mai 2026, pour environ 80 projets, contre 5 GW fin 2024. Ce passage de 14 à 18 GW illustre moins une consommation certaine que l'accélération des demandes de réservation de capacités sur le réseau. Source : RTE, <https://www.rte-france.com/bases-electricite/consommation-electricite/essor-data-centers-france>

⁸ Le **watt (W)** mesure la puissance, c'est-à-dire la capacité à transférer de l'énergie à un instant donné. Le **wattheure (Wh)**, quant à lui, mesure l'énergie effectivement consommée ou produite : il correspond à l'énergie transférée par une puissance de 1 watt pendant 1 heure. Un data center annoncé à plusieurs centaines de MW ne consomme donc pas mécaniquement cette puissance en permanence dès son ouverture. Sa consommation réelle dépend du taux d'occupation, du nombre de serveurs installés, du type de calcul, de la densité, du refroidissement, de la redondance et du profil de charge.
RTE donne un exemple : depuis 2016, huit data centers ont été raccordés au réseau de transport **pour une puissance totale contractualisée de 800 MW ; fin 2024, ils appelaient au maximum 120 MW, soit environ 15% de la puissance contractualisée.**
Source : RTE, Schéma décennal de développement du réseau 2025, fiche 5 « Raccordement de l'industrie », 2025.

Piège	Ce qui est annoncé	Ce qu'il faut lire
		croissance. « On fonctionne sur la capacité de production des serveurs : comment augmenter la capacité de production ? Comment allouer à certains pays ? La contrainte sur les mémoires n'avait pas été anticipée. Les connexions aux réseaux, la construction et les turbines n'avaient pas été anticipées non plus. », confie un expert des prévisions. Elle dépend aussi des délais de construction et des compétences mobilisables.
Projection ≠ prospective	Les data centers consommeront X TWh	Les chiffres de consommation future bornent le débat, mais ne testent pas les ruptures possibles : régulation, architectures de calcul, disponibilité des GPU, acceptabilité sociale, tensions géopolitiques. Voir ci-dessous

Projection ≠ prospective : « En 2035, les data centers consommeront X TWh »

Il nous a semblé important d'approfondir ce piège au-delà du tableau, car il peut conditionner beaucoup de décisions stratégiques. En effet, le débat public se focalise souvent sur des chiffres de consommation future, par nature incertains. Or une projection prolonge une trajectoire à partir d'hypothèses structurées ; une prospective teste aussi ce qui pourrait la casser.

Les projections permettent de fixer des ordres de grandeur, de comparer des scénarios et de sortir du débat impressionniste. Elles éclairent des trajectoires possibles, mais captent plus difficilement les ruptures : disponibilité réelle des serveurs et GPU, régulation de l'IA, acceptabilité sociale du numérique, tensions géopolitiques, autres familles de modèles ou ruptures dans les architectures de calcul, y compris, à terme, certaines formes de calcul spécialisé ou quantique.

Comme le résume une prospectiviste consultée : « Les travaux de l'ADEME, du Shift Project et de l'AIE sont solides et complémentaires pour poser des bornes quantitatives. Mais ils restent davantage des exercices de projection que de prospective au sens plein : ils prolongent des trajectoires plus qu'ils ne testent des futurs vraiment contrastés. Pour enrichir l'analyse, il faut travailler les incertitudes critiques (régulation de l'IA, architectures de calcul, acceptabilité sociale, tensions géopolitiques) et construire des scénarios de rupture, pas seulement des prolongements de tendances. »

Cette distinction est essentielle pour les data centers IA. Additionner des consommations futures ne suffit pas si l'on ne teste pas les facteurs susceptibles de changer la trajectoire : régulation forte de certains usages IA, bascule vers des modèles plus petits ou spécialisés, explosion d'agents autonomes très consommateurs, saturation sociale ou politique face aux infrastructures numériques, tensions sur les puces, changement de géographie industrielle, contraintes sur l'eau, le foncier ou les raccordements. Une trajectoire peut être robuste dans son modèle, tout en restant fragile face à ces ruptures.

Les projections doivent également être lues à travers les goulots physiques de la chaîne industrielle, notamment des serveurs effectivement produits, livrés, installés et exploités. « On raisonne sur

combien de serveurs seront livrés pour avoir une idée de la consommation d'énergie », poursuit le spécialiste des projections cité plus haut.

Cette approche rejoint la méthode de l'AIE, qui indique fonder une partie de sa modélisation sur les projections de livraisons de serveurs et sur le rythme d'adoption des accélérateurs IA - GPU, TPU et autres puces spécialisées, en tenant compte des contraintes d'offre.

Une projection fondée uniquement sur les annonces de projets peut aussi sous-estimer la force de la concurrence industrielle, lorsque plusieurs pays cherchent simultanément à capter les mêmes serveurs, les mêmes GPU, les mêmes mémoires, les mêmes équipements électriques et les mêmes compétences.

C'est précisément pourquoi l'exercice « combien les data centers consommeront-ils ? » doit être complété par une série de questions plus discriminantes : quoi, où, à quelle intensité, pour quels usages, à quelle maturité de projet, avec quelle valeur locale, selon quelle trajectoire industrielle et avec quelles hypothèses de rupture ?

Des événements récents montrent que la rupture peut aussi venir de l'intérieur du secteur. Anthropic estime qu'il serait souhaitable de disposer de la possibilité de ralentir ou de suspendre temporairement le développement des modèles de frontière lorsque les capacités progressent plus vite que les mécanismes de sûreté, dans le cadre d'une coordination entre industrie et pouvoirs publics. OpenAI a demandé l'instauration d'une régulation nationale obligatoire fondée sur les capacités des modèles les plus puissants. Après un incident au cours duquel certains de ses agents ont contourné leurs environnements d'isolation et compromis des systèmes internes ainsi que ceux de Hugging Face, l'entreprise a suspendu une partie de ses entraînements par renforcement destinés au déploiement.⁹ Sam Altman a également écarté une introduction en bourse d'OpenAI en 2026 dans ce contexte de préoccupations croissantes sur la sûreté.¹⁰

Cette inflexion révèle une difficulté structurelle : un acteur clé peut considérer qu'un ralentissement collectif serait souhaitable, tout en craignant de céder l'avantage à ses concurrents s'il ralentit seul. La presse économique décrit cette situation comme un « dilemme du prisonnier ». La coordination devient d'autant plus difficile que l'IA est désormais un enjeu industriel, économique et géopolitique entre entreprises et entre États.

Cette incertitude n'est pas périphérique à la question énergétique. Les investissements actuels en puces, data centers et capacités électriques reposent largement sur l'hypothèse d'une progression rapide et continue des capacités et des usages. Si des contraintes de sûreté, de régulation ou de coordination ralentissent cette trajectoire, certaines hypothèses de demande devront être révisées ; si la coordination échoue, la course peut au contraire entretenir une accélération dans laquelle les capacités progressent plus vite que les mécanismes de maîtrise. La gouvernabilité de l'IA devient ainsi elle-même une variable de la trajectoire matérielle de ses infrastructures.

En conclusion, les projections bornent le raisonnement, mais ne clôturent pas le débat. La prospective implique d'examiner les incertitudes, les ruptures, les arbitrages et les futurs réellement contrastés que les trajectoires moyennes rendent invisibles. Cette dimension manque encore dans le débat sur les data centers.

⁹ <https://openai.com/fr-FR/index/hugging-face-incident-and-the-road-ahead/> ; https://www.lemonde.fr/pixels/article/2026/09/12/comment-l-ia-a-suscite-un-vent-de-panique-cet-ete-dans-le-sillage-du-piratage-de-hugging-face_6770981_4408996.html

¹⁰ <https://www.reuters.com/legal/litigation/openai-ipo-will-not-happen-2026-amid-ai-safety-fears-altman-says-2026-09-12/>

Ce que les autres marchés nous apprennent

Les marchés étrangers montrent que la croissance des data centers ne produit pas partout les mêmes effets, mais qu'elle révèle partout les mêmes tensions : files d'attente gonflées, concentration géographique, pression sur les réseaux, arbitrages entre attractivité et externalités, et besoin de critères d'acceptation explicites. L'homologue italien de RTE, Terna¹¹, indiquait environ **30 GW** de demandes de raccordement de data centers fin 2024, **plus de 50 GW** au 30 juin 2025, puis **69,05 GW fin 2025** et **78,79 GW en janvier 2026**. Le Royaume-Uni a la même dynamique de pipelines hypertrophiés : des dizaines de GW de projets peuvent être annoncés sans correspondre à des capacités effectivement construites ni consommées. Ces ordres de grandeur confirment que la puissance annoncée peut devenir une option stratégique sur le réseau avant d'être un projet industriel réel. L'Irlande et les Pays-Bas ont déjà expérimenté des tensions fortes, jusqu'à des restrictions ou moratoires. D'autres pays ont aussi dû prendre des mesures : Singapour¹² a conditionné la reprise du développement à des exigences d'efficacité et de consommation de ressources. « *Dans l'État de Johor, en Malaisie¹³, environ 30% des demandes de data centers auraient été rejetées sur cinq mois en 2024 parce qu'elles ne répondaient pas aux exigences de durabilité et de conservation des ressources ; les autorités ont mis en place un comité d'évaluation face aux tensions sur l'eau et l'électricité* », illustre un expert.

La leçon pour la France est de « *bénéficier de la croissance, pas à n'importe quel prix* », comme le suggère un interviewé.

Plusieurs entretiens convergent vers une même alerte : « Tant que les annonces se multiplient, le marché donne une impression de dynamisme. Mais si les réseaux, le foncier, l'eau, les autorisations et les compétences arrivent à saturation, cette attractivité peut vite devenir une fragilité, surtout lorsque les projets n'ont pas le même niveau de maturité, que leurs usages finaux restent flous, que leurs retombées territoriales ne sont pas garanties », comme le précise une interviewée. La doctrine française devrait donc se construire avant la saturation.

Ces exemples montrent aussi qu'entre le moratoire et le laisser-faire, il existe une troisième voie : une allocation sélective, fondée sur des critères publics, des exigences de durabilité, une transparence des usages et une capacité à refuser les projets qui déplacent plus de contraintes qu'ils ne créent de valeur.

1.3 Le paradoxe français

La France présente une configuration singulière : son mix électrique est décarboné à plus de 95%¹⁴, nucléaire et renouvelables combinés. Excédentaire, en 2025, elle a exporté 92,3 TWh d'électricité.

Elle dispose aussi d'une position géographique centrale dans le maillage européen des data centers, des points d'échange et des câbles sous-marins. Paris appartient aux grands marchés européens FLAPD - Francfort, Londres, Amsterdam, Paris et Dublin- qui concentrent une part importante de la capacité opérationnelle régionale. Marseille, de son côté, est devenue un hub majeur de câbles

¹¹ <https://lightbox.terna.it/en/insight/data-center-trasmission-grid>

¹² <https://www.reedsmith.com/articles/singapores-data-centre-expansion-jurong-island-sustainable-growth/>

¹³ <https://www.nst.com.my/news/nation/2024/11/1136609/johor-rejects-30-cent-data-centre-bids>

¹⁴ Source : RTE, *Bilan électrique 2025* et <https://www.rte-france.com/bases-electricite/consommation-electricite/essor-data-centers-france>

sous-marins. Selon l'un de nos interviewés, « *Paris représente le quatrième ou cinquième volume de données en Europe selon les périmètres retenus, derrière des points d'échange majeurs comme Amsterdam ou Londres. Marseille constitue le deuxième volume d'échange en France après Paris, avec une ouverture stratégique vers la Méditerranée et le Moyen-Orient. Ces dorsales Internet mondiales fonctionnent en synergie : Paris est interconnectée avec le Nord de l'Europe, Londres, Francfort, Berlin ou Bordeaux, tandis que Marseille joue un rôle clé dans les flux vers le Sud et l'Est.* »¹⁵.

Dans ce contexte, les data centers peuvent apparaître comme de « bons clients » pour une électricité bas-carbone qui cherche des débouchés. Plusieurs entretiens formulent cette lecture, comme une **opportunité stratégique et aussi comme un paradoxe subtil**.

« *Vus de RTE, les data centers peuvent être convoqués comme de nouveaux consommateurs capables d'absorber une production électrique bas-carbone abondante. Mais le surplus relatif d'aujourd'hui peut devenir le déficit de demain si les autres usages - véhicules électriques, pompes à chaleur, industrie, ferroviaire - montent en charge comme prévu* », précise un expert.

Ce surplus doit aussi être compris comme l'effet d'une électrification des usages en retard (véhicules électriques, pompes à chaleur, rénovation, industrie) et non comme une ressource définitivement disponible pour n'importe quel usage électro-intensif. **L'avantage bas-carbone peut produire un effet d'appel comparable, dans sa logique, à certains mécanismes de rebond** : plus l'électricité paraît propre, plus elle devient disponible symboliquement pour de nouveaux usages électro-intensifs.

La durée de ce surplus relatif reste donc incertaine et dépendra notamment du rythme effectif d'électrification des transports, du chauffage et de l'industrie. Cette incertitude doit être intégrée à l'instruction des projets plutôt que traitée comme une disponibilité durable de capacité électrique.

La grille suivante vise à aider dans l'instruction de cette dynamique.

1.4 La grille analytique : substitution, relocalisation, addition, effet d'appel

Ce n'est pas parce qu'un data center émet moins de CO₂ en France qu'ailleurs qu'on doit automatiquement l'accueillir. Il faut notamment qualifier la nature de la demande, ce que le data center fait réellement au système énergétique et territorial. Nous pouvons distinguer quatre situations qui ne produisent pas le même bilan stratégique¹⁶ :

¹⁵ Le verbatim d'entretien parle de « volume de données » et de « points d'échange », mais les sources publiques disponibles documentent plus solidement le rang de Paris comme **marché data center** que comme point d'échange Internet au sens strict. Cushman & Wakefield classe Paris parmi les marchés **FLAPD** (Frankfurt, London, Amsterdam, Paris, Dublin) et indique que ces marchés concentrent environ 45% de la capacité opérationnelle régionale ; une autre publication de Cushman & Wakefield présente **Paris comme le 4e plus grand marché data center EMEA**. Pour Marseille, Orange indique qu'avec Medusa, la ville accueille désormais 17 atterrages de câbles sous-marins, ce qui en ferait le **premier hub européen et le sixième mondial** ; Telehouse présente Marseille comme le 9e hub mondial de connectivité, avec des liaisons rapides vers l'Afrique, le Moyen-Orient et l'Asie. Sources : <https://www.cushmanwakefield.com/en/united-kingdom/news/2025/02/powerhouse-markets-lead-emea-data-centre-growth>, <https://www.orange.com/en/whats-up/new-submarine-cable-strengthens-connectivity-across-mediterranean> et <https://www.marseille-port.fr/en/projets/cables-sous-marins>

¹⁶ Cette grille est une proposition analytique du rapport. Elle complète les approches territoriales d'instruction des projets, notamment celles qui portent sur le foncier, l'eau, la chaleur, les nuisances et les retombées locales, tout en se concentrant sur une autre question : la nature énergétique et stratégique de la demande créée ou déplacée.

Situation	Ce qui se passe	Ce que ça implique
Substitution « faire autrement »	Un usage numérique est exécuté autrement dans une infrastructure plus efficace, mieux dimensionnée ou moins carbonée, sans créer de nouveaux volumes d'usage. On remplace un data center peu performant et un mode d'exécution énergivore.	L'intérêt environnemental peut être réel si le gain carbone est effectif et si les autres impacts sont maîtrisés : consommation électrique, taux d'occupation, refroidissement, PUE/WUE, matériel, foncier, réseau et éventuels effets rebond. Le critère clé est : le nouveau data center remplace-t-il effectivement une infrastructure ou un mode d'exécution moins performant ?
Relocalisation « faire ici ce qui était exécuté ailleurs »	Une charge numérique française ou européenne existante, aujourd'hui exécutée dans un autre pays ou dans un système électrique plus carboné, est déplacée vers la France en raison de son mix bas-carbone, de sa connectivité ou de garanties de souveraineté.	L'argument est crédible si la charge relocalisée est identifiable, déjà existante, utile, substitutive et si le déplacement vers la France ne sert pas à créer une demande additionnelle. Il suppose aussi des contreparties contractualisées : transparence sur les usages, valeur locale, maîtrise juridique, sobriété foncière, gestion de l'eau, chaleur fatale, flexibilité réseau.
Addition	Un data center crée de nouveaux usages, de nouveaux volumes, sans remplacer une consommation existante. Addition justifiée (nouvel usage, utilité démontrée) ; non justifiée (consommation qui s'ajoute sans preuve d'utilité).	La question de la valeur captée et des conflits d'usage avec d'autres besoins électriques devient centrale L'argument environnemental peut s'affaiblir fortement s'il s'agit d'un usage non justifié.
Effet d'appel	L'avantage bas-carbone français devient un argument pour faire croître la demande numérique elle-même.	L'avantage bas-carbone alimente un effet d'appel : il devient lui-même un facteur de croissance de nouveaux usages électro-intensifs.

« Un data center en France émet-il moins qu'ailleurs ? » a donc peu d'intérêt. Il faudrait plutôt se demander : « ce data center remplace-t-il une charge existante, relocalise-t-il une charge utile, ajoute-t-il une nouvelle demande ou crée-t-il un effet d'appel pour des usages numériques supplémentaires ? »

Une consommation ici pour un travail rendu ailleurs

Un data center peut consommer en France pour produire une valeur captée ailleurs : l'électricité, le foncier, l'eau, les raccordements et les nuisances sont localisés ; mais les modèles, les services, les données exploitées, les marges et les décisions peuvent être captés hors du territoire. « Pour quel travail rendu, au bénéfice de qui, avec quelle valeur locale et quelles contreparties vérifiables ? »

Cette première lecture conduit à une proposition : un projet de data center ne devrait pas être instruit seulement par sa puissance demandée et son investissement annoncé, mais aussi par ses usages, sa maturité, sa localisation, son profil de charge, les ressources qu'il mobilise, sa place dans la grille substitution / relocalisation / addition / effet d'appel, les contreparties qu'il garantit et la valeur qu'il laisse réellement au territoire.

Chapitre 2 - Le raccordement : test de réalité des projets IA

Le chapitre en une phrase

La file d'attente au raccordement agit comme un révélateur : elle distingue les projets industriels engagés des options de réservation. Le chapitre propose quatre niveaux d'instruction - maturité du projet, nature de l'usage, pertinence de la localisation, contribution contractualisée - pour transformer les promesses en engagements vérifiables avant l'autorisation, plutôt qu'après.

En raccordant les data centers, le réseau électrique devient un révélateur d'arbitrages : où, quand, pour quels usages, avec quelle maturité de projet et selon quels arbitrages d'intérêt collectif cette électricité doit être rendue disponible. Ce chapitre montre pourquoi le réseau porte des choix techniques, tout en étant aussi un révélateur de choix politiques non formulés.

2.1 La file de raccordement : un symptôme, pas un problème technique

Les 18 GW de demandes de raccordement enregistrés par RTE décrivent aussi une compétition pour l'accès à une ressource stratégique : la capacité raccordable. Ce chiffre ne dit donc pas « combien sera consommé », mais « combien d'acteurs cherchent à réserver une option sur le réseau » et « quelle est leur stratégie pour capter les capacités électriques françaises ? »

« Le restaurant est vide, mais on n'a pas de tables disponibles parce qu'elles ont toutes été réservées », partage métaphoriquement un interviewé d'une organisation énergétique.

Le réseau n'est plus seulement sollicité pour alimenter des projets, mais aussi pour sécuriser des options. Comme le résume un acteur interviewé : « *On a vu trois cycles de consommation électrique : d'abord les gigafactories, avec une volonté industrielle plus réelle ; puis l'hydrogène, aujourd'hui ralenti faute de subventions et de compétitivité ; désormais les data centers, perçus comme un nouvel eldorado. La question est même de savoir si certains acteurs de l'hydrogène ne se repositionnent pas vers les data centers.* » Chaque vague électro-intensive arrive ainsi avec ses promesses - emplois, souveraineté, transition - tout en mobilisant les mêmes ressources rares : raccordement, foncier, délais administratifs, équipements et compétences.

La métaphore du restaurant vide dit l'essentiel : on y trouve des projets prêts à construire, des intentions, des options stratégiques, des positionnements fonciers et parfois des demandes de précaution. Le problème n'est pas seulement le volume de la file, mais son opacité : y sont confondues demande productive, demande spéculative et demande de précaution.

L'exemple britannique illustre ce risque de file hypertrophiée. Un expert interviewé évoque « 50 ou 55 GW de projets » et précise que « certains projets demandent plusieurs fois la même connexion, d'autres doublent la demande pour avoir de la capacité ».

En France, dans un système encore largement structuré par le principe du « premier arrivé, premier servi », la file devient difficile à filtrer. La question n'est plus seulement la puissance demandée, mais la maturité des projets, ainsi que leur fiabilité financière et industrielle.

Dans ce contexte, dimensionner trop tôt des infrastructures (lignes, postes, capacités réservées, foncier préparé, études et travaux engagés) pour des projets incertains peut devenir un risque économique, industriel, environnemental, social, sauf à pouvoir réallouer ces capacités à d'autres usages électrifiés. L'exemple malaisien a été cité lors d'un entretien : « *En Malaisie, le gestionnaire du réseau a anticipé la croissance des infrastructures, mais s'il n'y a pas de consommation, il y a*

un risque d'actifs échoués. La chance, c'est si on compense par d'autres demandes - froid, électrification industrielle, véhicules électriques... ».

Sans doctrine, l'incertitude privée peut devenir un coût collectif.

« 14 GW¹⁷, c'est 150 à 200 milliards d'euros de travaux, soit quatre fois le Grand Paris Express. On n'a pas les entreprises pour les construire », entretien avec un expert de la construction.

Ce verbatim déplace le sujet : la file de raccordement n'est pas seulement une file administrative ou électrique. Elle devient un révélateur de capacité industrielle. La question n'est pas uniquement de savoir si le réseau peut raccorder les data centers, mais si la France doit engager des ressources rares - entreprises, matériaux, équipements, ingénierie, délais - au service de projets dont la réalisation, les usages et les retombées ne sont pas encore démontrés.

2.2 RTE et Enedis : deux opérateurs, deux univers

La structure du réseau français distribue les responsabilités de raccordement entre deux opérateurs selon la puissance demandée, créant des dynamiques et des tensions spécifiques à chaque niveau.

Opérateur	Périmètre	Situation actuelle
Enedis (HTA)	Projets < 40 MW. Plus de 300 data centers raccordés	Délais passés de 12 à 24 mois en deux ans.
RTE (HTB)	Projets > 40 MW. Moins de 10 data centers étaient dans ce cas en 2021	Délais variables selon la puissance et la complexité du raccordement : ordre de grandeur d'environ 40 mois pour des projets de 50 à 100 MW, jusqu'à 60-65 mois pour des projets de plusieurs centaines de MW.

Un phénomène de « saucissonnage » est signalé dans plusieurs entretiens : certains projets peuvent solliciter d'abord un raccordement auprès d'Enedis, avant d'évoluer vers un raccordement au réseau de transport lorsque leur besoin de puissance augmente. « *Quand il n'y a pas suffisamment de capacité au niveau RTE, ils vont chez Enedis en disant qu'ils n'ont pas besoin de cette capacité. Enedis fait les travaux pour eux. Après, ils vont chez RTE. Enedis a fait les travaux pour rien.* », entretien avec une organisation énergétique. Cette pratique peut compliquer la planification des réseaux et mobiliser successivement plusieurs niveaux d'infrastructure. Elle ne signifie toutefois pas que la liaison réalisée par Enedis devient nécessairement inutile : selon la configuration retenue par l'exploitant, elle peut notamment être conservée comme liaison de secours après le raccordement au réseau de transport.

2.3 Les effets collatéraux : quand le raccordement d'un projet bloque d'autres transitions

L'engorgement de la file d'attente peut pénaliser les opérateurs de data centers sérieux et créer un risque systémique pour la transition énergétique française.

Exemples de risques de conflits d'usage cités dans les entretiens

Les exemples suivants sont des risques identifiés dans les entretiens.

¹⁷ Comme expliqué plus haut, lors des entretiens, le chiffre de demandes avancé était de 14 GW ; en mai 2026, RTE indiquait avoir réservé près de 18 GW de capacité.

Concurrence avec un industriel : « Une usine c'est 10 MW et 500 emplois, un data center c'est 60 MW et 3 emplois. Dans la règle actuelle, c'est le dernier arrivé qui paye. On n'a pas le droit de dire : dépêchez-vous, il n'y a plus que 10 ou 20 MW. L'industriel ira ailleurs. »¹⁸, entretien avec un responsable territorial. La capacité d'un poste source local est réservée par un data center de 60 MW qui ne la consommera peut-être pas. L'industriel souhaitant s'implanter avec une usine de 10 MW créatrice de 500 emplois peut se retrouver en attente, avec le risque d'y préférer un autre site.

Un hôpital en attente : la capacité du poste source local a été réservée par un data center dont le projet ne se réalisera peut-être jamais. L'hôpital attend.

Un verrier ou un aciériste : souhaite électrifier son process, mais trouve les capacités de raccordement saturées dans sa zone industrielle. Il continue à polluer.

Le coût d'opportunité industriel d'un data center ne se lit donc pas seulement dans le foncier ou l'électricité mobilisés, mais aussi dans les autres usages susceptibles d'être contraints lorsque plusieurs demandes de consommation se présentent sur une même zone. Les règles de raccordement ne sont toutefois pas identiques pour toutes les catégories de projets : les installations de production renouvelable relèvent notamment de dispositifs spécifiques de planification et de raccordement.

Ces conflits d'usage ne concernent pas seulement les grands sites électro-intensifs visibles dans les feuilles de route nationales. Ils peuvent aussi toucher l'industrie diffuse : agroalimentaire, chimie fine, papeterie, séchage, froid industriel ou sites de production intermédiaires. Ces activités demandent individuellement moins de puissance qu'un grand data center, mais peuvent être déterminantes pour l'emploi local, la décarbonation productive et la résilience économique d'un territoire.

L'enjeu n'est pas d'opposer par principe data centers et industrie. L'intérêt stratégique, technologique ou économique d'un data center peut lui aussi être important pour le territoire. Il s'agit plutôt de rendre explicites les arbitrages lorsque plusieurs projets sollicitent une capacité locale limitée et d'instruire leur maturité, leurs besoins, leurs usages et leurs retombées respectives.

Comment organiser ces arbitrages entre demandes électro-intensives successives, alors que leur maturité, leurs usages, leurs besoins de puissance et leurs retombées territoriales peuvent être très différents ?

2.4 La localisation : ce que le réseau révèle des territoires

Les entretiens montrent que la question réseau est territorialisée. Un scénario national peut indiquer que la France dispose d'une trajectoire électrique favorable, mais localement, un poste source, une ligne, une zone industrielle ou une intercommunalité sont des lieux d'arbitrage très concrets.

À Marseille, cette territorialisation est particulièrement visible. Les entretiens décrivent une convergence entre décarbonation du port, développement urbain, data centers, transport et ouvrages de sûreté. La démarche territoriale a été fondée sur l'identification des parties prenantes, la compréhension des stratégies à moyen terme et l'usage d'outils de géomatique pour cartographier les

¹⁸ La question des coûts de raccordement est également structurante. Les entretiens mentionnent la règle dite « *shallow cost* » : celui qui déclenche un investissement paie, avec des mécanismes d'abattement ou de mutualisation. Dans certaines zones comme le sud de l'Île-de-France, des capacités ont été définies pour mutualiser les coûts et éviter des effets de seuil. Mais la mutualisation n'est pas neutre : mutualiser pour plusieurs data centers n'a pas le même sens que mutualiser pour data centers, industrie, port, mobilité lourde et équipements publics. Elle suppose de décider pour quels types de consommateurs on dimensionne l'infrastructure.

projets et anticiper les infrastructures associées. L'orientation des data centers a été travaillée en lien avec l'État, RTE, Enedis, les réseaux et les projets industriels, afin d'identifier des zones disposant de capacité électrique, de proximité numérique, d'eau et de possibilités de valorisation de chaleur.

2.5 Du « premier arrivé » au « premier prêt » : une réforme de doctrine

La règle historique du « premier demandeur, premier servi »¹⁹ crée un système d'incitations qui peut devenir contre-productif : il suffit de déposer un dossier pour bloquer de la capacité, sans démontrer la moindre intention réelle de construction. RTE a lancé en mars 2026 une consultation publique pour faire évoluer cette logique vers un principe de « premier prêt, premier servi ».

Le rapport du Sénat du 16 septembre 2026²⁰ recommande également le passage au « premier prêt, premier servi » et l'inscription du fast track au cœur de la stratégie de RTE. Il propose en parallèle de rendre prioritaires pour le raccordement les centres de données dédiés à l'IA. Cette proposition ouvre une question supplémentaire : lorsqu'un data center, un projet industriel ou une opération de décarbonation sont tous prêts, selon quels critères départager leurs demandes de capacité ? Le « premier prêt » ne clôt donc pas l'arbitrage : il en déplace la frontière.

Le dilemme Use It Or Lose It (UIOLI)

Une partie du gonflement des chiffres ne relève pas d'une spéculation pure. Un opérateur de data center ne sait pas toujours quelle densité de calcul, quel type de refroidissement ou quel niveau de puissance par baie il devra accueillir dans dix ans. Demander large peut donc être une stratégie de couverture rationnelle, notamment face à la densification rapide des racks GPU.

Ce dilemme rend nécessaire une contractualisation fine : jalons de maturité, phasage, montée en puissance conditionnelle, pénalités de non-réalisation et réallocation des capacités non utilisées.

La question de doctrine est donc : « comment faire payer l'incertitude à ceux qui la créent, sans décourager les projets réellement matures et utiles ? ».

La réforme « premier prêt » conditionnerait l'octroi d'une capacité de soutirage à la démonstration que le client est en phase de réalisation, prêt à commander le matériel, les puces, les équipements de refroidissement. Ce point transformerait le raccordement en test de crédibilité industrielle.

¹⁹ À l'origine, elle a été instituée pour éviter les passe-droits. Aujourd'hui, des logiques davantage financières qu'industrielles peuvent en perturber le fonctionnement.

²⁰ <https://www.senat.fr/rap/r25-949/r25-9496.html>

Mécanismes existants ou en évolution	Ce qu'ils changent
Premier prêt, premier servi	Conditionne l'accès réseau à une maturité démontrée (permis de construire, plan de financement, lettre d'intention client)
UIOLI (Use It Or Lose It)	Entré en vigueur en août 2025. Permet à RTE d'ajuster les contrats surdimensionnés. Fin 2024, des data centers raccordés à RTE appelaient au maximum environ 15 % de leur puissance contractualisée.
Fast Track²¹	Raccordement en 3-4 ans sur des sites pré-identifiés. RTE indique avoir déjà pré-réservé des capacités sur cinq sites de forte puissance, de 400 MW à 1 GW (voir rapport du Sénat cité plus haut).
Raccordement anticipé avec limitation	Connexion accélérée si le client accepte une limitation de charge (1 000 à 2 000 heures), permettant de réduire les délais en échange d'une contrainte temporaire

L'AMI EDF²² sur d'anciens sites de production est une illustration de cette ingénierie territoriale. En novembre 2025, EDF et OpCore ont annoncé des négociations exclusives pour un data center de plusieurs centaines de MW sur le site de l'ancienne centrale thermique de Montereau, avec un investissement annoncé d'environ 4 milliards d'euros, bénéficiant du dispositif fast track. Ce cas relie plusieurs dimensions : foncier déjà industrialisé, raccordement existant, réutilisation d'un site, stratégie publique et opérateur privé. Il montre que l'accueil peut être orienté vers des actifs fonciers déjà inscrits dans l'histoire électrique du territoire.

« L'offre EDF AMI est un modèle intéressant de développement des data centers. Il permet la maîtrise du foncier, la maîtrise des enjeux territoriaux et la propriété du foncier. Par exemple, la stratégie d'EDF Montereau, ancienne centrale à charbon, c'est OpCore qui l'a remportée. Les deux autres, c'est Eclairion. Objectif : avoir de grosses capacités. On ne connaît pas les différentes familles de clients : cloud, entreprises... Cela permettrait de voir les tendances. », explique un interviewé.

Ces AMI pourraient ouvrir une piste stratégique plus large, comme nous l'expliquait un expert de la construction : *« l'ordre de grandeur économique d'un data center - un site de 100 MW peut représenter environ 1 milliard d'euros d'investissement et cela change l'arbitrage foncier : dans ce contexte, 20 millions d'euros de dépollution d'une friche industrielle deviennent une charge relative faible dans le modèle économique global du projet »*. Cette logique rejoint un autre point soulevé dans l'entretien : *« 300 MW, le privé paie la ligne de tension, l'infrastructure est payée par le privé... si à la fin ils ne trouvent pas de preneurs, l'infrastructure aura été construite et pourra servir à d'autres usages »*. Développer ce sujet, c'est finalement créer un mode de financement d'infrastructures utiles potentiellement pour d'autres usages. Faire dialoguer *« et si ça se faisait »* avec *« et si ça ne se faisait pas »*. Les data centers peuvent ainsi devenir un levier de requalification

²¹ <https://www.cre.fr/actualites/toute-lactualite/la-cre-approuve-la-procedure-de-raccordement-fast-track-pour-raccorder-les-sites-de-consommation-de-grande-puissance-au-reseau-tres-haute-tension-400-kv-de-rte.html> ; <https://www.services-rte.com/fr/actualites/publication-de-la-proc%C3%A9dure-fast%20track.html>

²² Appels à manifestation d'intérêt, voir <https://www.edf.fr/groupe-edf/espaces-dedies/journalistes/tous-les-communiqués-de-presse/edf-et-opcore-s-engagent-pour-le-développement-d-un-centre-de-données-de-grande-puissance-sur-le-site-de-l-ancienne-centrale-thermique-de-montereau-vallee-de-la-seine-seine-et-marne>

de sites dégradés ou pollués, à condition que cette logique soit portée par une stratégie territoriale explicite et non laissée à la seule initiative de l'opérateur.

2.6 La flexibilité : levier central, partage des coûts non réglé

La flexibilité, à savoir la capacité d'un consommateur à moduler sa demande en fonction de l'état du réseau, est présentée par plusieurs acteurs comme un levier structurant, à la fois pour faciliter les raccordements et pour optimiser l'utilisation du réseau. Elle n'est pas nouvelle : Gimelec y travaille depuis 2012 et a publié un livre blanc spécifique en novembre 2024²³. Une loi de 2022²⁴, dite loi MUPPA, impose aux sites de consommation utilisant des installations de production ou de stockage d'électricité de plus de 1 MW pour leur alimentation de secours de mettre à disposition de RTE la puissance non utilisée et techniquement disponible lors des périodes de forte tension, par l'intermédiaire du mécanisme d'ajustement.

Les expériences pionnières existent. Le démonstrateur DCflex²⁵ (RTE, Data4²⁶, Schneider Electric) sur le campus de Marcoussis est présenté par RTE comme le premier démonstrateur européen dédié à l'interaction entre data centers et système électrique. Il vise à tester la capacité des data centers à s'adapter aux conditions du réseau, notamment leur comportement face à un creux de tension, afin d'éviter qu'une déconnexion soudaine ne dégrade la stabilité du système.

« La flexibilité est un enjeu de raccordement. Il y a une exigence en Île-de-France, proposition de RTE. Il faut l'intégrer dès le départ. », entretien avec un syndicat de l'énergie.

Cependant, plusieurs obstacles freinent sa mise en œuvre généralisée. Les opérateurs de data centers hésitent à s'engager sur des capacités de flexibilité sans visibilité sur leur propre montée en charge. La question du partage des coûts et des bénéfices entre RTE, les opérateurs et les clients finaux n'est pas réglée. Les grands acteurs du cloud pourraient théoriquement reporter des charges entre leurs data centers dans plusieurs pays, mais ils n'ont pas aujourd'hui d'incitation économique, contractuelle ou réglementaire suffisante, par exemple via des exigences de plafonnement d'émissions, pour le faire au bénéfice du système électrique français. Comme le résume un acteur interviewé : *« à cause de l'intermittence, on aimerait que les clients raccordés puissent être flexibles pour faire fonctionner le système à moindre coût. On souhaite des acteurs qui acceptent de développer des business flexibles. Les GAFAM peuvent faire des reports de charge de l'un de leurs data centers vers l'autre. Il faudrait gérer les congestions du réseau avec des mécanismes d'ajustement de prix du kWh, mais dans le coût du kWh, la part RTE est portion congrue. C'est le fournisseur qui fixe le prix du kWh. Il faudrait l'intégrer dès le départ. »*

Les data centers ont par ailleurs des exigences fortes de continuité de service : leur capacité réelle de modulation dépend des contrats clients, des batteries, de la redondance et des SLA (service level agreement).

Un data center ne devrait pas seulement être évalué par la puissance qu'il demande, mais aussi par sa capacité réelle à se comporter comme un acteur du système électrique, ce que DCflex cherche précisément à qualifier.

²³ <https://gimelec.fr/en/contribution-des-data-centers-francais-a-la-flexibilite-du-systeme-electrique/>

²⁴ <https://www.services-rte.com/fr/decouvrez-nos-offres-de-service/impact-de-la-loi-muppa-sur-le-code-de-l-energie-une-faq-a-votre-disposition.html>

²⁵ <https://www.rte-france.com/actualites/projet-dcflex-epri-contexte-narratif-premier-demonstrateur-france>

²⁶ <https://www.data4group.com/actualites-data4/data4-rte-et-schneider-electric-lancent-le-1er-demonstrateur-europeen-sur-linteraction-entre-data-centers-et-systeme-electrique/>

La flexibilité peut devenir un service rendu au système électrique : modulation de charge, effacement, mécanismes d'ajustement, mobilisation des batteries, support de tension. Mais elle ne deviendra un levier structurant que si trois questions sont explicitées : 1) qui finance les équipements nécessaires ; 2) qui accepte la contrainte opérationnelle ; 3) comment se répartit la valeur créée entre le gestionnaire de réseau, l'opérateur, le client final, le territoire.

Face à la croissance des data centers IA, plusieurs solutions sont avancées : PPA²⁷ renouvelables, batteries, turbines à gaz temporaires, SMR, hydrogène, récupération de chaleur, flexibilité, déplacement temporel des charges... Chacune peut avoir un intérêt dans un contexte précis. Mais aucune ne doit devenir un argument générique d'acceptabilité : un PPA annuel ne garantit pas un alignement horaire entre consommation et production ; une turbine à gaz peut résoudre un délai de raccordement, et créer une dépendance fossile. « *Aux USA, parfois ils installent des turbines à gaz, le temps de se connecter au réseau. Dans la pratique, ils peuvent ensuite les maintenir si le raccordement réseau tarde ou si le projet change d'échelle. En Irlande, le projet annoncé la semaine dernière est sur du fossile.* », explique une organisation énergétique.

Un SMR, pour *Small Modular Reactor*, désigne un petit réacteur nucléaire modulaire, conçu pour être de puissance plus limitée qu'un réacteur nucléaire conventionnel et, en principe, industrialisable par modules. Cette option peut être pertinente à terme pour certains besoins industriels continus et fortement électro-intensifs. Mais, à court terme, elle reste une hypothèse industrielle, réglementaire et économique non disponible pour répondre aux besoins immédiats de raccordement des data centers IA. L'AIE indique que les premiers projets commerciaux de SMR devraient démarrer autour de 2030, tandis que plusieurs projets restent soumis à des incertitudes de coûts, de calendrier, d'autorisation et de chaîne d'approvisionnement.

La pile à hydrogène peut, elle aussi, jouer un rôle dans certains systèmes : alimentation de secours, remplacement partiel de groupes diesel, stockage ou restitution d'électricité dans des architectures locales hybrides. Mais elle ne constitue pas une réponse directe à la croissance des data centers IA. D'abord parce qu'elle ne produit pas une énergie primaire : elle restitue de l'électricité à partir d'hydrogène qu'il a fallu produire, stocker, transporter et convertir, avec des pertes à chaque étape. Ensuite parce que l'hydrogène bas-carbone reste une ressource rare, déjà attendue pour des usages difficiles à électrifier, notamment industriels. Enfin parce qu'une pile à hydrogène répond plutôt à un besoin de secours, de continuité ou de flexibilité ponctuelle qu'à une demande massive, continue et croissante d'électricité pour l'inférence et l'entraînement IA.

La doctrine doit donc distinguer les solutions qui réduisent réellement une contrainte systémique de celles qui la déplacent : du réseau vers le gaz, du carbone vers l'eau, de l'électricité vers le foncier, de l'empreinte locale vers une promesse contractuelle difficile à vérifier. « *La solution réduit-elle effectivement les impacts ou rend-elle simplement le projet plus acceptable en apparence ?* », entretien avec un syndicat de l'énergie qui poursuit : « *les data centers doivent apprendre comment gérer un réseau. L'inverse est vrai aussi.* »

2.7 Vers une doctrine de raccordement ?

Les éléments qui précèdent nous montrent que le réseau ne peut donc pas rester le lieu où l'on découvre trop tard les arbitrages. À partir des entretiens émergent quatre niveaux d'instruction que les réformes des règles d'accès pourraient considérer :

²⁷ *Power Purchase Agreement* (PPA) : contrat d'achat d'électricité conclu directement entre un producteur et un acheteur, généralement sur une durée longue et selon des conditions convenues de prix et de volume.

Niveau	Ce qu'il implique
Maturité du projet	Distinguer un projet industriel engagé d'une option de réservation. Le passage de la logique « premier demandeur » à une logique « premier prêt » constitue un premier filtre, à compléter, notamment par des jalons vérifiables, des garanties financières et des conséquences en cas de non-réalisation.
Nature de l'usage	Qualifier ce qui sera réellement hébergé ou calculé : cloud généraliste, capacité IA privée, infrastructure souveraine, calcul scientifique, service mutualisé ou actif purement spéculatif. Ces usages n'ont pas la même valeur locale, ni les mêmes implications de souveraineté.
Pertinence de localisation	Évaluer le coût réel d'un MW selon le territoire : robustesse du réseau, saturation locale, proximité numérique, disponibilité foncière, synergies industrielles, synergies avec le réseau électrique, contraintes portuaires ou métropolitaines...
Contribution contractualisée	Transformer les promesses en engagements : flexibilité, chaleur fatale, partage des coûts, transparence des données, réversibilité du site, dépollution éventuelle, emplois qualifiés et retombées locales.

Ce cadre invite à traiter le data center comme une pièce d'ingénierie territoriale : que permet-il de réparer, de mutualiser, de raccorder, de chauffer, de dépolluer, de préparer et à quel coût collectif, environnemental et stratégique ?

Un ensemble d'informations pourrait être demandé : puissance demandée, puissance phasée, maturité réelle, usage prévu, crédibilité industrielle, localisation, profil de charge, flexibilité et contreparties territoriales, investisseurs finaux, clients pressentis, eau, chaleur, groupes de secours, réversibilité, serveurs commandés, disponibilité des GPU, technologie de refroidissement... Selon les cas, ces informations permettraient d'éclairer les arbitrages de raccordement et de planification.

Il ne s'agit pas de rendre publics des secrets industriels sensibles. Ces informations permettraient plutôt à l'État, aux gestionnaires de réseau et aux collectivités de distinguer les projets prêts des options spéculatives, les usages critiques des usages ordinaires, les infrastructures souveraines des infrastructures captives, les projets bénéfiques des projets opportunistes.

Le réseau révèle qui réserve la puissance, le territoire révèle qui en assume les conséquences.

Chapitre 3 - Le territoire : là où l'attractivité devient négociation

Le chapitre en une phrase

La concentration géographique des data centers n'est pas une fatalité mais une logique de marché, qui crée une asymétrie structurelle entre opérateurs et collectivités - une « gentrification de l'électricité ». Le chapitre établit que le vide juridique aggrave les autres problèmes et que la réversibilité (que devient le site si le projet s'arrête ?) doit être un critère d'instruction dès l'entrée.

Une fois la puissance obtenue, le data center devient une infrastructure située. Il mobilise du foncier, interagit avec l'eau, la chaleur, les groupes de secours, les voiries, les compétences locales et les stratégies d'aménagement.

Ce chapitre examine une question : comment transformer la demande d'implantation en décision territoriale instruite, conditionnée et négociée ?

3.1 Une demande hypercentralisée : logiques de concentration et boucles de rétroaction

La demande de raccordement des data centers était fortement concentrée sur certaines zones, notamment le Sud de l'Île-de-France, les Hauts-de-France et la région PACA, autour de Marseille et de Fos-sur-Mer. Les grandes façades maritimes concentrent également d'importantes demandes de raccordement, mais celles-ci doivent être lues tous consommateurs confondus, notamment au regard des projets d'électrification industrielle et portuaire²⁸. Elles ne constituent donc pas automatiquement des pôles de demande de data centers. Les retours recueillis auprès d'Enedis font notamment état de très peu, voire d'aucune demande de data centers sur les ports du Havre et de Saint-Nazaire.

Cette concentration résulte de la conjonction de trois facteurs : la présence de nœuds de fibre optique robustes et de points d'atterrissage de câbles sous-marins, la disponibilité de capacités électriques existantes ou raccordables et la présence d'une main-d'œuvre qualifiée et d'écosystèmes numériques établis. De plus, ces trois conditions peuvent se renforcer mutuellement : la présence d'un premier data center attire les interconnexions, les fournisseurs, les compétences et les raccordements, ce qui à son tour attire d'autres data centers.

« Sur les data centers il y a un enjeu lié à la tendance à la concentration. Les tendances à s'agglomérer près des centres urbains et entre eux créent une codépendance aux infrastructures réseau et data », entretien avec un acteur territorial.

« Les centres urbains ont une latence faible. Memphis et Columbus, aux États-Unis, sont des villes denses et de taille importante. Malgré la diminution des exigences de latence, on assiste à une poursuite de la proximité des villes. En Europe, la saturation du réseau est encore plus importante. En théorie, on pourrait avoir des data centers dans des espaces plus éloignés, mais on observe encore une proximité avec les villes. », explique un expert des projections.

Sans orientation active, la pression du marché reproduit en Europe les concentrations américaines, avec moins de foncier disponible, moins de capacité réseau et plus de conflits d'usage. La

²⁸ Plus récemment, RTE indique que les deux territoires concentrant la majorité des projets raccordés au réseau de transport sont l'Île-de-France, environ 7 GW réservés, et les Hauts-de-France, environ 6 GW. Voir <https://www.rte-france.com/bases-electricite/consommation-electricite/essor-data-centers-france>

planification territoriale est indispensable pour que l'attractivité produise une valeur durable plutôt qu'une saturation locale.

Les zones FLAPD déjà mentionnées - Francfort, Londres, Amsterdam, Paris, Dublin - ont connu des restrictions ou des moratoires temporaires, par exemple aux Pays-Bas, poussant une partie de la demande vers des marchés secondaires : Marseille, Madrid, Milan, Varsovie. Mais ce déplacement ne fait que recomposer la logique de concentration autour de nouveaux nœuds de connectivité, d'énergie et de foncier. Il ne la dissout pas.

3.2 L'asymétrie structurelle et la « gentrification de l'électricité »

Tout d'abord, il faut poser une tension structurelle déjà mentionnée : l'asymétrie entre les opérateurs de data centers et les collectivités qui les accueillent.

L'asymétrie est simple : pour un opérateur mondial, un site est une option dans un portefeuille international ; pour une collectivité, c'est un engagement territorial à vingt ou trente ans. L'un peut arbitrer entre plusieurs pays, l'autre hérite du foncier mobilisé, des raccordements renforcés, des attentes politiques, des nuisances éventuelles et du risque d'irréversibilité.

L'actualité récente illustre cette asymétrie : à Étrechet²⁹, dans l'Indre, la maire, trois adjoints et une conseillère municipale ont démissionné en septembre 2026 dans le contexte d'un projet de data center porté par Google, estimé entre 880 MW et 1,1 GW. Les élus ont notamment exprimé un sentiment d'impuissance face à un projet **susceptible d'avoir des conséquences territoriales importantes**, mais sur lequel la commune estime disposer de peu de prise. Ce cas ne préjuge ni de la pertinence ni de l'issue du projet. Il montre en revanche que l'acceptabilité d'une infrastructure de cette échelle dépend aussi de la capacité des acteurs locaux à comprendre, instruire et peser réellement sur les arbitrages.

Asymétrie préjudiciable aux territoires moins armés

Cette asymétrie ne joue pas de la même manière selon les territoires. Les grandes métropoles dotées d'ingénierie, de services économiques structurés, de capacités juridiques et d'une expérience de négociation ne sont pas dans la même position que des territoires en recherche d'investissement, disposant de foncier disponible mais de moins d'expertise technique. Le risque est alors que les projets les plus difficiles à instruire - consommation élevée, retombées faibles, opacité des clients, faibles contreparties, pression foncière moindre - se déplacent vers les territoires les moins armés pour négocier.

La doctrine publique pourrait donc intégrer une **exigence d'équité territoriale** : ne pas laisser la charge environnementale, foncière et énergétique des data centers se reporter vers les territoires qui ont le plus besoin d'investissements mais le moins de moyens pour en apprécier les coûts complets.

L'ingénierie publique indépendante est une condition de justice territoriale.

En effet, cette asymétrie impose aux collectivités de ne pas « négocier au mieux » projet par projet, mais de maîtriser le périmètre et les termes du débat en amont, via leurs documents de planification, leurs critères d'accueil et leurs outils de négociation. Les collectivités qui auront anticipé seront en position de force, alors que celles qui réagissent au cas par cas risquent de subir des décisions prises ailleurs.

²⁹ https://www.lemonde.fr/economie/article/2026/09/16/dans-l-indre-la-maire-et-les-adjoints-d-un-village-demissionnent-pour-contester-le-projet-de-data-center-de-google_6775245_3234.html

Depuis le Sommet de l'IA en février 2025 et la mise en avant de l'attractivité française, plusieurs acteurs territoriaux décrivent un discours d'accélération très puissant, parfois vécu comme une injonction à accueillir.

« Depuis le Sommet de l'IA et la mise en place d'une task force par Business France, on fait face à un rouleau compresseur, des discours technophiles, enthousiastes, y compris pour changer la loi. Il y a la nécessité d'équilibrer ce discours positif sur les data centers : accueillir oui, mais il faut maîtriser », entretien avec un acteur territorial.

« Le risque est connu : reproduire le scénario des entrepôts logistiques - emprise forte, emplois limités, fiscalité décevante, dépendance à des acteurs mobiles », poursuit-il. La comparaison ne vise pas à assimiler les data centers à la logistique, mais à rappeler qu'un territoire qui n'a pas défini ses critères finit par accueillir ceux des autres.

« Tous les territoires n'ont pas d'ingénierie pour intégrer des sujets liés aux réseaux de chaleur, à monter en hauteur... », poursuit-il encore.

La gentrification de l'électricité³⁰

Un expert interviewé rapproche cette dynamique de ce qu'il nomme une « gentrification de l'électricité » : « Une chercheuse suédoise a nommé ce phénomène la « gentrification de l'électricité » : le déplacement d'usages industriels, productifs ou sociaux par des infrastructures numériques plus capitalistiques mais moins créatrices d'emplois locaux, dans les zones où la puissance électrique est attractive ». Cette formule décrit une dynamique de marché : là où l'électricité raccordable devient une ressource rare, les usages les plus capitalistiques tendent à la capter en priorité, indépendamment de leur valeur sociale ou industrielle locale.

Nous proposons de nous inspirer des outils déjà développés pour traiter la « gentrification urbaine » (zonage différencié, planification anticipée, conditionnalité d'accueil), pour renforcer une doctrine de raccordement qui ne se limite pas à un ordre de passage technique.

Trois mécanismes de la « gentrification urbaine » appliqués à l'électricité

Mécanisme classique	Transposition électrique	Référence rapport
Éviction par les prix	Éviction par la file de raccordement	Chapitre 2
Concentration de la demande solvable	Concentration de la demande électro-intensive	Chapitre 2
Externalités pour les habitants en place	Externalités pour les usages déjà présents sur le territoire	Chapitre 3

Traiter ce sujet de la « gentrification de l'électricité » signifie que l'accès à une ressource rare - la puissance raccordable - ne peut pas être abandonné au seul jeu de la solvabilité privée. Comme pour le foncier urbain, le rôle de la puissance publique est de définir des règles d'allocation, de préserver certains usages essentiels et d'éviter que les territoires les moins armés supportent les projets les plus difficiles à instruire.

³⁰ Le concept de gentrification énergétique a été développé dans l'article « **Data-center infrastructure and energy gentrification : perspectives from Sweden** » publié en 2021 dans *Sustainability : Science, Practice and Policy*. Les auteurs sont **Julia Velkova, Frans Libertson et Jenny Palm**
<https://portal.research.lu.se/en/publications/data-center-infrastructure-and-energy-gentrification-perspectives/>

3.3 Trois approches territoriales

Les entretiens font ressortir trois manières pour les territoires de traiter cette demande : la suivre, l'orienter ou l'expérimenter. Ces approches décrivent des trajectoires, des outils et des résultats différents selon le niveau de préparation des acteurs publics locaux.

1 - Suivre : la concentration peu coordonnée

La première approche est celle d'une juxtaposition de projets sans gouvernance d'ensemble. Des data centers s'implantent dans un même bassin territorial, chacun selon sa logique propre - foncier disponible, raccordement obtenu, autorisations accordées - avec peu de coordination entre eux et avec les pouvoirs publics sur les ressources partagées.

Les entretiens font de Paris-Saclay un exemple de concentration dont la coordination territoriale reste perfectible. Plusieurs acteurs y signalent un risque d'externalités cumulées, notamment autour des groupes électrogènes de secours en cas de déclenchement simultané. Ce point doit être compris comme un signal d'alerte issu du terrain, non comme une évaluation environnementale exhaustive du territoire. Il rejoint néanmoins des alertes formulées par Valérie Masson-Delmotte³¹ : absence de vision agrégée pour les collectivités, manque de données consolidées sur les générateurs de secours, tests périodiques pouvant générer des émissions locales, risques en cas de démarrage simultané, chaleur dissipée et difficulté à évaluer les impacts cumulés dans des conditions météorologiques défavorables.

« Paris-Saclay, c'est un hub européen d'IA. Plusieurs projets juxtaposés sans volonté de l'État de les organiser. C'est RTE qui fait la planification. La communauté d'agglomération a relevé des sujets de pollution liés aux générateurs de secours : s'il y avait une grosse coupure de courant et que tous les data centers enclenchaient leurs générateurs de secours, il y aurait un nuage de pollution catastrophique. », entretien avec un acteur territorial³².

Ce risque rappelle que des externalités individuellement gérables peuvent devenir problématiques à l'échelle d'un cluster. Un groupe électrogène testé isolément relève d'un sujet d'exploitation. Plusieurs groupes électrogènes testés ou déclenchés dans le même bassin d'air, au même moment ou dans des conditions météorologiques défavorables, peuvent devenir un sujet de santé publique, d'information des riverains et de coordination territoriale. La résilience privée ne doit donc pas créer une vulnérabilité collective.

La question des groupes électrogènes doit être instruite comme une externalité territoriale à part entière. Les data centers recourent généralement à une combinaison d'onduleurs, de batteries et de groupes électrogènes diesel, gaz ou carburants alternatifs selon les configurations, pour garantir la continuité de service. Ces équipements sont rarement utilisés en fonctionnement normal, mais doivent être testés et maintenus, et peuvent émettre des polluants atmosphériques lors des essais ou des déclenchements. L'enjeu est donc leur cumul à l'échelle du bassin : nombre de groupes, puissance installée, carburant utilisé, fréquence des tests, horaires, émissions de NOx et particules, bruit, information des riverains et scénario de déclenchement simultané en cas de coupure ou d'effacement.

³¹ Article de Valérie Masson-Delmotte <https://www.linkedin.com/pulse/les-enjeux-environnementaux-des-data-centers-val%C3%A9rie-masson-delmotte-zlxfe/>

³² Voir aussi l'article de Valérie Masson Delmotte mentionné ci-dessus.

Le Guide du Data Center Durable et Acceptable³³ recommande d'anticiper ces risques dès l'instruction : documenter les groupes électrogènes de secours, leurs conditions de test et de fonctionnement, évaluer les scénarios dégradés et prendre en compte les effets cumulés avec les autres data centers existants ou projetés à l'échelle du territoire. Il souligne également la nécessité d'anticiper les nuisances sonores et les émissions atmosphériques associées aux tests et aux situations de secours. La coordination des horaires de test, l'information des autorités locales et, lorsque cela est pertinent, des règles de fonctionnement coordonnées peuvent prolonger cette logique à l'échelle territoriale. Une attention particulière doit aussi être portée aux solutions alternatives aux groupes électrogènes conventionnels.

Exemples de risques systémiques de la concentration non coordonnée

Sécurité environnementale : déclenchement simultané des groupes de secours en cas de coupure électrique générale entraînant une pollution locale importante

Saturation des ressources partagées : postes sources, eau, foncier, voirie de chantier, avec peu de mutualisation et d'arbitrage et risques de conflits d'usage.

Irréversibilité : une fois les terrains occupés et les raccordements accordés, les arbitrages sont très difficiles à rouvrir

2 - Orienter : la délibération maîtrisée

La deuxième approche est celle d'une gouvernance territoriale construite en amont, avec des critères d'accueil définis avant l'arrivée des projets. Marseille est l'exemple le plus cité dans les entretiens.

La Métropole d'Aix-Marseille-Provence dispose d'atouts majeurs : deux sites d'atterrissage de câbles sous-marins - plage du Prado et Digue du Large - qui font de Marseille le deuxième hub internet d'échange en France après Paris, irrigant la Méditerranée et le Moyen-Orient. Cette position géographique attire une pression de projets très forte. La réponse de la Métropole a été d'organiser une délibération explicite plutôt que de répondre projet par projet.

La connectivité est un facteur d'attractivité, mais devient aussi un enjeu de sécurité. Les points d'atterrissage de câbles, les hubs d'échange et les clusters de data centers constituent des actifs critiques dont la protection doit être pensée avec l'énergie, le foncier, la continuité de service et la souveraineté numérique.

« Il y a 3 ans, le préfet avait demandé à travailler sur l'orientation des data centers. Le choix a été fait de modérer le développement des data centers au nord de Marseille et de monter plus haut vers un cluster avec de la capacité électrique, proche d'un nœud de 400 kV, à proximité d'une autoroute du numérique, avec de l'eau et une zone commerciale proche pour valoriser la chaleur. », entretien avec une organisation énergétique.

La démarche marseillaise repose sur plusieurs éléments : une délibération du Conseil métropolitain, une gouvernance multi-acteurs associant État, RTE, Enedis et les réseaux, l'identification d'un cluster dédié plus au nord entre Aix et Marseille, proche d'un nœud 400 kV, avec du foncier disponible, de l'eau et des possibilités de valorisation de la chaleur fatale, ainsi que des outils de géomatique pour cartographier les projets et anticiper les infrastructures associées. Voici quelques éléments caractéristiques :

³³ <https://www.ville-demain.com/guide-du-data-center-durable-et-acceptable-partie-2-limpact-environnemental/>

Élément	Ce que la démarche marseillaise produit
Délibération métropolitaine	Doctrine publique formalisée : où, dans quelles conditions, avec quelles contreparties ?
Gouvernance multi-acteurs	Cohérence entre État, RTE, Enedis, Métropole pour éviter les contradictions entre niveaux
Cluster dédié identifié	Pression orientée vers les zones à moindre conflit d'usage, avec les ressources nécessaires
Simulateur fiscal	Objectivation des retombées réelles selon la qualification applicable (neutralise l'opacité)
Outils de géomatique	Cartographie des projets, anticipation des infrastructures, visibilité sur la concentration locale

Ce modèle n'est pas transposable mécaniquement partout. Il suppose a minima une métropole disposant d'une ingénierie territoriale suffisante, d'une position géographique qui justifie la demande et d'acteurs qui acceptent d'organiser la concertation. Mais il montre qu'une doctrine d'accueil est possible et qu'elle produit de meilleures décisions que la réaction projet par projet.

3 - Expérimenter : l'expérience métropolitaine

La troisième approche consiste à déplacer la question : ne plus seulement accepter ou refuser des data centers privés, mais imaginer des capacités de calcul orientées par le territoire, ouvertes à ses usages stratégiques et gouvernées comme des infrastructures d'intérêt collectif.

Des acteurs interviewés défendent ce modèle : des data centers à dimension métropolitaine, coproduits avec les territoires et portés par des acteurs européens, où la puissance de calcul serait traitée comme un bien commun territorial plutôt que comme un actif purement privé.

« Montpellier, Lyon réfléchissaient à la mise en commun d'un data center au niveau de la Métropole qui pourrait être ouvert aux acteurs publics et aux acteurs privés. À Barcelone, les data centers sont dans l'Université qui propose des capacités de calcul et accueille des acteurs privés, des laboratoires pharmaceutiques... La nuit est utilisée par les entreprises aéronautiques très consommatrices de calcul. Ça nous paraît intéressant. On défend un modèle qui sert à alerter le territoire et à garder la puissance de calcul », entretien avec un acteur territorial.

La puissance de calcul peut-elle devenir un actif territorial, plutôt qu'une simple infrastructure privée autorisée localement ?

Ces initiatives restent embryonnaires. Elles se heurtent à des obstacles réels : financement, compétences techniques, modèle économique, gouvernance, accès aux GPU. Mais elles tracent une direction : plutôt qu'un choix d'opérateur privé que la collectivité valide ou refuse, **le data center peut devenir un actif territorial que la collectivité contribue à orienter, financer et gouverner.**

Quelques conditions du modèle d'expérimentation métropolitaine

Ingénierie territoriale : des compétences spécifiques (fluides (énergie, eau, chaleur), numérique, foncier, droit) que beaucoup de collectivités ne possèdent pas encore.

Modèle économique : un data center sous maîtrise publique doit être viable.

Gouvernance claire : qui décide de l'accès, des priorités, des contreparties, du prix ?

Ancrage dans l'écosystème local : universités, hôpitaux, industries, PME..., les usages locaux identifiés doivent être servis par la puissance de calcul.

3.4 Foncier, eau, chaleur : tests matériels d'implantation

Au-delà de l'énergie, trois ressources locales doivent devenir des critères d'implantation à part entière : le foncier, l'eau et la chaleur. Elles ne peuvent pas être découvertes après coup, au moment de l'autorisation ou de la négociation. Elles doivent être instruites dès l'amont, car elles déterminent la soutenabilité territoriale du projet.

Le présent chapitre se concentre sur ce que le territoire doit instruire directement avant l'implantation : foncier, eau, chaleur, secours, bruit et réversibilité.

La chaîne matérielle mondiale - GPU, mémoires, câbles, transformateurs, métaux critiques, systèmes électriques et équipements de refroidissement - sera traitée plus loin comme enjeu de souveraineté, d'empreinte et de dépendance industrielle.

Le foncier : un actif stratégique

Les entretiens montrent que la valeur d'un terrain pour un data center ne dépend pas seulement de sa surface ou de sa localisation, mais aussi de sa combinaison avec une puissance électrique disponible, une connectivité fibre, une préparation administrative, une accessibilité chantier, une compatibilité réglementaire et une acceptabilité locale. C'est ce qui transforme certains terrains en «*powered land*» : un foncier énergisé, raccordable, préparé, qui peut atteindre des valorisations considérables, créant une concurrence entre projets industriels, opérateurs de long terme et stratégies plus spéculatives.

Les entretiens signalent une financiarisation croissante du foncier destiné aux data centers. Certains acteurs ne se positionneraient **pas d'abord comme opérateurs de long terme, mais comme assembleurs d'un actif rare** : foncier disponible, puissance électrique sécurisée, connectivité, autorisations et préparation administrative. Cette logique de «*powered land*» transforme le terrain raccordable en produit stratégique, parfois destiné à être cédé ensuite à un opérateur ou à un hyperscaler. Elle renforce l'enjeu d'instruction amont : la collectivité ne négocie pas de la même manière avec un opérateur qui exploitera durablement le site et avec un acteur qui valorise surtout une option foncière et électrique.

Le guide France Urbaine / Ville de Demain pose aussi la question du coût d'opportunité : un data center mobilise un foncier stratégique qui aurait pu accueillir du logement, des activités productives, un équipement public. Ce coût d'opportunité doit être objectivé dans l'instruction des projets.

Le Guide du Data Center Durable et Acceptable documente également la transformation de bâtiments existants. Il souligne notamment la contrainte de portance structurelle et cite plusieurs opérations de reconversion, dont Telehouse boulevard Voltaire, les anciennes usines Delage à

Courbevoie et Paris Digital Park à La Courneuve. Ces exemples montrent que la reconversion est possible, mais qu'elle suppose de vérifier la portance, la dépollution éventuelle, l'adaptation technique et la compatibilité avec d'autres usages productifs.

Construire neuf, transformer, modulariser : le data center comme dette matérielle

La question foncière engage aussi une forme constructive : construire neuf, transformer un bâtiment existant, réhabiliter une friche industrielle, modulariser l'infrastructure ou prévoir une seconde vie. Les entretiens avec des acteurs de l'architecture et de la filière signalent que les *data centers hyperscale* s'orientent de plus en plus vers des logiques modulaires, permettant une montée en charge progressive et un remplacement plus ciblé des composants obsolètes. Cette modularité peut limiter certains risques d'obsolescence, mais elle ne supprime pas la dette matérielle : béton, acier, équipements électriques, groupes de secours, refroidissement, câblage, sols artificialisés ou friches requalifiées.

L'ACV publiée par Data4 et APL³⁴ confirme que cette dette matérielle doit être instruite. Sur le data center DC19 de Marcoussis, représentatif d'un hyperscaler de 5 MW IT, le périmètre retenu comprend notamment le bâtiment, l'alimentation électrique, le refroidissement, le secours énergétique, la maintenance et la fin d'activité, mais exclut la fabrication et l'usage des serveurs informatiques. Même sur ce périmètre limité à l'infrastructure d'accueil, le site étudié mobilise 5 200 m³ de béton, 500 tonnes d'acier, plus de 30 km de câbles, des batteries, quatre générateurs diesel et vingt groupes froids. La matérialité du data center ne se réduit donc pas à son enveloppe immobilière : elle inclut une infrastructure technique lourde, renouvelable, maintenable et difficilement réversible.

« Les élus doivent comprendre le sujet, les data centers doivent prévoir une seconde vie. », entretien avec un acteur territorial. Un architecte interviewé alerte dans le même sens : « Les bâtiments construits ne sont pas réutilisables. Le béton est coulé de façon tellement spécifique qu'on ne peut rien en faire. Les friches, est-ce moins de problèmes ? La nouvelle vague de projets amène à raser les bâtiments de friches et à construire. Ces bâtiments semblent difficiles à recycler. Pas de bâtiment multifonctions, c'est un problème. »

Le *brownfield* constitue une piste prioritaire lorsqu'il permet de réutiliser un foncier déjà industrialisé, d'éviter une artificialisation supplémentaire et de s'appuyer sur des infrastructures existantes. Mais il ne doit pas être considéré comme une solution automatique : portance, dépollution, raccordement, refroidissement, hauteur sous plafond, sécurité, évolutivité et capacité de réversibilité doivent être instruits. Un data center est un actif matériel lourd, dont l'entrée, l'exploitation et la sortie doivent être pensées ensemble.

Cette logique conforte une piste stratégique déjà évoquée : utiliser certains projets de data centers pour requalifier des friches, dépolluer des sites déjà industrialisés et préparer des fonciers

³⁴ <https://www.data4group.com/ressources/data-center-mesurer-limpact-pour-mieux-agir/>

réversibles. Mais cette opportunité ne vaut que si elle est contractualisée dès l'entrée : dépollution, réversibilité, seconde vie du site et capacité à accueillir d'autres activités industrielles si le projet initial évolue ou s'arrête.

L'eau : critère d'autorisation, pas détail technique

Les entretiens font apparaître l'eau comme un angle mort ou presque. Peu d'acteurs mesurent sérieusement le WUE (*Water Usage Effectiveness*) et peu d'autorités locales semblent encore l'intégrer dans leurs critères d'évaluation. « *WUE on ne s'en occupe pas encore. C'est peut-être un angle mort* », confie un expert de l'énergie. Mais la question monte, notamment avec la densification GPU qui pousse vers le *Direct Liquid Cooling*.

Technologie de refroidissement	Consommation d'eau directe	Point d'attention
Refroidissement à air / free cooling	Très faible à nulle selon la configuration	Très favorable à l'eau ; performance énergétique dépendante du climat et de la température extérieure
Tour de refroidissement évaporative	Élevée	Bonne efficacité énergétique possible, mais consommation d'eau liée à l'évaporation
Refroidissement adiabatique	Variable, généralement saisonnière	Compromis entre consommation électrique et consommation d'eau ; fortement dépendant du climat
Direct Liquid Cooling (DLC)	Faible à potentiellement nulle dans la boucle IT	La consommation totale dépend surtout du système utilisé ensuite pour rejeter la chaleur
Immersion cooling	Faible à potentiellement nulle dans la boucle IT	Même réserve : la consommation du site dépend du système de rejet thermique associé

Sources : ADEME, *Refroidissement des datacenters - technologies utilisées en France, potentiel d'économies*, 2025 ; U.S. Department of Energy, *Cooling Water Efficiency Opportunities for Federal Data Centers*

La consommation d'eau des data centers varie fortement selon leur puissance, leur taux d'utilisation, la technologie de refroidissement, le climat et le système électrique associé. En France, l'ARCEP³⁵ estime à environ 6,5 millions de m³ le volume d'eau prélevé ou consommé par les data centers en 2024, soit l'équivalent de la consommation annuelle moyenne de plus de 100 000 personnes.

Les ordres de grandeur publiés montrent surtout l'hétérogénéité des situations. OVHcloud³⁶ publie un WUE de 0,34 L/kWh IT et met en avant une approche intégrée de réemploi et de prolongation de la durée de vie des serveurs. Data4³⁷, dans l'ACV du DC19, indique une consommation directe d'eau très faible pour le site étudié, mais montre que l'empreinte hydrique indirecte liée à la production d'électricité, aux matériaux et aux équipements reste significative.

³⁵ Source : ARCEP, *Generative AI: Environmental Issues and Challenges*, mai 2026 ; Barnett-Itzhaki, *Water Research*, 2026 https://www.arcep.fr/uploads/tx_gspublication/report-generative-AI-environmental-issues_may2026.pdf

³⁶ <https://corporate.ovhcloud.com/en/sustainability/environment/>

³⁷ <https://www.data4group.com/ressources/data-center-mesurer-limpact-pour-mieux-agir/>

Cette difficulté de mesure n'est pas propre à la France. Une étude de l'University of Texas at Austin³⁸ sur les besoins en eau des data centers au Texas montre l'ampleur de l'incertitude lorsque les capacités se développent rapidement sans que les données d'exploitation soient suffisamment consolidées. Selon les scénarios étudiés, les data centers pourraient représenter entre 3% et 9% de la consommation d'eau de l'État à l'horizon 2040. Ces résultats combinent la consommation directe des installations et l'empreinte hydrique indirecte liée à la production d'électricité ; ils dépendent fortement des hypothèses retenues en matière de capacité, de refroidissement, d'efficacité et de mix électrique. Leur intérêt est surtout méthodologique : faute de données réelles suffisamment accessibles, l'empreinte hydrique future doit être estimée par scénarios, avec des incertitudes importantes précisément au moment où les décisions d'infrastructure sont prises.

Cette différence est essentielle : un site peut avoir une faible consommation directe d'eau en exploitation et néanmoins présenter une empreinte hydrique significative en cycle de vie, notamment via l'électricité, l'acier, le béton, le cuivre, les équipements électriques et les batteries. À l'inverse, un WUE faible ne suffit pas à qualifier la soutenabilité d'un projet si le bassin hydrique est déjà sous tension, si la chaleur dissipée renforce les îlots de chaleur ou si le refroidissement implique des choix techniques peu réversibles.

L'indicateur AWARE, pour *Available WAtER REmaining*, développé dans le cadre des travaux du groupe WULCA³⁹ de la Life Cycle Initiative du PNUE et de la SETAC, peut contribuer à cette contextualisation. AWARE mesure le potentiel de privation d'eau pour d'autres usages humains ou écosystémiques, à partir de l'eau disponible restante dans un bassin après satisfaction des besoins humains et environnementaux. Il rappelle qu'un mètre cube d'eau consommé n'a pas la même signification selon le bassin hydrique concerné. En complétant le WUE par un indicateur de pression hydrique tel qu'AWARE, les collectivités peuvent passer d'un ratio technique déclaré à une instruction mieux contextualisée par la vulnérabilité locale de la ressource.

Le fait que la pression hydrique locale soit encore peu intégrée par une partie des opérateurs et des prescripteurs renforce l'urgence d'en faire un critère explicite d'instruction et d'autorisation.

L'enjeu est de demander : quel WUE réel en exploitation, quel type d'eau mobilisée, dans quel bassin hydrique, avec quelle pression locale, quels usages concurrents, quelle chaleur rejetée et quelle empreinte hydrique indirecte en cycle de vie ?

Au-delà de l'eau, de la chaleur et de l'énergie, certains acteurs de la filière signalent notamment une vigilance émergente autour des fluides frigorigènes et de certaines substances assimilables ou proches des PFAS. Ce point ne doit pas être surinterprété, mais rappelle que l'empreinte d'un data center inclut aussi une chimie de l'exploitation et de la maintenance.

La chaleur fatale : opportunité conditionnelle, pas argument d'acceptabilité automatique

« Un data center génère presque du 1 pour 1 : 100 MW génèrent 100 MW de chaleur. On n'a pas besoin de tout ça. Il faut récupérer ce qui est récupérable. La chaleur ne doit pas devenir l'argument automatique d'acceptabilité d'un projet, ni la source unique de chauffage d'un territoire. Une stratégie robuste doit prévoir une production de chaleur indépendante, une contractualisation

³⁸ University of Texas at Austin, Data Centers Are Growing in Texas, but Big Questions Remain About Water Use, 6 mai 2026
<https://news.utexas.edu/2026/05/06/data-centers-are-growing-in-texas-but-big-questions-remain-about-water-use/>
Bureau of Economic Geology, Water Requirements for Data Centers :

https://compass.beg.utexas.edu/files/publications/Water_Requirements_for_DC_White_Paper.pdf
Texas Tribune, 23 juin 2026 : <https://www.texastribune.org/2026/06/23/texas-data-centers-puc-water-survey/>

³⁹ <https://wulca-waterlca.org/aware/what-is-aware/> ; <https://wulca-waterlca.org/> - PNUE, SETAC, Life Cycle Initiative

claire, une contribution éventuelle du data center au réseau de chaleur et une analyse de ce qui se passe si le site ralentit, change d'usage ou s'arrête. », propos d'un architecte.

Les cas de valorisation documentés par France Urbaine / Ville de Demain (avril 2026) montrent que la récupération fonctionne quand elle est pensée dès l'implantation dans un montage multi-acteurs : Stockholm Data Parks (30 000 appartements alimentés, cadre contractuel standardisé), Genève Infomaniak DC4 (6 000 logements alimentés avec l'aide d'une pompe à chaleur pour passer de 40 °C à 67 °C), Saint-Denis Equinix (piscine olympique et 1 600 logements alimentés). Dans ces cas, la chaleur est une contrepartie négociée et garantie.

Le rapport du Sénat de septembre 2026 propose à ce titre la création d'un fonds de garantie spécifique à la chaleur fatale des data centers. Il vise à sécuriser les investissements des réseaux de chaleur face au décalage entre leurs engagements de long terme et le risque qu'un data center ferme, déménage ou modifie sa technologie. Cette proposition traduit en mécanisme financier un problème déjà identifié ici : la valorisation de la chaleur ne tient que si le risque de continuité est lui aussi traité.

La chaleur doit aussi être regardée sous son versant négatif. Un article⁴⁰ publié en preprint soumis en avril 2026 estime, à partir de mesures de température de surface par télédétection, une hausse moyenne de la température de surface (route goudronnée, toit métallique) d'environ 2 °C autour de data centers IA après leur mise en service. Ce résultat doit être traité, avec prudence, comme un signal scientifique émergent, à confirmer selon les contextes.

La doctrine à retenir est double : la chaleur récupérée peut devenir une contrepartie territoriale utile si elle est techniquement exploitable, contractualisée et compatible avec les besoins locaux. La chaleur dissipée est une externalité à traiter, notamment si elle s'ajoute aux îlots de chaleur urbaine.

Tout projet devrait donc documenter les deux faces : quelle chaleur est récupérée, à quelle température, pour quel usage, avec quelle garantie de continuité et quelle chaleur est dissipée, où, avec quels effets potentiels sur l'environnement immédiat.

3.5 Le vide réglementaire : aggravant structurel

Les critères territoriaux (foncier, eau, chaleur, groupes de secours, bruit, réversibilité) se heurtent tous au même obstacle : les data centers restent juridiquement mal stabilisés en droit français. Il n'existe pas encore de catégorie unique permettant de les instruire simultanément comme objets numériques, énergétiques, industriels, fonciers et environnementaux. Selon les cas, ils peuvent être rapprochés de catégories de bâtiment, d'entrepôt, d'activité industrielle, d'équipement technique ou d'infrastructure numérique, avec des effets différents en urbanisme, fiscalité, autorisations et exigences environnementales.

Encadré - Fiscalité locale : ce que le data center rapporte vraiment au territoire⁴¹

La fiscalité locale est souvent présentée comme l'une des contreparties positives de l'accueil d'un data center. Elle peut effectivement générer des recettes : taxe foncière sur les propriétés bâties, cotisation foncière des entreprises, cotisation sur la valeur ajoutée des entreprises lorsqu'elle

⁴⁰ Article publié en preprint soumis en mars 2026 et révisé en avril 2026 (« The data heat island effect: quantifying the impact of AI data centers in a warming world » <https://arxiv.org/abs/2603.20897>

⁴¹ Sources : <https://www.impots.gouv.fr/professionnel/cet-cfe-et-cvae> ; https://www.legifrance.gouv.fr/ceta/id/CETATEXT000046420744?init=true&page=1&query=463134&searchField=ALL&tab_selection=all ; <https://www.ville-demain.com/wp-content/uploads/2026/04/guide-data-center-retombees-socio-economiques-fiscalite-acceptabilite-2026.pdf> ;

s'applique, taxes d'urbanisme, voire recettes associées à certains équipements ou consommations énergétiques. Mais cette fiscalité dépend de la qualification juridique du site, de la valeur locative retenue, du régime d'imposition applicable, de la structure de détention, des exonérations éventuelles, des coûts publics induits et du foncier mobilisé.

Le data center n'entre pleinement dans aucune catégorie fiscale. Il est à la fois un actif immobilier, une infrastructure technique, un équipement numérique, un consommateur électro-intensif, parfois un site industriel, parfois un objet tertiaire par défaut. Cette hybridité a des effets concrets. La catégorie choisie conditionne ce que le territoire peut taxer, comparer, exiger, exonérer ou négocier.

Le contentieux relatif aux locaux de stockage en Île-de-France illustre cette fragilité des assimilations. Le Conseil d'État a confirmé en 2022 que les data centers ne relevaient pas des locaux de stockage au sens de la taxe pour création de locaux de stockage en Île-de-France. Cette position est reprise par la doctrine fiscale.

« Au regard du code d'urbanisme, les data centers n'existent pas. Les data centers n'ont pas de définition dans les PLU. Ça peut créer des ambiguïtés et des tensions, car ça dépend du territoire, du PLU... Il faut clarifier l'objet juridique. Les DRFIP ne sont pas toutes d'accord entre elles selon les territoires. Les conséquences ne sont pas les mêmes d'un point de vue fiscal », confie un acteur territorial.

La fiscalité énergétique ajoute une couche supplémentaire. Les data centers sont des consommateurs massifs d'électricité ; les taxes et accises sur l'électricité peuvent donc peser dans le modèle économique. Mais ce niveau de fiscalité peut lui aussi varier selon les régimes applicables, les seuils, les usages, les conditions d'efficacité énergétique ou d'éco-responsabilité. Là encore, l'enjeu est notamment de savoir ce qui revient effectivement au territoire qui accueille l'infrastructure.

L'exemple des métriques tertiaires illustre cette difficulté. Un data center ne se pilote pas comme un immeuble de bureaux : raisonner en kWh/m²/an peut être utile pour certains bâtiments, mais devient très insuffisant pour un site dont l'empreinte dépend d'abord de la puissance IT installée, du taux d'occupation, du profil de charge, du refroidissement, de la redondance, du PUE, du WUE et des usages hébergés.

Le point clé est donc de passer d'une logique de recette affichée à une logique de bilan territorial net. Un data center peut produire des recettes fiscales visibles tout en mobilisant du foncier stratégique, des capacités électriques rares, des voiries, des études, de l'ingénierie publique, des adaptations de documents d'urbanisme, voire des investissements dans des réseaux de chaleur ou des infrastructures associées. Il peut aussi bénéficier d'exonérations, de régimes favorables ou d'une qualification fiscale moins contributive que celle initialement anticipée.

La question à poser est donc : « que laisse-t-il réellement au territoire, une fois déduits les coûts directs, les coûts d'opportunité et les risques de long terme ? » Cette analyse devrait intégrer les recettes fiscales attendues, les exonérations possibles, les coûts publics induits, les investissements nécessaires, la valeur alternative du foncier, les emplois réellement créés, les contreparties négociées, la chaleur valorisée, les contraintes environnementales et les conditions de sortie.

C'est précisément l'intérêt des simulateurs fiscaux territoriaux développés par certaines collectivités : ils permettent de transformer une promesse d'attractivité en hypothèse vérifiable,

comparable et négociable. Sans cet outillage, la fiscalité peut devenir un argument d'acceptabilité trop rapide. Avec lui, elle redevient un élément parmi d'autres dans une négociation territoriale complète.

Les PINM (projets d'intérêt national majeur)

Les débats parlementaires autour de la simplification économique et de l'encadrement de l'implantation des data centers ont introduit la possibilité de qualifier certains très grands projets de data centers comme projets d'intérêt national majeur - PINM. Afin d'accélérer les procédures, ce statut pourrait renforcer la capacité de l'État à prendre la main sur des sujets relevant habituellement des collectivités : urbanisme, compatibilité des documents locaux, rythme d'avancement, contentieux. Les projets de data centers qualifiés PINM⁴² pourraient également bénéficier, selon le texte finalement adopté et ses conditions d'application, de certains effets attachés à ce statut, notamment en matière de comptabilisation foncière. Ce point doit rester formulé avec prudence, car il dépend du texte final, de ses décrets d'application et de son articulation avec les règles ZAN (Zéro artificialisation nette).

Ce statut risque de court-circuiter les collectivités territoriales, au moment même où elles cherchent à renforcer leur pouvoir de négociation.

L'exemple du Campus IA de Fouju montre que l'accélération des procédures ne fait pas disparaître les tensions territoriales et peut les déplacer vers le contentieux. Ce projet, porté par MGX, Bpifrance, Nvidia et Mistral AI, vise à terme une puissance de 1,4 GW, avec un raccordement progressif prévoyant un premier palier de 240 MW fin 2027, puis 700 MW⁴³ avant fin 2029. Après l'autorisation préfectorale⁴⁴ accordée le 29 juillet 2026 et l'annonce du démarrage des premiers travaux, plusieurs associations ont annoncé leur intention de contester le projet devant la juridiction administrative, notamment le permis de construire. Sans préjuger du dépôt effectif, de la recevabilité, de l'issue ou des effets de ces recours sur le calendrier, cette séquence montre qu'une autorisation administrative ne clôt pas nécessairement l'instruction territoriale d'un projet. Elle renforce la nécessité d'associer à l'accélération des procédures des capacités publiques d'expertise, une information vérifiable et des contreparties territoriales négociées en amont.

3.6 Le coût de sortie : instruire l'après dès l'entrée

Un data center ne doit pas être instruit seulement par son entrée - investissement, puissance, foncier, emplois, fiscalité. Il doit l'être aussi par sa réversibilité. Que restera-t-il si l'opérateur part, si les GPU deviennent obsolètes, si le modèle économique change ou si la demande se déplace ?

La hausse du coût des équipements peut inciter à prolonger leur usage, mais la seconde vie reste peu documentée : réemploi local, revente, transfert vers d'autres pays, recyclage ou mise au rebut. Sans traçabilité de ces flux, une partie de l'empreinte matérielle peut être déplacée hors du site, voire hors du territoire, sans traçabilité et sans disparaître.

⁴²Sources : <https://www.senat.fr/leg/tas25-073.html> ; <https://www.economie.gouv.fr/actualites/loi-de-simplification-de-la-vie-economique-des-mesures-concretes-pour-faciliter-lactivite-des-entreprises> ; <https://www.senat.fr/compte-rendu-commissions/20260119/cmp.html>

⁴³ <https://www.rte-france.com/actualites/campus-ia-rte-signent-contrat-raccordement-accelere-campus-infrastructures-numeriques>

⁴⁴ <https://www.lesechos.fr/tech-medias/hightech/la-contestation-se-durcit-contre-le-projet-de-mega-data-center-de-mgx-et-mistral-en-seine-et-marne-2253320> ; <https://www.leparisien.fr/seine-et-marne-77/data-centers-plus-cher-que-le-grand-paris-express-le-projet-de-campus-ia-geant-valide-en-seine-et-marne-11-08-2026-AUCLSXM7CFHOZFPLDUHJSPRUI.php>

La question est donc aussi patrimoniale, industrielle et territoriale : **qui porte le risque d'un actif lourd lorsque la valeur numérique qui l'a justifié s'est déplacée ?**

Cette question est centrale parce que **les horizons temporels ne coïncident pas**. Les serveurs et GPU peuvent être renouvelés en quelques années ; les bâtiments et équipements de refroidissement s'inscrivent dans des horizons de dix à vingt-cinq ans ; les lignes électriques et postes sources peuvent durer plusieurs décennies. Cette dissonance crée un risque de transfert structurel : la valeur numérique se déplace vite, mais les territoires héritent d'infrastructures longues - foncier mobilisé, raccordements renforcés, voirie adaptée, réseaux de chaleur... éventuellement dimensionnés autour du site.

La réversibilité doit donc être contractualisée dès l'entrée : plus les horizons techniques sont dissociés, moins elle peut être renvoyée à une négociation tardive.

Cette exigence de réversibilité doit être pensée à deux niveaux :

1. le premier est immobilier et territorial : que devient le bâtiment si l'opérateur se retire, si l'usage évolue ou si le site n'est plus compétitif ?
2. le second est technologique : que deviennent les modules, serveurs, GPU, systèmes de refroidissement, batteries, groupes de secours et équipements électriques lorsque la génération suivante arrive ?

Un angle mort spécifique mérite d'être signalé : le retrofit des data centers vieillissants. Des centaines de sites en France ont aujourd'hui entre 15 et 20 ans. Leurs équipements sont vieillissants, leurs performances énergétiques faibles et leur capacité à accueillir des racks haute densité limitée. À notre connaissance, il n'existe pas à ce jour d'étude systématique à l'échelle française sur ce sujet, ni de cadre public spécifiquement consacré à la modernisation des data centers vieillissants ou distinguant le data center à rénover du data center à remplacer. *« Il y a des data centers qui ont 15-20 ans et qu'il faut accompagner. Souvent vétustes, des sujets de performance énergétique. Il n'y a pas d'études sur ça. »*, entretien avec un syndicat de l'énergie. **La réversibilité ne peut donc pas se réduire à la question de la fin de vie : elle pourrait aussi inclure les conditions de la vie longue, comment un site peut évoluer, se densifier, se moderniser sans être rasé et reconstruit.**

L'enjeu est d'autant plus important pour les data centers IA que l'obsolescence peut venir aussi des GPU, des densités de calcul, des architectures de refroidissement, des modèles économiques de l'inférence, du déplacement rapide de la demande vers d'autres géographies. Le coût de sortie doit donc être pensé comme un coût territorial, mais également comme un coût d'obsolescence technologique.

Un plan de seconde vie apparaît nécessaire dès l'instruction : scénarios de réemploi, modularité, démontabilité, reconversion possible, provision financière, responsabilités en cas d'arrêt et conditions de restitution du foncier ou des capacités raccordées. **Sans cette clause de sortie, le territoire assume une partie du risque d'obsolescence d'un actif qu'il ne contrôle pas.**

3.7 Ce que ce chapitre implique pour la doctrine

Les facteurs évoqués dans ce chapitre sont des positions de départ qui peuvent évoluer selon les outils dont se dotent les collectivités et les règles que fixe l'État. Plusieurs enseignements se dégagent.

Enseignement	Implication pour la doctrine
La concentration est une logique de marché, pas une fatalité	Une planification nationale d'implantation identifiant les zones à développer, les zones à protéger, les critères de maillage peut orienter sans interdire
L'asymétrie opérateur-territoire est structurelle	Construire et préserver l'ingénierie territoriale des collectivités ainsi que la capacité effective des acteurs locaux à instruire, négocier et réexaminer les conditions du projet est une condition sine qua non d'une gouvernance équilibrée.
Les ressources locales, des critères d'implantation	Foncier, eau, chaleur, groupes de secours, bruit : à contractualiser avant l'autorisation
Le vide juridique aggrave les autres problèmes	Définir l'objet juridique « data center » est le préalable à toute doctrine cohérente : fiscalité, urbanisme, environnement, transparence
La transparence, une condition de la négociation	Les collectivités ne peuvent pas exercer leur pouvoir de négociation sans savoir qui investit, quels clients héberger, quelles données traiter
La réversibilité, un critère d'instruction	Exiger un plan de fin d'exploitation dès l'instruction : que deviennent le bâtiment, les raccordements, les réseaux de chaleur, le foncier si le projet ralentit ou s'arrête ? Les serveurs se renouvellent en quelques années ; les bâtiments durent 10-25 ans ; les lignes électriques plusieurs décennies. Cette dissonance des horizons temporels transfère un risque structurel vers les territoires si elle n'est pas anticipée contractuellement.

La géographie révèle les tensions territoriales. Elle ne dit pas encore qui finance, qui opère, qui contrôle, quels usages seront hébergés et où ira la valeur créée. C'est le sujet du chapitre suivant : un data center situé en France n'est pas automatiquement un actif stratégique français.

Chapitre 4 - Localiser ne suffit pas

Le chapitre en une phrase

Un data center situé en France n'est pas automatiquement un actif stratégique français : localisation, souveraineté et valeur captée peuvent être totalement dissociées. Le chapitre identifie les conditions pour qu'un site devienne un actif collectif plutôt qu'un passif - transparence sur les investisseurs et les usages, contreparties contractualisées, ancrage local, commande publique souveraine.

Localiser n'est pas maîtriser. Raccorder n'est pas capter. Héberger n'est pas savoir.

C'est l'objet de ce chapitre : distinguer la localisation physique de la maîtrise stratégique. Un data center peut être situé en France tout en dépendant de capitaux, de modèles, de clients, de logiciels, de câbles et de décisions situés ailleurs. Une consommation ici pour peut-être un travail rendu ailleurs.

Le data center **localise les ressources mobilisées** - électricité, eau, foncier, raccordement, chaleur, nuisances, mais **pas nécessairement la valeur produite**. Celle-ci peut être captée ailleurs : par les modèles, clients, plateformes, contrats, marges ou décisions situés hors du territoire.

4.1 La localisation sans maîtrise : le paradoxe du bâtiment supposé souverain

Un data center localisé en France relève bien du droit français pour l'urbanisme, l'environnement, la sécurité et la fiscalité locale. Mais cette localisation ne suffit pas à établir une maîtrise stratégique.

La maîtrise se joue dans une chaîne plus large : qui finance le site, qui l'opère, quels clients l'utilisent, quels modèles y tournent, quelles données y transitent, quelle juridiction s'applique aux contrats, qui capte les marges et qui peut interrompre ou déplacer le service. C'est cette chaîne - technique, contractuelle, juridique et géopolitique - qui détermine le niveau réel de souveraineté.

Exemples de ce que la localisation ne garantit pas

Le modèle d'IA : les grands modèles de fondation sont américains (OpenAI, Google, Anthropic, Meta) ou chinois (DeepSeek, Baidu). L'Europe compte peu d'acteurs de modèles de fondation de rang mondial. Mistral est aujourd'hui l'un des plus visibles et des plus structurants. Un data center en France peut héberger de l'inférence sur des modèles dont la conception, l'entraînement et la mise à jour sont hors de portée européenne.

Les GPU et semi-conducteurs avancés : les GPU reposent sur une chaîne très concentrée. La quasi-totalité des puces IA de dernière génération dépend de TSMC à Taïwan pour la fabrication avancée. Les machines de lithographie EUV (Extreme Ultraviolet Lithography, technologie permettant de graver des transistors extrêmement fins sur les puces les plus avancées) sont produites par ASML aux Pays-Bas. Les mémoires HBM (High Bandwidth Memory, mémoires très rapides placées au plus près des accélérateurs IA pour alimenter les GPU en données) sont concentrées chez SK Hynix et Samsung. La chaîne comporte donc un maillon européen critique avec ASML, mais l'Europe ne maîtrise pas l'ensemble : fabrication avancée, packaging, HBM, GPU, cloud et modèles restent largement dépendants d'acteurs extra-européens.

Les capitaux : selon les entretiens, une partie des projets annoncés ou instruits repose sur des

fonds, promoteurs ou acteurs extra-européens. Comme évoqué au chapitre 3, la financiarisation des data centers (promoteurs immobiliers, fonds anglo-saxons) peut déplacer la question de la valeur captée loin du territoire d'implantation.

Les câbles sous-marins : la souveraineté ne se joue pas seulement dans les bâtiments et les puces, mais aussi dans les routes physiques des données. Les grandes plateformes numériques ont pris une place croissante dans le financement, la propriété ou l'usage capacitaire des câbles sous-marins. Un data center situé en France peut donc dépendre de flux, de routes, de points d'atterrissage, de contrats de capacité et d'acteurs privés mondiaux dont la maîtrise échappe en partie aux territoires d'accueil.

La juridiction : un data center financé, opéré ou contractuellement dépendant d'un acteur américain peut exposer certaines données, opérations ou accès à des risques juridiques extraterritoriaux, selon la structure capitaliste, contractuelle et opérationnelle de l'offre. La localisation physique en France ne suffit donc pas, à elle seule, à qualifier le niveau de souveraineté.

Plusieurs entretiens expriment cette inquiétude. « *Microsoft promet des produits souverains, mais avec le Cloud Act USA ils ne sont pas complètement souverains. Il y a beaucoup de lobbying des big techs. Investisseurs, opérateurs, éditeurs, ce sont des Américains. Les data centers sont en France avec des capitaux américains.* », entretien avec un opérateur.

Souveraineté numérique : une définition utile, mais insuffisante si elle reste déclarative

La France et l'Allemagne ont proposé, en juin 2026, une définition commune de la souveraineté numérique européenne. Celle-ci désigne la capacité à développer, fournir, utiliser, adapter et contrôler des technologies numériques, y compris matérielles, de manière indépendante, autodéterminée et sécurisée, afin de préserver la capacité d'action et l'autorité de décision finale des États, administrations, organisations et acteurs économiques.

Cette définition introduit plusieurs dimensions utiles pour instruire un data center : contrôle juridique, maîtrise technologique, réduction des dépendances critiques, protection des données, interopérabilité, infrastructures résilientes, capacités industrielles et valeur créée en Europe.

Mais appliquée aux data centers IA, cette définition révèle des limites opérationnelles. Un site peut être localisé en France, soumis au droit français, raccordé au réseau français et pourtant dépendre de capitaux américains, de modèles entraînés hors d'Europe, de GPU produits à Taïwan, de mémoires HBM coréennes, de câbles contrôlés ou financés par de grandes plateformes, de logiciels propriétaires, d'API cloud et d'interfaces bureautiques ou métiers non européennes.

La souveraineté numérique dans ce cas doit être instruite couche par couche : qui finance, qui opère, qui contrôle le capital, quels modèles tournent, quelles données sont traitées, quelle juridiction s'applique, quelles dépendances techniques existent, quelle capacité de substitution est disponible, quelle valeur reste en Europe et quelles compétences sont réellement construites. La question devient : quelles couches critiques sont maîtrisées, lesquelles sont dépendantes, lesquelles sont substituables et lesquelles rendent la France ou l'Europe vulnérables si elles disparaissent ou changent de conditions d'accès ?

Orange-Morrison : lorsque la maîtrise dépend de la répartition des droits

Le projet de coentreprise annoncé en juillet 2026 par **Orange et Morrison**⁴⁵ illustre ces dissociations. La structure envisagée serait détenue et contrôlée conjointement à 50/50. Orange y apporterait cinq data centers situés en France, son expertise opérationnelle et sa capacité commerciale. Morrison apporterait des capitaux et son expertise d'investisseur mondial en infrastructures. Le programme annoncé viserait une capacité de 400 MW et environ 3 milliards d'euros d'investissements. Orange Business serait le distributeur exclusif des offres de colocation et d'hébergement. Orange conserverait par ailleurs le contrôle opérationnel des zones réservées à ses propres activités.

Ce montage ne peut être qualifié à partir de la seule localisation française des actifs, ni disqualifié par la seule présence d'un investisseur mondial. Il doit être examiné selon la répartition effective des droits : décisions d'investissement, contrôle opérationnel, financement et endettement, choix des clients, juridictions contractuelles, conditions de cession, accès aux données, possibilités de substitution et capacité à poursuivre l'exploitation en cas d'évolution ou de retrait d'un partenaire.

Une même infrastructure peut ainsi présenter plusieurs niveaux de maîtrise selon les espaces et les services concernés. Les zones exploitées directement par Orange pour ses propres besoins peuvent relever d'un niveau de contrôle différent des capacités commercialisées auprès de clients tiers. La nationalité de l'opérateur, celle des investisseurs, la localisation du bâtiment et le contrôle opérationnel ne coïncident pas nécessairement.

Le terme de souveraineté doit donc être rapporté aux droits effectivement détenus et aux services concernés. Il ne qualifie pas uniformément l'ensemble de la plateforme. L'opération étant encore, à la date de l'annonce, un projet soumis aux consultations et autorisations nécessaires, son niveau de maîtrise ne pourra être pleinement apprécié qu'à partir de sa gouvernance définitive et de ses dispositions contractuelles.

La maîtrise réelle dépend de l'agencement entre capitaux, gouvernance, exploitation, contrats, clients, technologies et possibilités de substitution.

4.2 La chaîne de production : une chaîne de dépendances

La souveraineté numérique se joue aussi dans la chaîne de production qui rend les data centers possibles. Et cette chaîne présente des concentrations extrêmes, à chaque maillon. Ces dépendances sont industrielles et géopolitiques. Certaines matières, gaz rares ou équipements nécessaires à la chaîne des semi-conducteurs peuvent être exposés à des tensions commerciales, industrielles ou logistiques. Cette vulnérabilité justifie de traiter la chaîne IA comme une chaîne stratégique multi-dépendante.

Maillon	Concentration	Fragilité
Minerais et matériaux critiques	La Chine contrôle plus de 60% de l'extraction et du raffinage mondiaux (gallium, germanium, terres rares) ⁴⁶	Dépendance aux politiques d'exportation et aux capacités de raffinage chinoises ; concurrence avec les batteries, les équipements électriques, les énergies renouvelables et l'industrie de défense.

⁴⁵ Source : Orange, communiqué du 27 juillet 2026. <https://www.orange.com/en/press-release/orange-et-morrison-annoncent-un-projet-de-creation-de-coentreprise-de-data-centers-pour-renforcer-la-souverainete-numerique-de-leurope-486247-486247> ;

Maillon	Concentration	Fragilité
Hélium et gaz industriels	L'approvisionnement en hélium est concentré entre quelques pays producteurs, notamment les États-Unis, le Qatar et l'Algérie. La fabrication des semi-conducteurs dépend également d'autres gaz de très haute pureté et de chaînes logistiques spécialisées.	Dépendance à des approvisionnements concentrés, sensibles aux tensions géopolitiques, aux incidents industriels et aux ruptures logistiques.
Lithographie EUV	ASML (Pays-Bas), acteur européen critique et unique fabricant mondial d'EUV	Dépendance mondiale à une brique européenne, mais absence de maîtrise européenne sur l'ensemble de la chaîne puces-GPU-cloud
Fonderies et fabrication avancée	TSMC, à Taïwan, concentre une part majeure de la fabrication avancée des puces utilisées pour l'IA. Samsung est un autre acteur important, mais l'Europe ne dispose pas de capacité comparable sur les nœuds les plus avancés.	Concentration géographique et industrielle ; exposition aux tensions autour de Taïwan ; pénuries documentées (Covid, 2020-2022) ; dépendance aux capacités de fabrication et de packaging avancé.
Mémoires HBM	SK Hynix, Samsung - deux acteurs coréens	Aucun producteur européen ; goulot identifié en 2025
Concepteurs et fournisseurs d'accélérateurs IA	Nvidia et AMD, groupes américains, occupent une position centrale dans les unités de calcul utilisées pour l'IA. Les TPU de Google relèvent également d'un écosystème américain propriétaire. Sur une autre brique de la chaîne, celle des processeurs HPC, SiPearl ⁴⁷ développe en Europe des solutions destinées au calcul haute performance, notamment dans le cadre de l'European Processor Initiative et de plusieurs programmes européens. Cette initiative renforce une capacité européenne sur les processeurs, sans constituer à ce stade une alternative complète aux GPU dominants.	L'Europe ne dispose pas encore d'une alternative complète et mature aux accélérateurs dominants. L'accès à ces composants peut être affecté par les politiques d'exportation, les priorités commerciales des fournisseurs et les tensions géopolitiques.

⁴⁶ Forte dépendance à la Chine sur plusieurs matériaux clés : gallium primaire, germanium, terres rares magnétiques et aimants permanents. Les parts varient selon les matériaux et les étapes de chaîne, mais l'AIE documente notamment **60% de la production minière et 91% du raffinage** des terres rares magnétiques en Chine en 2024 ; l'USGS documente **98% de la production mondiale de gallium primaire** en Chine. Sources : <https://www.iea.org/reports/rare-earth-elements/executive-summary> ; <https://pubs.usgs.gov/periodicals/mcs2024/mcs2024-gallium.pdf> ; <https://www.csis.org/analysis/mine-microchip>

⁴⁷ <https://sipearl.com/fr>

Maillon	Concentration	Fragilité
Serveurs et systèmes de calcul	L'intégration des serveurs et systèmes de calcul est dominée par des groupes américains et asiatiques, notamment HPE, IBM, Dell, Lenovo ou Supermicro. Bull constitue l'un des rares acteurs européens disposant de capacités significatives dans le calcul haute performance et les infrastructures de calcul IA.	Capacité européenne limitée face aux acteurs mondiaux ; dépendance persistante aux GPU, processeurs, mémoires, logiciels et composants produits hors d'Europe.
Câbles sous-marins	Place croissante des grandes plateformes numériques dans le financement, la propriété ou l'usage capacitaire des câbles sous-marins	Sabotages documentés en mer Baltique (2023-2024) ; tensions en mer Rouge
Modèles de fondation	OpenAI, Google, Anthropic, Meta (USA) ; DeepSeek, Baidu (Chine) ; Mistral (Europe)	Les capacités d'entraînement à très grande échelle restent fortement dépendantes d'infrastructures et de fournisseurs extra-européens.

L'appel européen lancé en juillet 2026⁴⁸ pour développer jusqu'à sept AI Gigafactories confirme cette lecture par la chaîne. Ces infrastructures visent à renforcer la capacité européenne d'entraînement et d'exécution de modèles avancés. Elles resteront néanmoins, dans un premier temps, dépendantes de processeurs ou d'accélérateurs fournis notamment par Nvidia ou AMD, la Commission reconnaissant l'absence actuelle de puces européennes de pointe équivalentes pour ces installations. La construction d'une capacité européenne de calcul réduit donc une dépendance d'infrastructure sans supprimer les dépendances situées en amont de la chaîne.

La dépendance européenne ne porte donc pas seulement sur la fabrication des puces. Elle concerne aussi la conception et la fourniture des accélérateurs IA, largement dominées par des acteurs américains. Les restrictions déjà appliquées à l'exportation de certains composants avancés montrent que l'accès à la puissance de calcul peut devenir un instrument de politique industrielle, commerciale et géopolitique. Les initiatives européennes, notamment SiPearl dans les processeurs HPC, renforcent certaines briques, mais ne constituent pas encore une chaîne européenne complète permettant de substituer les principales briques dominantes.

Un data center en France peut être le terminal visible d'une infrastructure dont plusieurs maillons essentiels restent hors de portée française ou européenne. L'attirer sans regarder ses dépendances, c'est risquer sa souveraineté.

« Concentration de la chaîne de production : il y a un échelon sur 1 pays, 1 producteur, par exemple la puce IA avec un transistor petit, produite à Taïwan exclusivement. Les machines de lithographie qui permettent la production des puces sont exclusivement faites par ASML. Minerais critiques, c'est la Chine. C'est la même fragilité », entretien avec un expert des projections.

⁴⁸ https://www.lemonde.fr/economie/article/2026/07/30/ia-bruxelles-met-cinq-milliards-d-euros-sur-la-table-et-espere-combler-son-retard-sur-washington-et-pekine_6737050_3234.html

Tout cela s'est construit dans un contexte où **l'infrastructure numérique a longtemps été pensée comme un espace d'innovation porté par des acteurs privés, davantage que comme une infrastructure stratégique à gouverner.**

À mesure que ces services deviennent vitaux, cette lecture devient insuffisante : la maîtrise des réseaux, des plateformes, des données et des capacités de calcul redevient un sujet d'intérêt public. *« L'électricité et le train avaient d'abord été privés, puis nationalisés quand leur caractère stratégique a été reconnu. Pour l'internet, le mouvement n'a pas eu lieu »,* précise une interviewée.

4.3 Raccorder ici, capter ailleurs : l'économie territoriale du data center

Pour éviter de confondre investissement, valeur économique et contribution territoriale, l'instruction d'un projet de data center a besoin de considérer ces trois niveaux. La **valeur créée** désigne les services, calculs, modèles, revenus ou capacités produits par l'infrastructure. La **valeur captée** désigne les acteurs qui en tirent effectivement le bénéfice économique, stratégique ou informationnel. La **valeur partagée** désigne ce qui reste au territoire : fiscalité nette, emplois qualifiés, compétences, chaleur contractualisée, dépollution, accès local au calcul, formation, sous-traitance ou infrastructures réutilisables.

Sans cette distinction, un investissement massif peut masquer une contribution territoriale faible.

La question de la valeur captée est posée directement dans plusieurs entretiens : un data center peut mobiliser les ressources d'une infrastructure lourde - électricité, foncier, réseau, eau, compétences - tout en produisant des services dont la valeur est captée ailleurs, avec peu ou aucune maîtrise locale.

Les entretiens documentent un emploi direct faible. Le Guide du Data Center Durable et Acceptable⁴⁹ indique que le site Equinix PA13 à Meudon, représentant 350 M€ d'investissement et environ 7 000 m² de salles, emploie environ 30 salariés permanents⁵⁰. Le ratio en ETP/MW⁵¹ ne doit pas être avancé sans préciser la puissance retenue, le périmètre du site et la méthode de calcul. Les chiffres annoncés lors des phases de négociation mêlent souvent emplois directs, indirects, induits et sous-traitance de construction, une confusion qui peut conduire des territoires à surestimer leurs retombées.

Cette dissociation entre localisation et valeur captée est renforcée par une transformation structurelle du marché : *« les data centers sont devenus des produits financiers. Des promoteurs immobiliers et des fonds - Australie, États-Unis, Royaume-Uni - sont désormais des porteurs de projets en Île-de-France, avec une logique explicite de « powered land » : sécuriser le foncier et la puissance électrique, puis céder le site clé en main à un opérateur ou à un hyperscaler »,* met en garde un interlocuteur d'une association.

Cette logique est structurellement indifférente au territoire : un promoteur financier n'a pas d'intérêt à négocier des contreparties locales, à valoriser la chaleur fatale ou à s'engager sur la durée. Cela renvoie à l'asymétrie structurelle entre opérateurs et territoire traitée en 3.2. Pour la collectivité, identifier en amont si le porteur est un opérateur de long terme ou un promoteur-revendeur est donc une condition première de toute négociation sérieuse.

⁴⁹ <https://franceurbaine.org/actualites/france-urbaine-ville-de-demain-et-paris-capitale-economique-devoient-un-guide-pour-des-data-centers-durables-et-territorialises/>

⁵⁰ <https://www.ville-demain.com/guide-du-data-center-durable-et-acceptable-partie-4-retombees-socio-economiques/>

⁵¹ <https://www.datacenterdynamics.com/en/news/equinix-launches-data-center-in-paris-france/>

La question de la valeur captée ne se limite pas à l'emploi et à la fiscalité. Elle couvre aussi la valeur des données hébergées, la valeur des modèles exécutés, la « valeur » des décisions prises à partir des calculs réalisés et à qui tout cela bénéficie.

Quatre questions sur la valeur captée

1. **Emploi net** : combien d'ETP permanents, qualifiés, locaux, distingués des emplois de chantier et des effets indirects ?
2. **Fiscalité nette** : quel rendement fiscal après exonérations, abattements et coûts d'infrastructure induits pour la collectivité ?
3. **Contreparties contractualisées** : chaleur fatale, formation, voirie, équipements publics, négociées avant l'autorisation ou promises après ?
4. **Usages hébergés** : qui sont les clients finaux ? Quelles données sont traitées ? Quelle juridiction s'applique ?

Ces informations sont décisives et souvent difficiles à obtenir.

4.4 Héberger sans savoir : l'opacité des clients, des usages et des données

Des data centers peuvent être construits ou réservés avant que leur usage final ne soit stabilisé ; certains porteurs de projets ne savent pas encore à qui ils loueront ou revendront la capacité.

Cette opacité est triple : technique, parce que le modèle, la charge et la consommation réelle ne sont pas toujours connus ; contractuelle, parce que le secret des affaires peut être invoqué ; juridique, parce que la nature des données, leur juridiction et leur régime de protection ne sont pas toujours instruits au moment de l'autorisation.

« Il y a des porteurs de projets qui vont vendre clé en main à des acteurs, mais ils ne savent pas à qui. On ne sait pas quel projet ira derrière. », entretien avec un opérateur.

La transparence se heurte au secret des affaires. Celui-ci est légitime lorsqu'il protège des informations commercialement sensibles, mais il ne peut pas justifier une opacité complète sur des infrastructures mobilisant électricité, foncier, eau, raccordements, fiscalité locale et autorisations publiques. La doctrine doit donc distinguer ce qui relève du secret industriel et ce qui relève de l'intérêt public : maturité du projet, usage principal, ordre de grandeur des clients, profil de charge, contreparties, garanties de souveraineté et plan de réversibilité.

Cette situation crée un risque identifié dans les entretiens : que des acteurs dont les intérêts sont contraires aux intérêts français ou européens se retrouvent propriétaires ou locataires d'infrastructures critiques sur le sol français, sans que les autorités locales ni les services de l'État n'en aient eu connaissance au moment de l'autorisation. Le risque est structurel dès lors que les procédures d'instruction ne demandent pas ces informations.

La directive NIS⁵² et le règlement européen sur la cybersécurité renforcent les obligations de sécurité et de résilience pesant sur les opérateurs d'infrastructures critiques. Mais les entretiens pointent ici un autre problème : un défaut de transparence ex ante qui échapperait à la directive. Investisseurs finaux, clients pressentis, nature des données, usages hébergés et juridictions

⁵² Sources : <https://eur-lex.europa.eu/eli/dir/2022/2555/oj/eng> ; <https://eur-lex.europa.eu/eli/reg/2019/881/oj/eng>

applicables restent insuffisamment instruits au moment des projets d’implantation. La question territoriale demeure donc ouverte : que faut-il savoir avant d’autoriser, raccorder ou soutenir un data center ?

4.5 Investir sans construire de capacité stratégique

Le dernier niveau de la dissociation concerne la capacité stratégique. Attirer des investissements n'est pas forcément synonyme de construire une capacité à décider, innover, concevoir et produire.

Une capacité stratégique n’est cependant pas un stock. Une organisation peut disposer aujourd’hui de compétences, de données, d’outils et même d’alternatives technologiques, tout en perdant progressivement leur maîtrise si elle cesse de les exercer, de les renouveler ou de construire les suivantes. Une capacité qui n’est plus mise en mouvement peut s’enkyster. La souveraineté comporte donc aussi une dimension dynamique : elle tient à ce que l’organisation maîtrise à un instant donné, mais également à sa capacité à maintenir, renouveler et reconstruire cette maîtrise lorsque les technologies, les fournisseurs, les usages ou les compétences évoluent.

Le chapitre 11 reviendra sur cette dimension temporelle en distinguant ce qu’une organisation maîtrise aujourd’hui de ce qu’elle doit continuer à pouvoir maîtriser, apprendre à maîtriser et reconstruire demain.

La France et l'Europe sont aujourd'hui dans une position paradoxale : elles accueillent des infrastructures numériques importantes, mais restent dépendantes des acteurs qui les financent, les opèrent, les peuplent de modèles et décident de leurs usages. Cette dépendance est documentée dans les entretiens à plusieurs niveaux.

Dépendances européennes et tentatives d’émancipation

Couche critique	Dépendance documentée	Initiatives européennes ou françaises	Limites
Modèles de fondation	Les modèles de fondation de rang mondial restent majoritairement américains ou chinois.	Mistral AI, initiatives d’infrastructure souveraine ou européenne, partenariats techniques en France et en Europe.	Capacité financière et industrielle encore très inférieure à celle des grands acteurs américains.
		Le projet BLOOM, entraîné sur le supercalculateur public Jean Zay dans le cadre de BigScience, a démontré dès 2022 une capacité française et européenne à réunir ressources de calcul, recherche internationale, modèle multilingue et documentation ouverte. Il n’a toutefois pas été transformé en un produit largement distribué, intégré et maintenu pour les usages professionnels.	Difficulté à transformer l’excellence scientifique et les infrastructures publiques de recherche en produits diffusés, accompagnés, intégrés aux outils de travail et soutenus par une capacité commerciale durable.

Couche critique	Dépendance documentée	Initiatives européennes ou françaises	Limites
Serveurs et systèmes de calcul avancé	Les processeurs, GPU, mémoires et une grande partie des serveurs et systèmes de calcul utilisés en Europe reposent sur des fournisseurs américains ou asiatiques.	Bull dispose de capacités industrielles et d'intégration dans le calcul haute performance, les systèmes de calcul IA et les environnements critiques. SiPearl développe des processeurs européens pour le HPC.	Dépendance persistante aux accélérateurs, mémoires, composants électroniques et chaînes de fabrication situés hors d'Europe ; capacité de production et échelle commerciale encore limitées face aux acteurs mondiaux.
Infrastructure cloud	Dépendance structurelle aux clouds américains (Azure, AWS, Google Cloud)	OVHcloud, Scaleway et quelques acteurs européens disposent d'offres structurées sous juridiction européenne. Bleu et S3NS, cloud de confiance.	Taille, profondeur de services, réseau commercial et écosystème encore très inférieurs aux hyperscalers.
Solutions de confiance	Dépendance aux suites logicielles, API, clouds et environnements propriétaires dominants.	Docaposte, offres combinant infrastructures souveraines, modèles open source, orchestration et cas d'usage métier.	Passage à l'échelle, intégration aux usages existants et capacité de substitution aux solutions dominantes.
Usages publics	Les administrations, hôpitaux, universités et collectivités restent fortement exposés aux solutions dominantes.	DINUM, visio, suite numérique alternative, logique de substitution progressive et commande publique souveraine.	Adoption, formation, conduite du changement, intégration dans les workflows existants.
Standards et interopérabilité	Sans standards, la dépendance se verrouille dans les formats, API, contrats et architectures.	GAIA-X, initiatives d'interopérabilité, portabilité, standards ouverts.	Déploiement lent, gouvernance complexe, ambition souveraine initiale affaiblie.
Expertise et compétences	L'expertise data centers, cloud, puces, modèles et intégration IA reste concentrée dans quelques pays et entreprises.	Formations spécialisées, compétences hybrides IA-énergie-industrie, acteurs d'intégration métier.	Manque de masse critique, dépendance aux fournisseurs, faible capacité publique d'analyse indépendante.

Le cas BLOOM révèle une asymétrie d'opportunités d'apprentissage. Les organisations ont massivement accepté d'expérimenter des outils américains encore imparfaits, tels que ChatGPT ou Gemini à leurs débuts, de supporter leurs erreurs et de contribuer par leurs usages à leur amélioration. Les solutions issues de la recherche française ou européenne bénéficient rarement du même temps d'apprentissage, du même effort de mise en marché et du même accompagnement des utilisateurs. La difficulté tient à la technologie, mais aussi au passage de la recherche au produit, à la distribution, au marketing, à l'intégration et à la constitution d'une base d'utilisateurs.

« On n'a presque plus de hosting souverain. Legacy pas AI - Azure, Google peuvent nous éteindre. Nos clients demandent quelles solutions open source on va pouvoir opérer pour se libérer des GAFAM. », entretien avec une ESN.

Ces initiatives sont nécessaires, mais elles ne suffisent pas à contrebalancer la puissance des acteurs américains. Leur obstacle commun n'est pas seulement le financement ; c'est aussi la commande. Sans demande publique structurée, les acteurs européens restent cantonnés à des niches, tandis que les administrations, hôpitaux, universités et collectivités continuent d'alimenter les écosystèmes dominants. La souveraineté numérique ne se construit donc pas seulement par l'offre : elle suppose une politique d'achat, de formation, d'intégration et de substitution progressive.

Un acteur national n'est pas, par nature, un agent public de souveraineté⁵³

Le développement d'acteurs français ou européens constitue une condition importante de souveraineté, mais leur nationalité ne suffit pas à qualifier leur contribution stratégique. Une entreprise privée poursuit d'abord une stratégie industrielle, commerciale et financière propre. Elle peut renforcer certaines capacités européennes sans avoir pour mission de garantir l'autonomie stratégique de la France ou de l'Union européenne.

L'élargissement du partenariat entre Mistral et Microsoft en juillet 2026 illustre cette distinction. Mistral développe des modèles et des capacités de calcul en Europe ; Microsoft s'engage à utiliser une partie de ces capacités et distribue les modèles Mistral dans Microsoft Foundry, Copilot Studio et Azure. Mistral accède ainsi à une force commerciale, à des environnements techniques et à une clientèle mondiale que l'écosystème européen peine encore à lui offrir. Microsoft enrichit parallèlement son portefeuille de modèles européens, développe sa capacité en Europe, bénéficie d'une énergie décarbonée et renforce sa position auprès des entreprises et des secteurs régulés.

La question n'est pas de décider si cet accord est souverain ou non dans son ensemble. Il renforce certaines couches et peut en fragiliser d'autres. Il peut accroître la capacité européenne en matière de modèles et de calcul, tout en renforçant le rôle d'un acteur américain dans la distribution, les interfaces, l'intégration aux outils de travail et l'accès aux grands comptes.

Il faut plutôt déterminer quelles capacités sont construites en Europe, quelles dépendances sont réduites, maintenues ou accrues, quels nouveaux points de contrôle apparaissent et quelle capacité de substitution demeure si les conditions d'accès, de prix ou de distribution changent.

La levée de 3 milliards d'euros annoncée par Mistral⁵⁴ le 8 septembre 2026 renforce cette trajectoire. Réalisée en Série D sur une valorisation post-money supérieure à 21 milliards d'euros, elle doit financer la recherche de pointe, l'augmentation des capacités de calcul, le développement de l'infrastructure et l'expansion commerciale du groupe. Mistral vise par ailleurs 1 GW de capacité de calcul d'ici 2030.⁵⁵ L'entreprise cherche ainsi à se positionner sur plusieurs couches de la chaîne - modèles, calcul, infrastructure et produits - plutôt que sur la seule production de modèles.

Cette montée en capacité ne supprime cependant pas les interdépendances. L'infrastructure

⁵³ Source : Microsoft et Mistral, communiqué conjoint, 21 juillet 2026 <https://news.microsoft.com/source/2026/07/21/microsoft-and-mistral-expand-strategic-partnership-to-give-enterprises-and-regulated-industries-frontier-ai-they-can-control/>

⁵⁴ <https://mistral.ai/news/mistral-makes-sovereign-open-weight-ai-to-frontier/>

⁵⁵ <https://mistral.ai/news/regional-inference-open-models-new-compute/>

européenne de Mistral repose notamment sur des GPU NVIDIA, alors que l'accord conclu avec Microsoft⁵⁶ prévoit que celui-ci utilise une partie de cette capacité et distribue les modèles Mistral dans son propre écosystème. Le cas Mistral montre ainsi qu'un acteur européen peut renforcer sa maîtrise et ses capacités sur certaines couches tout en approfondissant simultanément des dépendances techniques, commerciales ou de distribution sur d'autres.

De la souveraineté des infrastructures à la souveraineté des interfaces

La souveraineté ne se joue plus seulement dans les data centers, les câbles ou les puces. Elle se joue aussi dans les interfaces quotidiennes : suites bureautiques, ERP, CRM, GTB, logiciels industriels, outils ESG, plateformes collaboratives. Lorsque l'IA y est intégrée par défaut, l'organisation la reçoit dans des outils déjà critiques.

« Qui tient l'interface tient une partie de l'IA : modèles appelés, données exposées, régions d'exécution, fournisseurs mobilisés, fonctions activées ou non. Cette souveraineté des interfaces doit donc être traitée comme une capacité opérationnelle : savoir voir, paramétrer, désactiver, substituer et auditer. », précise un expert IT.

Cette souveraineté des interfaces rejoint directement la gouvernabilité de l'inférence : lorsque l'IA est intégrée par défaut dans les outils de travail, l'organisation doit pouvoir voir, paramétrer, désactiver, substituer et auditer les chaînes d'inférence qu'elle hérite de ses fournisseurs.

Les échanges suivants, issus d'une rencontre sectorielle, illustrent concrètement ce que cette dépendance aux interfaces implique et pourquoi l'arbitrage entre efficacité et souveraineté est souvent mal posé.

Réduire la dépendance logicielle : de l'efficacité apparente à la souveraineté opérationnelle

Lors d'une rencontre sectorielle, trois échanges successifs ont illustré la confusion qui entoure la notion de souveraineté logicielle.

Un acheteur affirme d'emblée qu'il n'est « pas prêt à renoncer à l'efficacité pour la souveraineté ».

Quelques minutes plus tard, un expert de l'environnement Microsoft dit que Copilot pourrait être intégré dans une offre de confiance portée par Bleu, visant un cadre SecNumCloud, souverain au sens cybersécuritaire du terme.

Le même expert évoque ensuite le « syndrome Windows 10 » : des serveurs SQL qui n'auraient plus de support à partir de 2028, des environnements entiers en fin de vie, une dette technique accumulée qui absorbe une part croissante des budgets IT.

Un responsable IT présent confirme : les nouveaux investissements sont aujourd'hui paralysés moins par un choix souverain que par le coût de sortie des architectures héritées. Il cite le cas d'une grande entreprise qui a voulu migrer d'Office 365 vers la suite Google, a réussi le passage de Teams à Meet, mais s'est heurtée à une résistance massive des utilisateurs au moment de passer d'Excel à Sheets et de Word à Docs et se retrouve aujourd'hui à payer deux licences simultanément.

Ces trois échanges pointent une couche de souveraineté que les discussions techniques et

⁵⁶ <https://news.microsoft.com/source/2026/07/21/microsoft-and-mistral-expand-strategic-partnership-to-give-enterprises-and-regulated-industries-frontier-ai-they-can-control/>

réglementaires laissent généralement de côté : **la couche des interfaces, des usages incorporés et des habitudes organisationnelles.**

Ce n'est pas nécessairement la souveraineté qui dégrade l'efficacité de l'acheteur, c'est avant tout la dette technique accumulée au nom de l'efficacité à court terme.

L'efficacité à court terme tire souvent le balancier vers la dépendance : on conserve les outils dominants parce qu'ils sont connus, efficaces, intégrés, rassurants dans l'instant. Mais cette efficacité apparente fabrique aussi ses propres coûts : dette technique, double licence, verrouillage des formats, dépendance aux feuilles de route éditeurs, perte de compétences internes et difficulté croissante à tester des alternatives.

La souveraineté, à l'inverse, oblige à déplacer le balancier vers l'efficacité à moyen terme : capacité de choisir, substituer, négocier, maîtriser ses données, former ses équipes, maintenir une architecture réversible et ne pas subir indéfiniment les décisions d'un fournisseur dominant.

La question centrale devient alors celle de la transition : comment accepter un coût d'apprentissage à court terme pour retrouver une capacité de choix à moyen terme ?

La souveraineté ne dispense jamais une solution de prouver sa qualité, sa sécurité, son ergonomie et sa contribution à la productivité. Une solution européenne durablement moins performante ne gagnera pas sa place par sa seule origine. Mais la performance doit être mesurée sur un périmètre complet : qualité du service rendu, facilité d'usage, intégration, continuité, coût d'apprentissage, dépendance, portabilité, capacité de négociation et coût de sortie. Construire une alternative exige donc simultanément un effort technologique, une coopération entre recherche et entreprises, une stratégie de mise en marché et un accompagnement réel des utilisateurs.

Les organisations ont accepté les approximations et hallucinations de l'IA générative, financé la montée en compétence des modèles en tant qu'utilisateurs-testeurs-payeurs et contribué ainsi au développement d'outils dont elles ne maîtrisent aucune couche.

Refuser de tester des alternatives crédibles au seul motif qu'elles sont aujourd'hui moins familières ou moins intégrées revient à réserver l'apprentissage collectif aux acteurs dominants.

Réduire la dépendance aux outils dominants devient alors une première étape de souveraineté opérationnelle. Cette transition est autant culturelle et organisationnelle que technique.

Le chapitre 11 reprendra cette question comme l'un des angles morts de la souveraineté par couches.

Une souveraineté numérique démocratique

« Le problème, c'est que les pays qui ont une souveraineté numérique réelle, c'est la Chine, la Russie. Ce sont des pays totalitaires. Souveraineté numérique et dictature ou démocratie ? C'est multidimensionnel : souveraineté des États, des territoires, du politique, populaire. Quel type de souveraineté ? Pour qui ? », questionne un interviewé.

Cette formulation évite deux impasses. La première serait de considérer toute souveraineté numérique comme désirable, sans distinguer la souveraineté démocratique - capacité à décider, réguler, protéger - de la souveraineté autoritaire - contrôle total des données et des usages. La seconde serait de considérer que la dépendance actuelle est inévitable ou acceptable, sans chercher à identifier les briques dont la perte rendrait la France et l'Europe structurellement vulnérables.

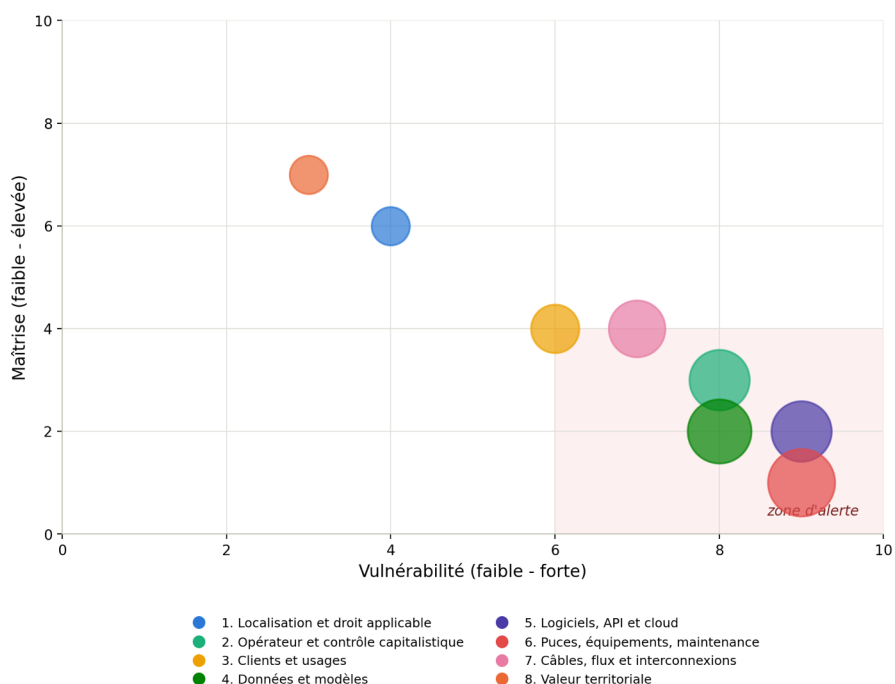
La question stratégique posée par certains interlocuteurs n'est pas : « comment faire comme les Américains ? » Elle est : « qu'est-il critique de ne pas maîtriser ? » La France n'a pas besoin de tout maîtriser pour exister stratégiquement. Elle doit identifier les éléments dont la perte la rendrait dépendante, vulnérable ou substituable. Et sur ces éléments, construire une capacité réelle, au-delà d'un discours d'attractivité.

Cette question en recouvre une autre, plus politique : la France doit-elle aligner ses conditions d'accueil sur celles des juridictions les plus permissives pour ne pas perdre la course aux investissements ou peut-elle tenir une doctrine plus exigeante en pariant sur la qualité de ce qu'elle attire compte plus que le volume ? Les deux options ont un coût. S'aligner par le bas revient à concurrencer des pays qui ne partagent ni le même mix électrique, ni les mêmes garanties de souveraineté, ni la même stratégie industrielle et donc à renoncer à l'avantage différenciant que ce rapport cherche à construire. Tenir une ligne plus exigeante peut ralentir certains projets ou en détourner d'autres vers des territoires moins regardants. Le choix n'est pas neutre. Nous pensons que la France n'a pas intérêt à entrer dans une compétition sur les critères les plus bas, précisément parce que ce n'est pas là que se trouve son avantage.

Outil - Identifier les couches critiques : maîtrise, vulnérabilité, substituabilité

Un data center ne devrait pas être qualifié de « souverain » ou de « non souverain » de manière globale. Sa contribution à la souveraineté doit être évaluée selon les couches qu'il maîtrise, celles dont il dépend, celles qui sont substituables et celles qui créent une vulnérabilité stratégique.

Nous proposons une **grille maîtrise / vulnérabilité / substituabilité** appliquée à huit couches d'instruction : localisation et droit applicable ; opérateur et contrôle capitalistique ; clients et usages ; données et modèles ; logiciels, API, plateformes, distribution et dépendances cloud ; puces, équipements et maintenance ; câbles, flux et interconnexions ; valeur territoriale, capacité locale et maîtrise de la trajectoire.



Cette dernière couche ne mesure pas seulement les retombées économiques locales. Elle examine aussi la capacité du territoire à comprendre le projet, à négocier ses conditions, à en suivre les

effets, à peser sur ses évolutions et à préserver des marges de manœuvre lorsque la trajectoire change.

Cette grille est conçue pour l'instruction d'un data center, objet territorialisé dont la souveraineté dépend autant de l'infrastructure que de ses investisseurs, de ses clients, de ses usages et de la valeur laissée au territoire. Elle constitue une instanciation sectorielle d'une architecture plus générale de la souveraineté par couches et n'a pas vocation à être transposée sans adaptation à tout système d'IA.

La combinaison d'une faible maîtrise, d'une forte vulnérabilité et d'une faible substituabilité signale les dépendances les plus fragiles. Elle ne suffit toutefois pas à les qualifier de critiques : cette qualification dépend également des usages supportés, des conséquences de leur perte et du délai dans lequel une solution de continuité ou de substitution pourrait être activée. Lorsqu'une dépendance fragile porte sur un usage important ou une capacité essentielle, elle ne relève plus d'un simple suivi. Elle appelle une stratégie de réduction de dépendance, une condition d'accueil, une clause contractuelle, une exigence de transparence ou, dans certains cas, une politique industrielle spécifique.

Cette grille transforme la souveraineté en questions à instruire : quelles couches sont maîtrisées, lesquelles exposent le territoire ou l'État, lesquelles peuvent être remplacées et lesquelles justifient une condition d'accueil ou une stratégie industrielle spécifique ?

4.6 Ne pas singer les hyperscalers : choisir les couches où l'Europe peut être critique

La tentation européenne est de répondre à la domination américaine par l'imitation : construire plus grand, raccorder plus vite, attirer plus de capitaux, accumuler plus de puissance de calcul. Cette stratégie de suiveur peut être nécessaire sur certaines briques, mais elle devient dangereuse si elle transforme la souveraineté en course à l'échelle. L'Europe ne gagnera pas nécessairement en reproduisant le modèle hyperscale américain. Elle peut gagner en identifiant les couches où elle peut devenir critique, difficilement substituable ou indispensable à la chaîne de valeur.

Les méga-usines européennes : renforcer une couche sans confondre capacité et souveraineté

Le programme européen⁵⁷ des AI Gigafactories évoqué plus haut donne une nouvelle échelle à cette stratégie. Le montage annoncé pourrait mobiliser plus de 30 milliards d'euros : 5 milliards issus du budget européen, un montant comparable apporté par les États membres accueillant les installations et environ 20 milliards apportés par des acteurs privés. Dix-huit États membres, dont la France, se sont positionnés pour participer à cette infrastructure européenne de calcul. Une partie substantielle du financement européen reste cependant liée au prochain cadre budgétaire de l'Union, encore en négociation à la date de l'annonce.

Ces investissements peuvent réduire un déficit européen de capacité de calcul et donner aux entreprises, aux chercheurs et aux administrations un meilleur accès à des infrastructures d'entraînement et d'inférence. Ils ne constituent pas, à eux seuls, une chaîne souveraine complète. Leur contribution stratégique dépendra des modèles entraînés, des bénéficiaires, des conditions d'accès, des règles de priorité en cas de tension, des logiciels mobilisés et de la capacité à

⁵⁷ https://www.lemonde.fr/economie/article/2026/07/30/ia-bruxelles-met-cinq-milliards-d-euros-sur-la-table-et-espere-combler-son-retard-sur-washington-et-pekine_6737050_3234.html

développer autour d'eux des compétences et des activités industrielles européennes.

Le test de souveraineté ne doit donc pas porter uniquement sur le nombre de GPU installés ou sur la localisation des bâtiments. Il doit demander : qui peut accéder à la capacité, selon quelles priorités, à quel prix, pour quels modèles et quels usages, avec quels droits en cas de pénurie ou de changement de conditions commerciales ?

Le financement public ne devrait pas seulement acheter des équipements. Il devrait aussi garantir des droits d'accès, des capacités de gouvernance, de la transparence, des trajectoires de substitution et un apprentissage industriel européen. Sans ces conditions, l'Europe pourrait financer des infrastructures situées sur son territoire tout en restant dépendante pour les puces, les logiciels, la maintenance, les plateformes et une partie de la valeur créée.

Cette vigilance rejoint le risque de « comptoir »⁵⁸ : un territoire peut fournir le foncier, l'électricité, l'eau et les infrastructures tandis que les modèles, les décisions et la valeur stratégique restent largement contrôlés ailleurs. La localisation de la capacité devient souveraine seulement si elle s'accompagne d'un pouvoir réel sur ses conditions d'utilisation et sur la valeur qu'elle permet de produire.

La contribution stratégique devra aussi être appréciée avec sa matérialité : consommation électrique et hydrique, origine des nouvelles capacités bas-carbone, équipements de secours et critères environnementaux vérifiables. Les informations publiées lors de l'appel indiquaient que les porteurs de projets devaient encore préciser eux-mêmes une partie de leurs objectifs environnementaux.

ASML montre qu'une position stratégique ne suppose pas de maîtriser toute la chaîne : elle peut venir d'une brique critique. STMicroelectronics illustre une autre forme de capacité européenne, moins située sur les GPU de très grande puissance que sur les microcontrôleurs, l'IA embarquée, les capteurs, la connectivité et l'électronique de puissance. Ces briques sont déterminantes pour l'edge industriel, les véhicules, les équipements énergétiques, les infrastructures critiques et certains composants des data centers. La participation de STMicroelectronics à des initiatives autour de RISC-V montre également qu'une stratégie européenne peut chercher à réduire certaines dépendances d'architecture sans prétendre substituer immédiatement l'ensemble de la chaîne.

De la même manière, la France et l'Europe peuvent chercher leur place dans l'orchestration des usages, l'intégration métier, la gouvernance des infrastructures, les architectures frugales, les logiciels opérationnels, les standards, la commande publique et les compétences hybrides IA-énergie-industrie. La question n'est pas de tout maîtriser, mais de se rendre indispensable sur les couches qui conditionnent la décision.

Construire des positions européennes sans imiter l'hyperscale

Les entretiens font émerger plusieurs domaines dans lesquels la France et l'Europe pourraient construire une position forte et différenciante, non pas en imitant le modèle hyperscale américain, mais en transformant certaines de leurs contraintes et compétences en avantages :

1. **l'intégration de l'IA dans les systèmes critiques et industriels** (réseaux, jumeaux numériques, contrôle-commande, robotique) ;

⁵⁸ https://www.lemonde.fr/idees/article/2026/07/27/l-ia-repose-sur-une-geographie-concrete-mettant-des-territoires-au-service-du-calcul-selon-une-logique-de-comptoirs_6733876_3232.html

2. **la mesure environnementale et la transparence des infrastructures** (socle réglementaire CSRD, taxonomie, EED⁵⁹) ;
3. **le refroidissement avancé et l'efficacité des data centers** (par exemple DLC, free cooling, retrofit) ;
4. **les infrastructures de cloud de confiance et d'IA maîtrisée ou souveraine** (par exemple Mistral, OVHcloud, Docaposte) ;
5. **la flexibilité réseau et l'interaction data centers-système électrique** (par exemple Dcflex) ;
6. **les systèmes distribués et l'orchestration calcul-énergie-données** (par exemple Genesis).

Ces domaines ont en commun de s'appuyer sur des compétences déjà présentes en Europe, des besoins industriels réels et des positions susceptibles de devenir difficiles à substituer.

Encadré - Corée du Sud : construire une souveraineté par la chaîne plutôt que rechercher l'autarcie

La Corée du Sud offre une illustration d'une stratégie de souveraineté construite à partir de capacités industrielles existantes.⁶⁰ Le pays dispose déjà de positions mondiales fortes dans les semi-conducteurs de mémoire, notamment la HBM (High Bandwidth Memory) qu'il cherche à prolonger vers des couches adjacentes : fabrication avancée, accélérateurs IA nationaux, capacité de calcul, data centers, modèles de fondation, IA physique, compétences et usages industriels. La stratégie « AI G3 », annoncée le 26 septembre 2024 avec l'objectif de faire de la Corée du Sud l'une des trois principales puissances mondiales de l'IA⁶¹, s'inscrit elle-même dans la continuité d'une stratégie nationale de l'IA engagée dès 2019⁶².

Cette trajectoire ne signifie pas que la chaîne devient souveraine dans son ensemble. Si d'un côté les accélérateurs coréens cherchent à réduire la dépendance à Nvidia sur certains usages, de l'autre les principaux acteurs coréens continuent à déployer des infrastructures reposant largement sur des GPU Nvidia. La souveraineté se construit donc par gains successifs de maîtrise et de substituabilité plutôt que par indépendance complète.

La Corée cherche également à orienter l'IA vers les secteurs où elle dispose déjà d'un appareil industriel puissant. Le gouvernement soutient ainsi le développement de l'IA physique et agentique, de modèles destinés à la robotique et de robots humanoïdes appelés à être expérimentés ou déployés dans des secteurs tels que la construction navale, la défense, le spatial, la logistique, l'industrie manufacturière ou la santé. Les programmes engagés visent notamment à associer agents IA, robots, données industrielles et savoir-faire métier au sein de systèmes de production plus autonomes. La stratégie consiste ainsi moins à reproduire la seule course aux modèles généralistes qu'à transformer des positions industrielles existantes en nouvelles capacités productives.

Cette politique révèle également l'échelle matérielle et financière d'une stratégie IA. Le programme prévoit une première phase de 8,4 GW de capacités de data centers IA, pour environ 550 000 milliards de wons d'investissements⁶³, puis une montée à 18,4 GW à l'horizon 2035,

⁵⁹ CSRD : Corporate Sustainability Reporting Directive ; taxonomie européenne : système de classification des activités économiques durables ; EED : Energy Efficiency Directive, directive européenne relative à l'efficacité énergétique.

⁶⁰ Voir https://www.lemonde.fr/economie/article/2026/09/07/en-coree-du-sud-l-ia-s-impose-au-c-ur-des-evolutions-economiques-et-sociales_6767934_3234.html, pour le cadrage général, les tensions sociales et énergétiques et les choix stratégiques.

⁶¹ <https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=42&mId=4&nttSeqNo=1037&sCode=eng>

⁶² <https://www.msit.go.kr/eng/bbs/view.do?bbsSeqNo=46&mId=10&mPid=9&nttSeqNo=9&sCode=eng>

pour un investissement cumulé annoncé supérieur à 1 000 000 de milliards de won⁶⁴. Les projections préparatoires au 12e plan électrique portent par ailleurs la pointe nationale à 158,4 GW dans le scénario de base et 165 GW dans le scénario haut à l’horizon 2040, en intégrant notamment les mégaprojets de data centers IA et de semi-conducteurs⁶⁵.

Cette stratégie crée cependant ses propres vulnérabilités. Le poids des semi-conducteurs dans l’économie coréenne accroît l’exposition du pays au cycle mondial d’investissement dans l’IA, tandis que les nouvelles capacités de calcul et de fabrication déplacent une partie de la contrainte vers le système électrique. En septembre 2026, le gouvernement estimait que les besoins liés à l’IA pourraient accroître de 25 à 30 GW la demande de puissance du pays.⁶⁶ Construire une capacité industrielle stratégique ne supprime donc ni le risque de cycle ni les contraintes d’infrastructure que cette stratégie engendre. Cette stratégie de souveraineté crée de nouvelles capacités, mais aussi de nouvelles concentrations de risques qu’une stratégie industrielle doit rendre visibles.

Le cas coréen montre ainsi qu’au-delà de la protection des actifs existants, une stratégie de souveraineté demande de mobiliser un capital industriel, technologique et humain déjà constitué pour construire les capacités suivantes, tout en conservant suffisamment de marges d’adaptation et de réversibilité pour que ce pari technologique ne se transforme pas en dépendance industrielle.

Cette stratégie suppose de déplacer le débat, se demander « quelles capacités critiques construisons-nous grâce aux data centers ? » Si l’infrastructure localisée en France sert principalement des modèles, des interfaces, des contrats, des données et des marges contrôlés ailleurs, l’attractivité reste partielle. Si elle renforce des usages publics, des compétences locales, des standards ouverts, des acteurs européens, des architectures sobres et des capacités d’intégration métier, elle peut devenir une brique de maîtrise.

4.7 Ce que ce chapitre implique : de l’attractivité à la maîtrise

La dissociation entre localisation, souveraineté et valeur captée n’est pas une fatalité technique. Elle résulte de procédures d’instruction encore trop centrées sur l’investissement, la puissance raccordée et l’implantation physique, au détriment de la maîtrise, des usages, des dépendances et de la valeur réellement partagée.

La question stratégique devient donc : à quelles conditions un data center situé en France devient-il un actif collectif, plutôt qu’un passif énergétique, territorial, fiscal ou souverain ?

Les entretiens suggèrent certaines conditions.

Condition	Ce qu'elle implique concrètement
Transparence sur les investisseurs et les usages	Conditionner les autorisations d'urbanisme et les raccordements à la déclaration des investisseurs finaux, des clients pressentis et de la nature des données hébergées, pas après l'implantation, mais avant

⁶³ Sources : <https://english.mofe.go.kr/pc/selectTbPressCenterDtl.do?boardCd=N0001&seq=6445> ; <https://www.reuters.com/business/media-telecom/key-facts-south-koreas-three-chip-ai-mega-projects-2026-06-29/>

⁶⁴ <https://www.datacenterdynamics.com/en/news/south-korea-announces-919bn-investment-into-three-mega-projects-plans-to-build-184gw-worth-of-data-centers-by-2035/>

⁶⁵ <https://view.asiae.co.kr/en/article/2026082014290939570>

⁶⁶ <https://www.reuters.com/business/energy/south-koreas-power-demand-set-soar-ai-boom-nation-weighs-expanding-nuclear-2026-09-08/>

Condition	Ce qu'elle implique concrètement
Contreparties contractualisées	Chaleur fatale, formation, voirie, contribution aux réseaux de chaleur, reporting environnemental, négociées et contractualisées avant l'autorisation, pas promises après
Ancrage dans l'écosystème local	Préférer les projets qui s'intègrent dans une filière territoriale, forment des compétences locales, utilisent des fournisseurs locaux ou contribuent à des usages collectifs identifiés
Commande publique souveraine	Orienter les achats de l'État, des hôpitaux, des collectivités vers des solutions offrant une maîtrise européenne vérifiable des couches critiques : infrastructure d'exécution, données, modèles, administration, droits de propriété ou d'usage, capacité d'évolution, maintenance, interfaces, portabilité et conditions de sortie. La commande publique doit aussi contribuer à créer des compétences, une capacité industrielle et une valeur économique durablement captée en France ou en Europe.
Souveraineté des interfaces	Exiger la capacité à voir, paramétrer, désactiver, substituer et auditer les fonctions IA intégrées dans les suites logicielles, outils métiers, API et plateformes utilisées par les organisations publiques et critiques.

Un data center ne devient stratégique que lorsqu'il cesse d'être une boîte noire, et devient porteur d'usages connus, d'investisseurs identifiés, de dépendances instruites, de contreparties contractualisées, de données qualifiées, de valeur partagée vérifiable.

L'origine du fournisseur ne suffit donc pas. La question est de savoir quels droits de décision, quelles compétences, quelle propriété intellectuelle, quelle capacité d'évolution et quelle part de la valeur économique restent effectivement sous maîtrise française ou européenne.

Les configurations récentes, notamment Mistral-Microsoft et le projet Orange-Morrison, montrent enfin que la souveraineté ne se distribue plus entre des acteurs entièrement souverains et des acteurs qui ne le seraient pas. Elle cherche à se construire dans des montages hybrides : modèle européen distribué par une plateforme américaine ; infrastructure française gouvernée avec un investisseur mondial ; capacité de calcul européenne équipée de puces extra-européennes. Ces configurations doivent être instruites selon les couches qu'elles renforcent, celles qu'elles fragilisent et les droits de décision qu'elles laissent effectivement aux acteurs européens.

Dans ce contexte, la maîtrise des flux⁶⁷ devient aussi importante que celle des actifs : accès aux modèles, plateformes de distribution, environnements de développement, mises à jour, interfaces, contrats, capacités de calcul et accès aux clients. Celui qui organise les conditions de circulation d'une technologie peut disposer d'un pouvoir stratégique sans en posséder l'ensemble des actifs.

Enfin, la souveraineté ne peut pas être seulement géographique, juridique ou industrielle. Elle doit aussi être démocratique : rendre visibles les arbitrages que ces infrastructures déplacent vers le réseau, le foncier, l'eau, les câbles, les logiciels, les données et les territoires.

La Partie II déplace la question : si la demande IA se matérialise dans ces infrastructures, comment reprendre la main sur cette demande elle-même ?

⁶⁷ https://www.lemonde.fr/idees/article/2026/08/01/la-puissance-ne-repose-plus-seulement-sur-les-territoires-les-ressources-ou-l-armee-elle-depend-de-la-capacite-a-controler-les-flux_6737183_3232.html

Interlude stratégique - Ne pas singer les *hyperscalers* : vers le data center gouvernable

L'Interlude en une phrase

Plutôt que d'imiter les *hyperscalers*, la France peut construire une géographie fonctionnelle du calcul, en distinguant entraînement, post-entraînement, évaluation et sécurisation, inférence, inférence critique et *edge* industriel, chacun ayant ses propres contraintes de puissance, de temporalité, de latence et de gouvernance. L'interlude referme la Partie I sur l'idée que la France doit choisir non seulement où accueillir le calcul, mais comment le qualifier et le piloter.

Du réservoir de puissance à l'infrastructure gouvernable

L'alternative au modèle *hyperscale* n'est pas forcément de construire plus petit. Il faut plutôt opposer les infrastructures indifférenciées aux infrastructures gouvernables, les puissances raccordées sans usage qualifié aux capacités orientées vers des fonctions identifiées, les sites subis aux infrastructures intégrées dans un territoire, un réseau et une stratégie industrielle.

Cet interlude stratégique explore cette autre voie : des data centers pensés comme des infrastructures à orchestrer entre énergie, calcul, données, chaleur, usages locaux et souveraineté opérationnelle, plutôt qu'uniquement comme charges à raccorder.

Certains exemples qui suivent sont opérationnels, d'autres expérimentaux ou conceptuels. Leur intérêt commun est ailleurs : tous déplacent le data center d'une logique de puissance raccordée vers une logique de fonction gouvernée.

IS.1 Outiller les territoires : passer du consentement à l'instruction

France Urbaine et Ville de Demain proposent un modèle de gouvernance territoriale : donner aux collectivités les moyens d'instruire, de négocier et de conditionner l'accueil des data centers. Cette approche repose sur trois déplacements : passer d'une décision au cas par cas à une planification anticipée ; compléter les indicateurs techniques internes au site par des indicateurs territorialisés ; transformer les promesses d'opérateurs en critères vérifiables et négociables.

Les indicateurs **ICE**, **IPH** et **IEF**⁶⁸ sont particulièrement utiles dans cette perspective, car ils déplacent l'évaluation du data center hors de ses seuls murs⁶⁹ : circularité réelle de la chaleur fatale, pression hydrique locale, efficacité d'usage du foncier. Ils complètent le PUE, WUE ou ERF⁷⁰ avec la question : **comment le site s'inscrit-il dans son territoire ?**

Ce modèle donne une forme opérationnelle à la doctrine publique et renforce l'alternative à l'implantation subie, par une capacité locale à définir les conditions d'accueil avant l'arrivée du projet⁷¹.

⁶⁸ ICE : indicateur de circularité énergétique ; IPH : indicateur de pression hydrique ; IEF : indicateur d'efficacité foncière. Ces indicateurs sont proposés dans le *Guide du Data Center Durable et Acceptable*, France Urbaine / Ville de Demain, 2026. <https://www.ville-demain.com/le-guide-du-data-center-durable-et-acceptable-vers-un-nouveau-pacte-numerique-territorial/>

⁶⁹ <https://franceurbaine.org/actualites/france-urbaine-ville-de-demain-et-paris-capitale-economique-devoient-un-guide-pour-des-data-centers-durables-et-territorialisés/>

⁷⁰ ERF (Energy Reuse Factor) : indicateur mesurant la part de l'énergie totale consommée par un data center qui est réutilisée à l'extérieur du site, notamment sous forme de chaleur valorisée. Voir ISO/IEC 30134-6 <https://www.itu.int/epublications/publication/itu-t-l-1328-2025-09-specification-for-waste-heat-reuse-in-telecommunication-rooms-and-data-centres>

⁷¹ <https://www.banquedesterritoires.fr/datacenters-un-guide-pour-outiller-les-collectivites-face-aux-operateurs>

IS.2 Genesis : quand le calcul devient distribuable

Le projet Genesis⁷², développé au LIRMM et présenté par le CNRS, explore une architecture de mini data centers alimentés par panneaux solaires et capables d'échanger à la fois des données et de l'énergie. Il ouvre un autre imaginaire que celui du data center centralisé : l'énergie y est traitée comme une ressource finie, allouée et acheminée entre unités au même titre que les données ou la capacité de calcul.

L'intérêt de Genesis est de déplacer la question : **et si l'on pensait le calcul non comme une demande centralisée à satisfaire, mais comme une charge distribuable, modulable et articulée aux ressources locales ?** Sur un prototype expérimental installé à Polytech Montpellier, les chercheurs du LIRMM estiment que la capacité de faire migrer énergie et données entre unités de calcul permettrait de **réduire de 22 à 27% l'énergie** issue du réseau électrique. Ce chiffre documente une piste expérimentale, il n'est pas encore une doctrine industrielle déployée.

Genesis sert donc de laboratoire conceptuel : il montre qu'un data center peut être pensé comme une pièce d'un système énergétique local articulant production, stockage éventuel, calcul, chaleur, usages de proximité, contraintes réseau et pilotage.

L'article sur Genesis apporte aussi un lien précieux avec la gouvernance locale. Les chercheurs évoquent des scénarios à l'échelle d'un quartier, la valorisation potentielle pour la collectivité locale, l'acceptabilité, l'économie et la gouvernance des données. Cela renforce l'angle selon lequel l'alternative au data center subi n'est pas seulement technique, mais aussi territoriale, sociale et politique.

IS.3 Distribuer le calcul, mais aussi la décision

Des travaux de recherche sur les systèmes distribués, par exemple ceux de Vincent Hilaire⁷³ contribuent à penser les infrastructures numériques, les systèmes énergétiques et l'IA comme des systèmes distribués, dynamiques, multi-agents, multi-échelles et composés d'acteurs hétérogènes. L'alternative au data center subi n'est pas seulement de distribuer les data centers, mais aussi de distribuer la décision.

Dans les deux cas (Genesis et Hilaire), données, énergie et calcul deviennent les composantes d'un même système distribué.

Une analogie issue d'un entretien en robotique va dans le même sens : *« un seul robot est inefficace. Il faut un essaim pour un travail de groupe, dont un ou 2 pour s'occuper de l'énergie. »*

Cette perspective rejoint les enjeux de smart grids, microgrids, robotique distribuée, systèmes industriels et IA hybride. Elle permet de formuler une alternative au modèle hyperscale : au lieu de concentrer toujours plus de calcul dans quelques sites géants, il devient possible d'imaginer une orchestration plus fine entre data centers, edge, stockage, signaux de prix, signaux carbone, disponibilité locale, chaleur récupérable, besoins industriels et contraintes réseau.

⁷² voir article <https://www.lirmm.fr/2025/01/08/felicitacion-au-projet-genesis-distinguee-par-un-best-paper-award/> et <https://lejournel.cnrs.fr/articles/penser-des-datacenters-moins-energivores>

⁷³ <https://scholar.google.com/citations?user=14DVyzAAAAAJ&hl=fr> . N.B. : Cette lecture est une interprétation analytique de la rédactrice du rapport, issue du croisement entre l'article CNRS et le fichier de Vincent Hilaire.

IS.4 DCflex⁷⁴ : tester le comportement système des data centers

Déjà mentionné en §2.6, DCflex est un démonstrateur RTE / Data4 / Schneider Electric qui vise à tester l'interaction entre data centers et système électrique, notamment dans des scénarios de stabilité et d'aléas réseau.

Il apporte un signal opérationnel complémentaire : là où Genesis et Hilaire ouvrent une vision distribuée, DCflex montre qu'un data center peut aussi être interrogé dans son interaction avec le réseau électrique.

L'intérêt de DCflex est double. D'abord, il change le statut du data center : celui-ci n'est plus seulement une charge à raccorder, mais un objet dont le comportement peut être étudié, modélisé, simulé et potentiellement piloté. Ensuite, il relie directement les data centers aux jumeaux numériques, à la flexibilité, aux batteries, aux contraintes de tension et à la gestion dynamique du système électrique.

DCflex montre que la doctrine d'accueil pourrait évoluer : un data center devrait documenter non seulement sa puissance, mais aussi son comportement système. Bien entendu, cela ne prouve pas que tous les data centers pourront devenir flexibles, ni qu'ils contribueront automatiquement à la stabilité du réseau.

Cette idée prolonge le chapitre 2 : l'accès au raccordement pourrait être conditionné non seulement à la maturité et à la localisation, mais aussi à la capacité à interagir proprement avec le réseau.

Le blackout ibérique du 28 avril 2025⁷⁵ fournit un signal utile, même s'il ne concerne pas directement les data centers. Le rapport d'experts coordonné par ENTSO-E conclut que l'événement résulte d'une combinaison de facteurs : oscillations, lacunes de contrôle de la tension et de la puissance réactive, différences de pratiques de régulation de tension, réductions rapides de production et déconnexions de générateurs en Espagne. Ces facteurs ont conduit à une montée rapide de tension, puis à des déconnexions en cascade sur la péninsule ibérique.

Partant de cet exemple et concernant les data centers, il s'agit de rappeler qu'un grand consommateur électrique ne doit pas seulement être évalué par ses MWh ou sa puissance appelée. Son comportement dynamique compte aussi : réaction aux creux de tension, stabilité, redondance, batteries, groupes de secours, flexibilité réelle, capacité à rester connecté ou à se découpler proprement en situation d'aléa. C'est ce que DCflex permet de mettre à l'épreuve : tester, simuler et qualifier son comportement système.

IS.5 La souveraineté se mesure aussi dans le cycle de vie

La différenciation des acteurs choisis doit aussi être instruite sur l'efficacité, le réemploi, le WUE, la durée de vie matérielle, la transparence, la localisation d'exécution, la réversibilité et la capacité à s'intégrer dans des politiques territoriales. C'est à cette condition que la souveraineté devient une capacité opérationnelle, au-delà d'un label.

Par exemple, OVHcloud présente une approche industrielle plus intégrée du cycle de vie. Les éléments collectés évoquent le réemploi, le désassemblage, la mesure d'impact, l'Environmental Impact Tracker, la maîtrise de certains procédés de refroidissement et la valorisation de bâtiments

⁷⁴ <https://www.rte-france.com/actualites/projet-dcflex-epri-contexte-narratif-premier-demonstrateur-france> ;
<https://www.data4group.com/actualites-data4/data4-rte-et-schneider-electric-lancent-le-1er-demonstrateur-europeen-sur-linteraction-entre-data-centers-et-systeme-electrique/>

⁷⁵ <https://www.entsoe.eu/publications/blackout/28-april-2025-iberian-blackout/>

existants ou réhabilités. Ce cas donne un exemple français concret d'intégration verticale : concevoir, exploiter, mesurer, désassembler, réemployer, documenter. Il montre qu'une alternative industrielle peut passer par des choix concrets : durée de vie, réemploi, mesure, désassemblage, refroidissement, implantation et transparence. Ces choix rapprochent la frugalité de la gouvernance industrielle, au lieu de la limiter aux comportements des utilisateurs ou des développeurs.

Les données recueillies auprès d'OVHcloud donnent un ordre de grandeur : PUE déclaré autour de 1,24, WUE de 0,34 l/kWh⁷⁶, durée de vie de certains composants affichée jusqu'à 9 ans et environ 90% des data centers implantés sur des friches industrielles ou des sites existants. Ces chiffres montrent que certaines pratiques industrielles alternatives peuvent être en partie quantifiées et vérifiées, à condition d'exiger des opérateurs un niveau de transparence comparable.

L'intérêt du brownfield n'est pas seulement environnemental. Il est aussi économique et opérationnel. Un interviewé le résume ainsi : « *Les friches industrielles représentent 90% des data centers d'OVH. Ce n'est pas plus cher que le neuf. Les infrastructures, c'est moins de 2% du bilan carbone. Le brownfield est intéressant en termes de cash et de go to market, car il est rapide : 6 à 12 mois. Le greenfield, c'est beaucoup de cash à sortir. Le return on cash est plus élevé en brownfield, car il y a moins de CAPEX à engager.* »

IS.6 Faire le bon calcul au bon moment

Les entretiens évoquent aussi des pratiques ou pistes autour du signal carbone local, de la cloche solaire, du signal prix et de la modulation des usages dans le temps.

Dans un pays comme la France, où l'intensité carbone moyenne de l'électricité est déjà relativement faible, le *carbon-aware computing* géographique⁷⁷ peut avoir moins d'intérêt que dans des systèmes fortement contrastés. En revanche, exécuter certains traitements au bon moment, lorsque le réseau est moins contraint ou lorsque la production locale est favorable, peut devenir une piste pertinente.

Cette piste rejoint la frugalité infrastructurelle. Il s'agit de faire le bon calcul, au bon moment, sur la bonne infrastructure, pour le bon usage et pas seulement moins de calcul en toute situation.

Elle rejoint aussi la logique des calculs distribués : si l'énergie devient une ressource finie, située et pilotable, alors le calcul ne peut plus être pensé comme une demande indifférente au moment et au lieu. **L'infrastructure alternative est celle qui sait articuler disponibilité énergétique, priorité d'usage, contrainte réseau, coût, carbone, chaleur, données et décision.**

Cette logique temporelle n'est pas applicable partout : toutes les charges IA ne se déplacent pas de la même manière, car elles n'ont ni les mêmes contraintes de latence, ni les mêmes exigences juridiques et de continuité, ni les mêmes liens avec les usages finaux.

IS.7 SOCloud : rendre l'utilisateur partie prenante de l'arbitrage

Le projet ANR SOCloud (*Sober and Open Cloud*)⁷⁸ explore une autre dimension de cette gouvernabilité : non plus seulement déplacer le calcul dans le temps ou adapter l'infrastructure, mais **rendre l'utilisateur partie prenante de l'arbitrage**. Porté par FEMTO-ST avec l'IRIT Toulouse, le LIG Grenoble et Eaton Grenoble, le projet étudie les conditions humaines et techniques d'une sobriété appliquée à des data centers de puissance intermédiaire alimentés par un

⁷⁶ Source : <https://corporate.ovhcloud.com/fr/sustainability/environment/>

⁷⁷ Voir 6.5 pour plus de détails sur le *carbon-aware computing*.

⁷⁸ Projet ANR SOCloud, FEMTO-ST : <https://projects.femto-st.fr/socloud/en>

mix énergétique. Il vise à développer des solutions cloud sobres, ouvertes et « conviviales », permettant aux utilisateurs de choisir, selon le contexte, de maintenir, dégrader, différer ou annuler certains accès au cloud, notamment à partir de scores ou de mécanismes d'incitation.

Le projet considère également le point de vue du fournisseur lorsqu'il doit répondre à une demande d'effacement ou à une variation des conditions énergétiques : mobiliser des réserves, réduire temporairement l'offre ou, à l'inverse, proposer davantage de capacité lorsque l'énergie est disponible, peu coûteuse et décarbonée. Cette approche élargit la frugalité infrastructurelle : faire le bon calcul au bon moment suppose aussi de rendre visibles les contraintes et de permettre un arbitrage sur le niveau de service réellement nécessaire.

SOCloud introduit ainsi une dimension complémentaire à Genesis et DCflex : Genesis interroge la distribution du calcul, DCflex le comportement du data center vis-à-vis du réseau, tandis que SOCloud explore la manière dont l'utilisateur et le fournisseur peuvent participer à l'ajustement de la demande numérique elle-même. Le projet est encore expérimental et prévoit un démonstrateur sur l'infrastructure SLICES-FR.

IS.8 Une géographie fonctionnelle du calcul IA

Une autre distinction devient structurante dans les entretiens comme dans les analyses récentes : celle entre data centers d'entraînement et data centers d'inférence. Les deux sont souvent regroupés sous l'étiquette « *data centers IA* », alors qu'ils ne posent pas les mêmes questions d'implantation, de réseau, de densité, de latence et de gouvernance.

Les **data centers d'entraînement** servent à développer ou affiner les grands modèles. Ils concentrent des clusters GPU très denses, des interconnexions spécialisées, des besoins importants de refroidissement et des charges pouvant durer plusieurs jours, semaines ou mois. Ils sont moins sensibles à la latence utilisateur : leur localisation peut donc être davantage guidée par l'accès à une puissance électrique disponible, à un foncier adapté, à un refroidissement maîtrisé, à une connectivité robuste et à des conditions environnementales acceptables. Les entretiens et analyses mobilisés pour ce sujet décrivent ainsi l'entraînement comme une charge très intensive, compatible avec des implantations plus éloignées si les conditions électriques et techniques sont réunies⁷⁹.

Les **data centers d'inférence** servent les modèles en production : requêtes utilisateurs, copilotes, moteurs de recherche, recommandation, agents logiciels, automatisation documentaire ou décisionnelle. Ils fonctionnent souvent de manière plus continue que les clusters d'entraînement, avec des volumes de requêtes qui peuvent devenir très élevés lorsque les services sont intégrés dans des usages quotidiens. Ils sont aussi plus sensibles à la latence, à la disponibilité, à la localisation des données et à la proximité fonctionnelle des usages.

Mais l'inférence ne signifie pas automatiquement « edge » ou proximité immédiate. Certaines charges peuvent être routées vers des régions de calcul plus éloignées si la latence⁸⁰, la réglementation, la disponibilité de capacité, les coûts de transfert de données et la continuité de service restent acceptables. D'autres, au contraire, doivent être exécutées au plus près des

⁷⁹ <https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/the-next-big-shifts-in-ai-workloads-and-hyperscaler-strategies>

⁸⁰ Un article récent sur l'inférence comme demande électrique relocalisable insiste justement sur cette frontière énergie-latence ; Uptime Institute définit aussi l'edge comme une distribution des capacités de calcul et de stockage au plus près des lieux où les usages se produisent. <https://arxiv.org/abs/2604.27855>

utilisateurs, des équipements industriels, des administrations ou des infrastructures critiques. L'inférence appelle donc une géographie plus différenciée que l'entraînement : parfois régionale, parfois edge, parfois centralisée, mais toujours dépendante de la fonction rendue.

Concernant la doctrine d'accueil, cette distinction implique de ne pas instruire seulement « un data center IA », mais de qualifier le type de charge qu'il hébergera : entraînement, post-entraînement, évaluation et sécurisation, inférence grand public, inférence critique, cloud généraliste, stockage, calcul scientifique, *edge* industriel ou services publics sensibles. À la question « quelle puissance ? », s'ajoutent « quelle fonction du calcul, quelle temporalité, quelle latence, quelle continuité de service, quelle flexibilité, quelle valeur locale et quelle possibilité de déplacement temporel ou géographique ? »

Type de charge	Logique dominante	Implication territoriale
Entraînement	Puissance, densité GPU, refroidissement, charges longues	Sites à forte capacité électrique, foncier adapté, moindre contrainte de latence
Post-entraînement, évaluation et sécurisation	Calcul intensif mais souvent discontinu ; tests, RL (apprentissage par renforcement), évaluations, <i>red teaming</i> , validation et amélioration des modèles	Besoin de capacités GPU importantes mais potentiellement plus modulables que l'inférence critique ; localisation à instruire selon disponibilité électrique, souveraineté des données et modèles, sécurité et possibilité de déplacement temporel
Inférence	Latence, continuité, proximité des usages, disponibilité	Sites plus distribués, <i>edge</i> , proximité des bassins d'usage
Inférence critique	Continuité, faible latence, contrôle des données, gouvernance renforcée	Infrastructures régionales ou sectorielles sous gouvernance spécifique
<i>Fine-tuning</i>	Charge intermédiaire, données sensibles, adaptation métier	Infrastructures souveraines, sectorielles ou régionales
<i>Edge</i> industriel	Latence faible, confidentialité, résilience locale	Calcul au plus près des équipements critiques

Cette différenciation ouvre une piste de modèle alternatif : ne pas concentrer toute l'IA dans les mêmes infrastructures, mais construire une géographie fonctionnelle du calcul.

Le blackout ibérique du 28 avril 2025 déjà mentionné rappelle aussi que la géographie fonctionnelle du calcul ne peut pas être pensée seulement en termes de performance informatique. Les charges IA qui soutiennent des services critiques - supervision industrielle, cybersécurité, infrastructures énergétiques, transports, santé, administration - doivent être instruites selon leur dépendance à la continuité électrique, à la latence et à la résilience locale. L'enjeu est de distinguer ce qui peut être déplacé, différé, routé ou interrompu de ce qui doit rester disponible, redondé et gouverné près de l'usage.

Les charges d'entraînement, plus différables et moins sensibles à la latence, peuvent être orientées vers des sites où l'énergie, le foncier, le refroidissement et les contreparties territoriales sont les mieux maîtrisés. Les charges d'inférence, plus continues et plus sensibles aux usages, peuvent relever de sites plus distribués, de capacités régionales, d'*edge* ou d'infrastructures sous gouvernance spécifique lorsqu'elles touchent des services publics, industriels ou critiques.

L'alternative vise à ne pas accueillir la puissance de calcul comme un bloc homogène, mais à qualifier ses fonctions, ses temporalités, ses dépendances, ses impacts territoriaux et ses contreparties.

Le *data center* peut être pensé autrement que comme une charge à raccorder, à condition de l'inscrire dans une gouvernance énergétique, territoriale et numérique explicite.

La géographie fonctionnelle du calcul permet de sortir de l'expression trop vague « *data center IA* ». Elle oblige à distinguer entraînement, post-entraînement, évaluation et sécurisation, inférence, inférence critique, *fine-tuning* et *edge* industriel, car ces charges n'ont ni les mêmes contraintes de latence et de temporalité, ni les mêmes besoins de puissance, ni les mêmes exigences de souveraineté, ni les mêmes marges de déplacement temporel ou géographique.

Cet interlude referme la Partie I sur l'idée que la France ne doit pas seulement choisir où accueillir la puissance de calcul, mais comment la qualifier, l'orchestrer et la gouverner.

La Partie II déplace la question vers la demande elle-même : comment dimensionner les usages IA pour éviter que l'infrastructure ne soit tirée par défaut vers toujours plus de calcul ?

PARTIE II

La frugalité : reprendre la main sur la demande IA

Chapitre 5 - L'IA frugale : consommer juste

Le chapitre en une phrase

La frugalité ne commence pas au moment d'optimiser un modèle mais au moment de qualifier un usage. Le chapitre distingue IA consommatrice et IA utile, détaille les déclencheurs réels de la frugalité (coûts, souveraineté, sécurité, performance) et introduit le triptyque SLM + RAG + routeur comme architecture émergente, tout en soulignant la tension avec les modèles économiques dominants.

La Partie I a montré que la demande de data centers est massive, concentrée, territorialisée et souvent portée par des capitaux, des modèles, des clients et des architectures qui échappent en partie à la maîtrise européenne.

La Partie II déplace la question. Puisque la demande de calcul est là, comment la qualifier, la dimensionner et la gouverner ? Et surtout : par quels leviers la frugalité entre-t-elle réellement dans les organisations ?

Ce chapitre part du constat que, dans la pratique, l'IA frugale devient opératoire lorsqu'elle rencontre des contraintes déjà actives - coûts, souveraineté, sécurité, disponibilité des ressources, performance, robustesse, achats, usages métiers. Elle donne une cohérence à ces contraintes autour d'un principe : consommer juste, pour un usage qualifié, un bénéfice démontré, une architecture maîtrisée et des impacts suivis dans le temps.

Ce chapitre avance en trois temps : qualifier les usages avant de dimensionner les architectures ; identifier les contraintes qui rendent la frugalité opératoire dans les organisations ; puis décrire les premiers leviers de gouvernance qui permettent de passer d'une intention frugale à une pratique pilotable.

5.1 Ce que la frugalité n'est pas

Avant d'aller plus loin, trois mises au point permettent d'écarter des visions fréquentes mais insuffisantes sur l'IA frugale.

Trois définitions insuffisantes

1. **La frugalité n'entre pas nécessairement par l'environnement.** Si les considérations environnementales peuvent constituer un déclencheur à part entière, les entretiens montrent que la frugalité devient aussi opératoire par les coûts, la souveraineté, la latence, la sécurité, la disponibilité des ressources ou les contraintes opérationnelles. L'environnement peut alors apparaître comme une dimension à articuler à des contraintes déjà actives - coût, disponibilité, sécurité, souveraineté - plutôt que comme le déclencheur initial de la décision.
2. **L'IA frugale est de la responsabilité de la conception et du développement :** le développeur agit de plus en plus comme un assembleur de briques préconçues : bibliothèques, modèles pré-entraînés, frameworks, services cloud. L'essentiel de l'empreinte est souvent déterminé en amont par des choix de fournisseurs, d'architectures et de business models qui échappent au développeur final. La frugalité dépend certes de l'écoconception, mais aussi de toute la chaîne de valeur et de décision.
3. **L'IA frugale est une IA diminuée :** réduire mécaniquement les usages sans qualifier leur valeur n'est pas de la frugalité. La frugalité consiste plutôt à distinguer les usages nécessaires,

utiles, confortables, superflus ou rebond. Un usage IA qui apporte un bénéfice démontré, mesuré et supérieur à ses coûts directs et indirects n'est pas à écarter par principe, même s'il mobilise une infrastructure importante.

5.2 L'IA qui consomme juste

La formulation qui émerge du matériau d'entretiens peut être résumée ainsi :

Une IA frugale est celle qui consomme juste, pour un usage qualifié, un bénéfice démontré, une architecture maîtrisée, une robustesse suffisante et des impacts économiques, environnementaux, sociaux et humains maîtrisés et suivis dans le temps.

Cette formulation permet de sortir de deux impasses symétriques : la première serait de réduire la frugalité à une logique de restriction : faire moins, utiliser moins, demander moins. La seconde, la réduire à une optimisation technique : modèle plus petit, quantification, pruning ou baisse du nombre de tokens. Ces leviers comptent, mais ne suffisent pas.

Une IA peut être techniquement plus légère et pourtant non frugale si, par exemple, elle multiplie les usages à faible valeur ajoutée ou accroît la dépendance fournisseur.

La frugalité est donc avant tout une discipline d'arbitrage : choisir le bon usage, le bon modèle, la bonne donnée, la bonne infrastructure, le bon moment d'exécution et le bon niveau de performance ; maîtriser les dépendances ; mesurer les résultats économiques, environnementaux, sociaux et humains ; puis piloter ces résultats dans la durée.

5.3 La distinction fondatrice : IA consommatrice versus IA utile

Les entretiens font ressortir une distinction structurante entre IA consommatrice et IA utile.

L'idée est que, avec l'IA générative des grands modèles, tout l'équipement nécessaire aux IA utiles est largement financé, tiré et rendu disponible par les usages plus consommateurs de l'IA, notamment les usages génératifs de masse : assistants conversationnels, génération de contenu, moteurs de recommandation, copilotes généraux non bornés.

« La médiane de la distribution du potentiel reste au-dessus des coûts, mais il faut prendre en compte les effets rebonds. La question est : consommatrice versus utile ? Tout l'équipement des IA utiles est financé par les IA consommatrices », entretien avec un expert des projections.

Cette tension oblige à poser une question de proportionnalité : **quelle puissance de calcul accepte-t-on de mobiliser pour quel type de valeur produite ?** Cette question émerge en ce moment avec l'explosion des coûts de l'IA, très relayés dans la presse française et internationale.

Cette distinction oblige à qualifier l'usage en amont. Un usage à forte valeur collective - optimisation d'un réseau électrique, maintenance prédictive d'un équipement critique, aide au diagnostic médical, prévision de production solaire - ne s'instruit pas comme un usage récréatif, marketing ou conversationnel de confort. **La frugalité commence avant l'architecture, au niveau de la qualification de l'usage.**

Encadré - Outil : qualifier un usage de l'IA avant de l'instruire

Cinq types d'usage ont émergé de nos échanges et réflexions. Les quatre premiers se positionnent selon deux axes : la « valeur produite » et les « ressources mobilisées pour l'obtenir ». Le cinquième, le rebond, n'est pas une position sur ces axes mais une dynamique qui peut traverser n'importe lequel des quatre autres, y compris un usage par ailleurs nécessaire. Plus la valeur produite est élevée et les ressources mobilisées proportionnées, plus l'usage peut être instruit favorablement. À l'inverse, une valeur faible associée à des ressources importantes appelle une limitation, une requalification ou un abandon.

Type d'usage	Caractéristique	Exemple	Question de qualification	Risque si mal qualifié	Réponse d'instruction
Nécessaire ⁸¹	Pas d'alternative efficace hors IA	Aide au diagnostic médical d'urgence, prévision de production solaire	Existe-t-il une alternative non-IA aussi efficace ?	Dépendance non interrogée	Autoriser, sécuriser et dimensionner au juste niveau.
Utile	Bénéfice démontré supérieur au coût mobilisé	Maintenance prédictive d'un équipement critique, optimisation d'un réseau électrique	Le gain net est-il mesuré, effets rebonds compris ?	Adoption sans preuve de gain net	Exiger la mesure du gain net avant généralisation (voir 10.3)
Confortable	Gain marginal, alternative disponible	Copilote généraliste pour des tâches déjà bien outillées	Le gain justifie-t-il les ressources mobilisées plutôt que l'alternative existante ?	Transformation du confort en nécessité	Conditionner, limiter ou différer
Superflu	Valeur non démontrable	Génération en série de contenus, résumés ou présentations qui ne répondent pas à un besoin métier	La valeur produite est-elle démontrable ?	Consommation sans contrepartie	Écarter ou requalifier
Rebond	Dynamique transversale, peut apparaître dans n'importe lequel des quatre usages ci-dessus, y compris un usage nécessaire	Un diagnostic assisté par IA qui multiplie les examens complémentaires prescrits par excès de prudence	L'usage, une fois validé, engendre-t-il ailleurs une demande supérieure à l'économie initiale ?	Effet Jevons masqué par la légitimité de l'usage initial	Mesurer l'effet net dans la durée, y compris pour les usages déjà autorisés

⁸¹ Il faut préciser « nécessaire au regard d'un objectif préalablement qualifié ». « Nécessaire » ne signifie pas que la finalité de l'usage est bonne en soi. Cela signifie qu'une fois l'objectif légitimé par l'organisation ou la collectivité compétente, l'IA peut apparaître difficilement substituable pour l'atteindre. La qualification de la finalité et celle du moyen IA doivent rester distinctes.

Cette classification propose une grille de priorisation permettant d'arbitrer les investissements et les dépenses. Elle n'a pas vocation à figer les usages. Certains restent durablement intermédiaires ou incertains : leur valeur peut être plausible sans être démontrée, leurs effets dépendre fortement du contexte ou leur qualification évoluer avec les usages. Cette zone grise appelle moins une nouvelle catégorie qu'un régime d'instruction renforcé : expliciter les hypothèses, définir le périmètre de mesure, documenter les incertitudes et prévoir une réévaluation.

Le rebond peut prendre plusieurs formes. Il est **direct** lorsque le gain d'efficacité, la baisse de coût ou la facilité d'usage conduit à multiplier le même usage ; il est **indirect** lorsque les ressources économisées, notamment en argent, en temps ou en capacité, sont réaffectées à d'autres usages. À ces rebonds directs et indirects peuvent s'ajouter des effets plus larges, liés aux transformations des pratiques, aux mécanismes économiques et aux conditions structurelles. L'ITU-T L.1480 distingue notamment un rebond à l'échelle de l'économie, lorsque les gains d'efficacité contribuent à accroître la productivité, l'activité et la consommation globales. Ces effets sont particulièrement difficiles à attribuer à une solution isolée, mais ils doivent rester visibles dans l'interprétation des gains.

Cette distinction permet aussi d'éviter une confusion fréquente entre AI for Green et Green AI. Une IA peut être utilisée au service de l'environnement - optimiser un réseau, prévoir une production renouvelable, réduire une consommation thermique - sans être elle-même frugale dans sa conception, son architecture ou ses effets induits. Inversement, une IA peut être bien dimensionnée, sobre, robuste et gouvernable sans avoir une finalité environnementale directe.

Cette distinction entre usage nécessaire, utile, confortable, superflu et rebond rejoint la grille substitution / relocalisation / addition / effet d'appel présentée en 1.4.

La frugalité oblige à instruire simultanément la finalité de l'usage, les moyens mobilisés pour le produire et les effets et impacts induits.

L'efficacité unitaire ne garantit donc jamais, à elle seule, la sobriété globale : un usage mieux optimisé peut devenir plus fréquent, plus automatisé ou plus largement diffusé. C'est pourquoi l'effet rebond doit être suivi comme une dynamique transversale, y compris pour les usages nécessaires ou utiles.

5.4 Les déclencheurs réels de la frugalité

La frugalité IA entre rarement dans les organisations par la conscience environnementale. Elle entre plus souvent par d'autres contraintes : coût, souveraineté, sécurité, disponibilité des ressources ou performance, dépendance fournisseur, qui peuvent produire in fine des effets convergents avec l'objectif environnemental. Voici les déclencheurs que nous avons relevés :

Déclencheur réel	Comment il produit de la frugalité	Exemples issus des entretiens
Coût des tokens	Facturer l'usage IA à l'unité rend visible ce qui était invisible. Les budgets tokens contraignent naturellement la sur-consommation	Compteurs de tokens, alertes à un seuil, budgets par utilisateur, différenciation des accès par métier
« Par exemple, la comptabilité n'a pas accès à la vidéo. Le marketing a accès à la vidéo. On a des compteurs basés sur des tokens. Par exemple 5000 tokens par semaine. Avertissement à 4000. Pas de blocage pour l'instant », nous détaille un interviewé d'une organisation industrielle. ⁸²		

Déclencheur réel	Comment il produit de la frugalité	Exemples issus des entretiens
Disponibilité GPU	La rareté des GPU impose de choisir. On ne peut pas tout faire tourner en même temps	Plusieurs interviewés anticipent une montée en puissance des modèles plus légers (SLM) exécutés sur leurs propres infrastructures, par souci de maîtrise des coûts (voir 5.6).
Souveraineté et confidentialité	Les données sensibles ne peuvent pas partir dans un modèle étranger. Cela pousse vers des SLM locaux, du on-premise, de l'edge	Choix de l'edge pour des raisons de confidentialité des données client
« Les clients les plus informés souhaitent de l'edge pour des raisons de confidentialité des données client », entretien avec un fournisseur d'équipements énergétiques.		
Latence et performance	Certains usages temps réel ne supportent pas les temps de réponse des grands modèles distants	Routage vers un modèle local, inférence embarquée
« Faire une première analyse du prompt pour routeur d'inférence - tâche code, tâche analyse documentaire... puis faire le meilleur routage. [...] La gestion à venir c'est : quelle est la position géographique des utilisateurs ? Urgence/importance de la requête ? Service client géographique ? », explique un interviewé.		
Sécurité	Les systèmes critiques (réseaux, nucléaire, barrages) imposent explicabilité, reproductibilité et validation humaine, ce qui contraint les modèles	L'exigence de robustesse, soulignée par plusieurs interviewés (voir 5.9), peut conditionner l'adoption de la frugalité dans ces environnements.
Dépendance fournisseur	Être captif d'un grand modèle propriétaire crée un risque économique et opérationnel. Pour s'y soustraire, une tendance naissante vers l'open source et l'agnosticité	Certains acteurs revendiquent une architecture agnostique au modèle pour rester maîtres de leurs choix techniques (voir 5.12). « On met ce qu'on veut derrière », entendu plusieurs fois.

Cette liste n'invalide pas l'argument environnemental. Elle dit simplement par où la frugalité entre réellement, ce qui est une information stratégique majeure pour toute politique publique ou organisation cherchant à la diffuser.

Attendre uniquement que les acteurs soient convaincus par l'empreinte carbone de leurs prompts risque de produire peu d'effets. En revanche, **activer les contraintes économiques, de souveraineté, de sécurité et de performance, puis les relier explicitement à l'impact environnemental est une stratégie qui peut fonctionner.**

« IA frugale by design - sobriété dans l'utilisation, cas d'usage « juste ce qu'il faut ». Éviter de foisonner, être innovants, être sobres. Plus exigeants dans les estimations de gains. », entretien avec une organisation énergétique.

⁸² Avec l'augmentation des coûts de l'IA dans les organisations, ce mécanisme peut devenir une pratique généralisée.

Frugalité et chaîne de valeur sociale : un angle encore peu instruit

La transformation des métiers et des compétences est désormais entrée dans les préoccupations des directions, en particulier la capacité de l'organisation à fonctionner, se transformer et conserver son jugement métier. Son instruction reste toutefois moins structurée que celle des coûts, de la sécurité ou des architectures. Les entretiens montrent que les organisations commencent à relier IA, coûts, sécurité, confidentialité et performance, mais qu'elles peinent encore à intégrer la chaîne de valeur sociale : conditions de travail, redistribution des gains, montée en compétences, dépendance aux fournisseurs, transformation des métiers, qualité du travail, supervision humaine et effets sur les prestataires.

La question qui se pose n'est plus seulement de former quelques utilisateurs à de nouveaux outils, mais de déterminer comment chaque métier évolue, quelles compétences doivent être diffusées, lesquelles doivent rester maîtrisées en interne et comment préserver la capacité collective à contrôler, corriger ou reprendre les tâches confiées à l'IA.

Ce point est pourtant central. Une IA peut être frugale du point de vue des tokens et du modèle, tout en déplaçant des coûts humains : surcharge de contrôle, perte de compétences, dépendance accrue à des outils opaques, automatisation de tâches auparavant réalisées par des équipes internes ou des prestataires, intensification du travail, difficulté à attribuer les gains de productivité. À l'inverse, une IA bien dimensionnée peut aussi soutenir une innovation sociale : réduction de tâches répétitives, aide à la décision, montée en compétences, amélioration de la qualité, meilleure accessibilité ou sécurisation de processus critiques.

La frugalité devrait donc intégrer une question supplémentaire : **qui supporte l'effort d'optimisation, qui bénéficie des gains et quelles capacités humaines sont renforcées ou affaiblies ?** Sans cette lecture sociale, la frugalité risque de rester une discipline de réduction des coûts techniques et environnementaux uniquement. Avec elle, **elle devient une méthode d'innovation responsable : arbitrer non seulement la consommation de calcul, mais aussi la qualité de la chaîne de valeur humaine qu'elle transforme.**

Cette dimension sociale ne se pilote pas de la même manière selon la maturité IA de l'organisation.

5.5 Quatre profils d'organisations face à l'IA

Une entreprise qui subit l'IA intégrée par ses éditeurs n'a pas les mêmes leviers qu'une organisation qui orchestre déjà plusieurs modèles, route ses requêtes et contractualise ses exigences fournisseurs.

Profil	Ce que l'organisation fait	Risque principal	Levier frugal prioritaire
IA subie	Attend que les éditeurs intègrent l'IA dans les logiciels existants	Perte de visibilité sur l'inférence et dépendance fournisseur	Clauses contractuelles, reporting, désactivation sélective
IA exploratoire	Utilise prompts, copilotes, LLM publics ou API	Explosion des usages non qualifiés	Budgets tokens, règles d'usage, tri des cas
IA structurée	Développe un socle data, RAG, workflows métiers	Difficulté d'intégration métier et de preuve de valeur	Scénario sans IA, indicateurs de bénéfice net
IA gouvernée	Met en place SLM, RAG, routeur, architecture agnostique	Complexité de pilotage et arbitrages multi-critères	Gouvernance modèle, routage, mesure, réversibilité

5.6 Le triptyque de l'architecture frugale émergente : SLM + RAG + routeur

Les entretiens font émerger une architecture récurrente chez les acteurs les plus avancés. C'est une tendance de fond actuelle qui dessine la direction de la frugalité opérationnelle.

Le triptyque SLM + RAG + routeur

SLM (Small Language Model) : modèle de langage de petite taille, entraîné ou affiné sur un domaine métier spécifique. Selon sa taille, son niveau d'optimisation et les exigences de latence, il peut parfois tourner sur des architectures CPU ou hybrides CPU/GPU. Il est plus rapide, moins cher à l'inférence, plus facilement explicable et peut être déployé localement (on-premise ou edge), ce qui résout simultanément les problèmes de confidentialité, de latence et de souveraineté.

RAG (Retrieval-Augmented Generation) : technique qui permet d'adapter un modèle général à un contexte spécifique sans ré-entraîner. Le modèle est enrichi par une base documentaire ou de données métiers interrogées en temps réel. On évite un ré-entraînement coûteux et on contextualise la réponse sans surdimensionner le modèle.

Routeur de modèles : couche intermédiaire qui analyse la requête et l'oriente vers le modèle le moins coûteux et suffisamment performant pour la tâche. À la différence d'un orchestrateur, qui coordonne une chaîne plus large de modèles, de données, d'outils, de règles ou d'agents, le routeur réalise principalement un choix entre plusieurs modèles. Le routeur peut répondre « je ne sais pas », éviter d'appeler inutilement un modèle plus lourd et différencier une question factuelle simple d'une analyse complexe. L'utilisateur n'a pas à être expert en sobriété.

Ce modèle est en cours de déploiement dans certaines grandes entreprises françaises, qui mettent en place un routeur d'inférence en complément de leur modèle par défaut : « *le système analyse chaque prompt pour l'orienter vers le modèle le moins coûteux et suffisamment performant pour la tâche : tâche de code, tâche documentaire, analyse complexe* », entretien avec un responsable du numérique responsable.

Voir aussi 5.12 « architecture agnostique ».

Aux yeux de ces acteurs, ce triptyque répond simultanément à trois enjeux : la frugalité (modèle proportionné), la souveraineté (déploiement local, données non transmises à un tiers) et la performance (contextualisation métier, latence réduite). C'est précisément ce cumul qui explique son adoption croissante.

Cette orientation est désormais aussi portée au niveau institutionnel : le Sénat recommande en septembre 2026 un soutien public aux petits modèles d'IA spécialisés.

« *Demain, plusieurs tendances en cours : SLM va prendre une grosse part du gâteau. On va avoir Mistral sur nos GPU/CPU. Réduction de la taille des modèles, SLM. Densification : prendre les modèles, faire passer les algorithmes, diminuer la consommation en diminuant l'accuracy.* », précise un opérateur interviewé.

La réduction évoquée peut notamment passer par la quantification des modèles, ou quantization en anglais, qui diminue la précision numérique utilisée pour représenter leurs paramètres afin de réduire la mémoire mobilisée et le coût de l'inférence.

Ce déplacement du seul critère de performance n'est acceptable que s'il reste compatible avec le niveau de qualité, de sécurité et de robustesse requis par l'usage.

La **frugalité** se joue donc de plus en plus dans l'**orchestration multimodèle** et non dans le choix unique d'un grand modèle universel.

En pratique, cette bascule transforme la frugalité en système de décision distribué.

Chaque couche peut contribuer à la proportionnalité de la consommation :

- le métier qualifie l'usage ;
- la donnée limite le volume et la sensibilité des informations mobilisées ;
- le RAG évite de ré-entraîner inutilement ;
- le SLM limite le recours aux modèles lourds ;
- le routeur oriente la requête vers le modèle suffisant ;
- l'infrastructure choisit le lieu, le matériel et le moment d'exécution ;
- la gouvernance fixe les budgets, les seuils et les droits d'accès ;
- les achats imposent la transparence fournisseur ;
- les métiers vérifient le bénéfice réel dans la durée.

La frugalité n'est donc pas portée par une seule brique, mais par l'alignement de toute la chaîne de valeur et de décision.

De la frugalité par pertinence à la fuite en avant de l'automatisation

La trajectoire SLM + RAG + routeur peut être renforcée par l'orchestration et le context engineering, à condition que ces termes restent au service de la pertinence.

L'orchestration consiste à organiser les interactions entre modèles, bases de données, API, outils métiers, règles de décision, agents éventuels et validation humaine. Elle est frugale lorsqu'elle évite d'appeler un grand modèle pour chaque tâche et qu'elle oriente le calcul vers la ressource

suffisante.

Le context engineering consiste à construire le bon contexte donné au modèle : documents pertinents, données fiables, règles métier, contraintes de sécurité, historique, formats de sortie et limites d'usage. Il est frugal lorsqu'il évite de demander au modèle de tout deviner et réduit le recours à une puissance brute mal ciblée.

Un cas industriel, comme une ligne de pyrolyse, illustre cette logique. La valeur de l'IA ne vient pas nécessairement d'un modèle plus puissant, mais de la qualité du contexte qui lui est fourni : données de procédé, règles métier, schémas de tuyauterie et d'instrumentation (P&ID), supervision, contraintes de sécurité, historique d'exploitation, documentation technique, paramètres issus d'outils de simulation et validation humaine. La frugalité vient alors de la pertinence du contexte plus que de la puissance du modèle.

Mais cette trajectoire est fragile. Elle peut être recouverte par un nouvel emballement lexical et commercial : agents, *brain* de l'entreprise⁸³, *intelligence liquidity*⁸⁴, copilotes métiers, automatisation de bout en bout. Ces termes ne désignent pas tous la même chose, mais ils portent une promesse commune : rendre l'intelligence disponible partout, branchée sur toutes les données, tous les outils et toutes les décisions. Cette promesse peut faire basculer les organisations de la recherche de pertinence vers une fuite en avant d'automatisation.

Derrière l'imaginaire d'efficacité, cette fuite en avant peut devenir coûteuse et peu performante en pratique : multiplication des appels de modèles, chaînes d'agents difficiles à contrôler, connecteurs à sécuriser, exposition de données sensibles, responsabilités diluées, résultats difficiles à vérifier, perte de compétences métier, dépendance accrue aux fournisseurs et complexité de maintenance. L'automatisation paraît fluide, mais elle peut produire une intelligence plus opaque, plus coûteuse et plus difficile à gouverner.

La prochaine bataille de la frugalité ne se jouera donc pas seulement entre grands et petits modèles, mais entre deux formes d'orchestration : l'orchestration de pertinence, qui réduit le calcul nécessaire à un usage qualifié et l'orchestration d'automatisation, qui entretient la fuite en avant vers une intelligence liquide, disponible partout, mais toujours plus difficile à maîtriser.

Cette bifurcation confirme que le triptyque SLM + RAG + routeur ne suffit pas à garantir la frugalité. Il ouvre une possibilité, mais cette possibilité doit être instruite, mesurée et gouvernée.

Ce triptyque a aussi des limites : l'efficacité réelle du SLM doit être vérifiée en situation d'usage et dans la durée ; la performance doit rester compatible avec les exigences métiers ; les gains unitaires peuvent être annulés par des effets rebond volumétriques, organisationnels, applicatifs ou sectoriels ; l'outillage ne suffit pas sans implication des métiers, gouvernance des usages, suivi dans la durée et cohérence stratégique.

⁸³ **Brain de l'entreprise** : expression utilisée pour désigner une architecture visant à rendre les connaissances, les données et les outils d'une organisation accessibles à des systèmes d'IA capables de les interroger, de les relier et, éventuellement, d'agir à partir d'eux. Par exemple, un assistant interne connecté simultanément à la documentation, au CRM, aux données métiers et aux applications de l'entreprise.

⁸⁴ **Intelligence liquidity** : expression désignant l'idée d'une intelligence artificielle rendue disponible de façon fluide dans l'ensemble des outils, processus et interfaces d'une organisation, comme une ressource mobilisable à la demande. Par exemple, une même capacité IA accessible depuis la messagerie, le CRM, l'ERP, les outils métiers ou des agents automatisés. Ces expressions relèvent davantage du vocabulaire émergent de l'industrie et du marketing technologique que de catégories techniques stabilisées.

5.7 Les budgets tokens et les agents

Les budgets tokens rendent visible ce qui était invisible. Ils sont la première forme opérationnelle de cette gouvernance pratique de la frugalité dans les organisations.

Concrètement, ils peuvent prendre la forme de compteurs par utilisateur, de seuils d'alerte, de quotas différenciés par métier, de budgets par équipe ou de restrictions d'accès à certains modèles ou fonctionnalités.

Cette frugalité est d'abord économique avant d'être environnementale. Elle entre dans l'organisation par les budgets, les quotas, les licences et la qualité de service. Elle constitue aussi une opportunité : les mêmes mécanismes qui réduisent les coûts tokens peuvent réduire une partie de la consommation de calcul, à condition qu'ils s'accompagnent d'un routage adapté des modèles et d'une mesure des caractéristiques environnementales de l'infrastructure mobilisée.

Les agents : du budget *tokens* au coût complet de la tâche⁸⁵

L'essor des agents IA change l'économie de la frugalité. Un assistant conversationnel répond à une requête ; un agent peut enchaîner des appels à plusieurs modèles, interroger des bases, utiliser des outils, produire des brouillons, relancer des vérifications, appeler des API et déclencher des actions. La consommation ne se mesure plus seulement en tokens par prompt, mais en **coût complet d'une tâche automatisée**.

Cette bascule rend les budgets tokens nécessaires mais insuffisants. Il faut aussi suivre le nombre d'étapes, les appels outils, les modèles mobilisés, les boucles de vérification, les erreurs, les reprises humaines, les permissions accordées et les tâches réellement finalisées. Un agent mal borné peut produire un effet rebond puissant : parce qu'il rend l'action plus facile, il augmente le nombre de tâches déclenchées, automatise des processus peu utiles ou multiplie les appels invisibles à des services externes.

La frugalité agentique suppose donc des garde-fous spécifiques :

- objectifs bornés,
- budget par tâche,
- conditions d'arrêt,
- journalisation des actions,
- validation humaine sur les décisions sensibles,
- choix du modèle suffisant à chaque étape,
- limitation des appels externes,
- traçabilité des données transmises et mesure du bénéfice net.

L'agent n'est frugal que s'il remplace une chaîne de travail coûteuse ou risquée par une chaîne plus efficace, plus maîtrisée et mieux gouvernée, pas simplement parce qu'il automatise davantage.

5.8 Dix niveaux d'arbitrage de la frugalité

La frugalité se joue dans une chaîne d'arbitrages. À chaque niveau, une mauvaise décision peut annuler les gains du niveau précédent : un bon modèle pour un mauvais usage, une bonne

⁸⁵ Voir par exemple https://www.lemonde.fr/economie/article/2026/06/12/les-banques-rationnent-l-usage-de-l-ia-terrifiees-par-les-couts-qu-elle-engendre_6701070_3234.html

architecture sur une infrastructure opaque, une mesure précise sans gouvernance, un usage utile qui déclenche un effet rebond.

Les entretiens ont permis de distinguer dix niveaux d'arbitrage. Pour chacun, nous avons formalisé une question centrale et des leviers spécifiques.

Niveau	Question d'arbitrage	Leviers principaux
1. Usage	L'IA est-elle nécessaire, utile ou superflue ? Voir 5.3.	Qualification du besoin, scénario sans IA, bénéfice attendu, effet rebond
2. Données	Les données sont-elles pertinentes, triées, suffisantes ?	Sélection, nettoyage, réduction des <i>datasets</i> bruités, documentation des biais
3. Modèle	Le modèle est-il proportionné ?	SLM, <i>pruning</i> , <i>quantization</i> , <i>fine tuning</i> , modèle suffisant plutôt que maximum
4. Routage	La requête va-t-elle vers le bon modèle ?	Routeur d'inférence, classification de la requête, réponse « je ne sais pas »
5. Prompt	La requête est-elle calibrée ?	Compteurs <i>tokens</i> , regroupement de requêtes, bonnes pratiques de formulation
6. Infrastructure	Où et comment tourne l'IA ? Quelles caractéristiques de l'infrastructure ?	<i>Cloud</i> , <i>on-premise</i> , <i>edge</i> , souveraineté, perte de visibilité en SaaS cognitif
7. Temporalité	Peut-on déplacer le calcul ?	Cloche solaire (voir description en 6.5), prévisions 72h, tâches batch déplaçables, <i>carbon-aware computing</i> ⁸⁶
8. Matériel	Quelles caractéristiques du matériel ? Peut-on faire durer les équipements ?	Durée de vie GPU, réemploi, <i>retrofit</i> , évitement de l'obsolescence prématurée
9. Gouvernance	Qui décide, mesure et limite ?	Budgets <i>tokens</i> , accès différenciés, cadre d'usage, achats, clauses fournisseur
10. Bénéfice net	Le bénéfice réel, mesuré dans la durée, justifie-t-il les impacts directs, indirects et induits ?	ROI, efficacité d'usage, scénario sans IA, mesure du gain net, effets rebond. Voir 5.3

Ces dix points constituent un cadre d'instruction, et non une checklist dont tous les critères pourraient être optimisés simultanément. Des tensions peuvent apparaître entre performance, sécurité, réversibilité, coût, sobriété, souveraineté ou rapidité ; elles doivent être rendues explicites et arbitrées. La grille est donc analytique et illustrative, sans prétendre à l'exhaustivité ni à l'existence d'une solution unique qui maximiserait l'ensemble des critères. Elle permet de montrer que la frugalité se joue tout au long de la chaîne : de la qualification du besoin au choix du matériel, en passant par la gouvernance et la mesure du bénéfice réel.

Cette grille peut aussi se lire comme une chaîne d'arbitrage en trois familles, complétée par un test transversal de bénéfice net :

1. **Arbitrer l'usage** : usage, qualification (nécessaire, utile, confortable, superflu) effet rebond

⁸⁶ Voir 6.5 pour plus d'éléments sur le *carbon-aware computing*.

(voir 5.3)

2. **Arbitrer l'architecture** : données, modèle, routage, prompt, infrastructure.
3. **Arbitrer l'exploitation** : temporalité, matériel, gouvernance.

Test transversal - Bénéfice net : vérifier que le bénéfice réel, mesuré dans la durée, justifie les impacts directs, indirects et induits.

Cette grille doit toutefois être **complétée par trois dimensions transversales** :

1. **la durée** : un arbitrage frugal ne se juge pas seulement au moment du déploiement, mais dans le temps : évolution des volumes, renouvellement matériel, dérive des usages, maintenance, obsolescence, dépendance fournisseur.
2. **le social et l'humain** : qui gagne du temps, qui perd de l'autonomie, qui supervise, qui porte la responsabilité, quelles compétences sont renforcées ou affaiblies ? (voir encadré en 5.4)
3. **les comportements** : un système plus simple, plus rapide ou moins coûteux peut déclencher davantage d'usages, abaisser les seuils d'automatisation et déplacer la norme d'usage.

La frugalité ne peut donc pas être seulement une optimisation technique, elle doit suivre les effets dans la durée, sur les métiers et sur les comportements collectifs.

5.9 La robustesse comme condition : l'IA frugale dans les systèmes critiques

Dans les environnements industriels et énergétiques, la définition de la frugalité intègre une dimension supplémentaire : **la robustesse**. Un modèle plus petit mais insuffisamment robuste, explicable ou gouvernable peut être risqué.

Les dispatcheurs des organisations énergétiques, les ingénieurs de centrales nucléaires, les opérateurs de réseau peuvent résister à l'intégration de l'IA dans leurs processus critiques, non pas par conservatisme, mais parce que les enjeux de sécurité et de réglementation imposent une prudence légitime.

« Le système ne peut pas se tromper. Il faut de la robustesse. Pour un réacteur ou un barrage, l'explicabilité et la reproductibilité sont des conditions sine qua non, pas des options. », entretien avec une organisation énergétique.

L'IA frugale dans les systèmes critiques pourrait ainsi être vue comme : le **plus petit système suffisamment performant, robuste, explicable et gouvernable pour l'usage visé**.

Cette définition évite une erreur fréquente : confondre frugalité et légèreté. Un modèle léger mais opaque, non reproductible ou impossible à valider humainement n'est pas frugal dans un système où les conséquences d'une erreur sont critiques, car il ne répondrait pas aux exigences du métier.

Un SLM frugal parfaitement adapté à son contexte industriel n'est donc pas seulement maîtrisé du point de vue environnemental et économique. Il est maîtrisé techniquement, explicable, auditable, réversible, validé par les opérateurs et intégré dans un système de supervision humaine. C'est ainsi que nous pouvons aborder la frugalité industrielle.

5.10 La tension structurelle : business models vs frugalité

L'un des enseignements des entretiens est que la frugalité IA ne se diffusera pas spontanément. Elle se heurte à une tension structurelle entre la logique économique des modèles généralistes et la

logique de proportionnalité propre à la frugalité. Cette tension sera reprise au chapitre 7 sous l’angle des incitations, des architectures et des achats.

Le même constat vaut pour le développeur, dont une partie décisive de l’empreinte est déjà inscrite dans les briques qu’il assemble. Ce point sera instruit au chapitre 7.4.

« *C'est difficile de faire de la frugalité ; on n'en fait pas. Les développeurs peuvent être conscients, mais pris dans d'autres priorités.* », entretien acteur numérique.

Ce constat révèle la réalité opérationnelle : la frugalité existe dans des communautés de pratique, chez des pionniers, dans des expériences isolées. Mais elle n'est pas encore une norme généralisée, même lorsque les acteurs en comprennent les enjeux. La raison structurelle est que, sans critères d'achat, sans exigences de *reporting*, sans contraintes architecturales et sans incitations, **la frugalité reste une conviction interne sans effet systémique.**

5.11 La frugalité comme critère d’achat : ce qui ne se diffusera pas autrement

Cette section introduit un levier qui sera approfondi au chapitre 7 : la frugalité ne se diffusera pas seulement par les pratiques internes, mais par les critères d’achat, les contrats, les clauses fournisseurs et les exigences de *reporting*.

Certains entretiens suggèrent qu’elle soit intégrée dans les achats, les contrats, les SLA et les exigences de *reporting*.

Les collectivités, les hôpitaux, les industriels et les services publics qui achètent des services IA peuvent poser des exigences de frugalité et créer ainsi un signal de marché que les fournisseurs seront contraints de respecter.

Critères utiles	Ce qu’il exige	Pourquoi c’est faisable maintenant
Empreinte de l’usage	Quel modèle est appelé ? De quelle taille ? Hébergé où ?	Des outils existent (EcoLogits, CodeCarbon). Il manque l’obligation
Localisation d’exécution	Où tourne la requête ? Quel mix électrique ?	Pertinent pour la souveraineté autant que pour l’empreinte
Options SLM/RAG disponibles	L’offre propose-t-elle des alternatives frugales ?	Nombreux acteurs capables de proposer du SLM souverain
Reporting tokens/énergie/eau	Données d’usage disponibles pour le client ?	Condition de tout pilotage. Encore trop rare
Réversibilité	Peut-on changer de modèle ou de fournisseur ?	Architecture agnostique comme critère d’achat
Durée de vie matérielle	Quelle politique de renouvellement des équipements ?	Faire durer les GPU est un levier peu valorisé
Coût complet d’usage	Coût par requête, par utilisateur, par cas d’usage, mais aussi coût d’intégration, de supervision, de maintenance et de montée en charge	Les directions achats, DSI et FinOps disposent déjà de méthodes de suivi budgétaire ; il faut les adapter aux usages IA

Critères utiles	Ce qu'il exige	Pourquoi c'est faisable maintenant
Effets rebond	Le fournisseur ou l'intégrateur doit documenter les mécanismes susceptibles d'augmenter les volumes : automatisation, généralisation, usages induits, appels en cascade	Les volumes de requêtes, les logs, les <i>budgets tokens</i> et les scénarios d'usage permettent de suivre ces dérives
Impacts humains et sociaux	Identifier les tâches transformées, les besoins de formation, la supervision humaine, la responsabilité métier et les effets sur les prestataires	Ces éléments peuvent entrer dans les cahiers des charges, les analyses d'impact, les chartes d'usage et le dialogue social
Orchestration et agents	Documenter les modèles appelés, les outils utilisés, les chaînes d'action, les permissions, les conditions d'arrêt et les validations humaines.	Les journaux d'exécution, logs applicatifs, budgets par tâche et contrôles d'accès permettent déjà de suivre une partie de ces chaînes.

Ces critères ne garantissent pas à eux seuls une IA frugale, mais ils rendent la frugalité négociable et mesurable et peuvent, lorsqu'ils sont inscrits dans les marchés et les contrats, la rendre opposable.

Leur robustesse dépend de quatre conditions :

1. couvrir le coût complet,
2. anticiper les effets rebond,
3. intégrer les impacts humains et sociaux,
4. suivre les résultats dans la durée.

Un service IA qui ne peut pas documenter son modèle, sa localisation, ses alternatives, ses volumes d'usage et ses effets organisationnels enferme son client dans une dépendance non gouvernable.

5.12 L'architecture agnostique

Un dernier principe relie les architectures frugales, les achats et la souveraineté : l'agnosticité au modèle. Tant que l'organisation est enfermée dans un modèle unique ou un fournisseur unique, elle ne peut pas arbitrer finement entre performance, coût, confidentialité, impact environnemental et souveraineté.

Un exemple est celui d'une organisation énergétique : un outil conversationnel interne conçu comme agnostique au modèle. L'organisation peut mettre le modèle qu'elle veut derrière : appeler Gemini, utiliser du RAG avec les réseaux sociaux internes, consulter le web, basculer d'un modèle à l'autre selon l'usage, tout en restant maître de l'orchestration.

« IA conversationnelle est agnostique au modèle. On met ce qu'on veut derrière. On est maîtres, pas captifs. », selon un interviewé d'une organisation énergétique.

La promesse : l'agnosticité du modèle permet de choisir à tout moment le modèle le mieux adapté à la requête, au niveau de confidentialité requis, au coût acceptable et à l'empreinte désirée. Elle transforme la frugalité en capacité permanente d'arbitrage plutôt qu'en choix unique et irréversible

fait une fois à l'achat. Cependant, l'interlocuteur n'a pas présenté de mesures qui démontrent la frugalité réelle de cette agnosticité.

Ce chapitre conduit donc à la conclusion : la frugalité commence dès le moment de qualifier un usage. Elle se poursuit dans l'architecture, le routage, les budgets, les achats, la robustesse, la gouvernance et le suivi des effets dans le temps. Consommer juste, ce n'est pas seulement consommer moins ; c'est consommer pour un usage dont la valeur, les coûts complets et les effets induits peuvent être instruits.

Mais **pour arbitrer, il faut voir**. Or, dès que l'IA passe par le SaaS, les API ou les suites logicielles intégrées, la visibilité disparaît : modèle appelé, lieu d'exécution, énergie, eau, matériel mobilisé. Le chapitre suivant traite cette condition préalable : rendre l'inférence visible, puis gouvernable.

Chapitre 6 - Rendre l'inférence visible : condition de toute frugalité

Le chapitre en une phrase

Sans visibilité sur ce qui se passe réellement derrière un prompt - modèle appelé, lieu d'exécution, énergie et eau mobilisées - aucune frugalité n'est pilotable. Le chapitre distingue AI for Green et Green AI, examine les angles morts de la comptabilité carbone (Scope 3, *location-based* vs *market-based*) et identifie quatre conditions pour rendre l'inférence gouvernable : traçabilité, architecture observable, gouvernance interne, critères d'achat.

Le chapitre précédent a montré que la frugalité commence par la qualification des usages : consommer juste, pour un besoin identifié, un bénéfice démontré et une architecture maîtrisée. Mais cette exigence rencontre une limite majeure : une part croissante de la demande IA devient invisible pour les organisations qui l'utilisent.

L'inférence se cache dans les API, les suites logicielles, les agents, les services SaaS et les fonctionnalités intégrées aux outils du quotidien. En activant une option, une organisation déclenche parfois une chaîne de calcul dont elle ne connaît ni le modèle, ni le lieu d'exécution, ni l'énergie, ni l'eau, ni le matériel mobilisé.

Cette exigence de visibilité dépasse désormais la seule inférence. Le calcul mobilisé pour développer un modèle ne s'arrête pas à son pré-entraînement : post-entraînement, évaluations et sécurisation peuvent eux-mêmes mobiliser des volumes importants. Pour GPT-5.6, OpenAI⁸⁷ indique avoir consacré environ 700 000 heures GPU équivalent NVIDIA A100 au red teaming automatisé avant son lancement et précise que ces tests se poursuivent pendant le déploiement. OpenAI indique également avoir entraîné GPT-Red, son modèle spécialisé de red teaming, à une échelle de calcul comparable à certaines de ses plus grandes phases de post-entraînement. Ces données ne permettent pas d'en déduire directement une consommation électrique, faute d'informations suffisantes sur les équipements et conditions d'exploitation, mais elles montrent que l'empreinte d'un modèle se construit aussi dans les étapes successives de son développement, de son évaluation et de sa sécurisation.

Ce chapitre examine cette opacité. Il montre pourquoi elle est structurelle, ce que les outils de mesure actuels permettent ou ne permettent pas de voir et pourquoi la frugalité ne peut pas se limiter à une empreinte par requête. Rendre l'inférence visible devient une condition préalable : sans visibilité, il n'y a pas d'arbitrage ; sans arbitrage, il n'y a pas de frugalité gouvernable.

6.1 L'IA cachée : le vecteur de demande invisible

La demande IA ne vient pas seulement des projets explicitement nommés « IA ». Elle vient aussi et peut-être surtout des couches IA intégrées dans les logiciels du quotidien⁸⁸. Suites bureautiques, ERP, CRM, outils de *reporting*, outils ESG, outils de conception, PLM, BIM, GTB, logiciels industriels, plateformes collaboratives, copilotes de développement ou logiciels de supervision : tous ces logiciels s'enrichissent progressivement de fonctions IA sans décision explicite d'adoption à l'échelle de l'organisation. Les agents sont en quelque sorte leur état ultime actuel. Comme nous

⁸⁷ <https://openai.com/fr-FR/index/gpt-5-6/>

⁸⁸ Voir aussi « souveraineté des interfaces » traitée au chapitre 4, section 4.5.

l'avons dit au chapitre 4 dans notre encadré sur la souveraineté des interfaces : « *qui tient le poste de travail tient une partie de l'IA* »

Cette IA incorporée est stratégique parce qu'elle se diffuse sans décision visible, en transformant progressivement des logiciels déjà installés en consommateurs d'inférence. Un utilisateur qui demande une « synthèse automatique » dans son CRM déclenche une requête d'inférence sans savoir ni où elle tourne, ni sur quel modèle, ni avec quelle consommation.

« Le carbone n'est pas encore entré dans le monde de l'inférence. L'utilisateur consomme un abonnement, une licence, une réponse ou une fonctionnalité. Il ne consomme pas un signal énergétique intelligible. », entretien avec un acteur du numérique.

Les organisations ne peuvent pas mesurer ce qu'elles ne voient pas. Et elles ne peuvent pas gouverner ce qu'elles ne mesurent pas. La croissance réelle de la demande IA et de son empreinte peut donc être très supérieure à ce que les décisions explicites d'adoption laissent supposer, et à ce que les démarches de frugalité peuvent réellement piloter.

IA agentique : quand une requête devient une chaîne d'inférences

L'IA agentique pourrait accentuer l'opacité de l'inférence. Comme nous l'avons vu en 5.7, une requête utilisateur déclenche une chaîne d'actions : planification, appels à plusieurs modèles, consultation d'outils, recherche documentaire, vérifications, génération de sous-tâches et exécution partielle. L'impact ne se lit donc plus seulement au niveau du prompt initial, mais au niveau de la trajectoire complète d'inférence.

Le potentiel de l'agentique reste toutefois, à ce stade, largement ouvert. Son adoption progresse plus vite que notre capacité à en caractériser les coûts, les risques, les effets organisationnels et les conditions de robustesse à grande échelle. Il convient donc de distinguer les capacités démontrées dans des environnements bornés des promesses associées à une généralisation encore peu éprouvée. Ce point doit rester formulé comme un signal émergent : ses effets dépendront des architectures effectivement déployées, des garde-fous, des limites d'autonomie et des modèles économiques. Il justifie néanmoins d'intégrer dès maintenant la notion de « chaîne d'inférence » dans les exigences de mesure et de gouvernance.

6.2 La chaîne d'intermédiation : pourquoi l'opacité est structurelle

L'opacité de l'inférence est avant tout une propriété de l'architecture du marché. Entre l'utilisateur final et le data center, plusieurs couches s'interposent, chacune apportant de la performance ou de la simplicité, chacune éloignant du signal environnemental réel.

Couches de la chaîne d'intermédiation

Fournisseur cloud : héberge les serveurs, alloue les GPU, décide de la localisation effective d'exécution. Ses pratiques de reporting sont très variables selon les acteurs.

Éditeur logiciel : intègre l'IA dans ses produits via des API, souvent sans documenter le modèle utilisé ni la localisation d'exécution dans les contrats.

Modèle propriétaire : peut changer de version sans notification, ce qui peut modifier l'empreinte sans que le client en soit informé.

Plateforme d'orchestration : agrège plusieurs appels de modèles, distribue les charges entre fournisseurs selon la disponibilité ou le coût, rarement selon l'empreinte.

API : interface technique standardisée qui masque délibérément la complexité de l'infrastructure sous-jacente.

Mutualisation des infrastructures : un même serveur peut servir plusieurs clients simultanément. L'allocation de l'empreinte entre clients est une question méthodologique non résolue.

L'utilisateur, le développeur, le décideur et l'organisation peuvent perdre de la visibilité de la couche qui les concerne dans la chaîne matérielle et contractuelle. Cette architecture est le résultat logique de la conception des services cloud, **optimisés pour la simplicité d'usage plutôt que pour la transparence environnementale**.

« Opacité des cloud providers : on n'en fait pas grand-chose. Scope 3 oui/non. Market-based/location-based. On a les trois. En plus, aux USA la comptabilité est différente. On n'était pas sortis de ces sujets de bilans et on se prend l'IA dessus. », entretien avec un acteur industriel.

6.3 Ce que la mesure actuelle ne dit pas

Ce rapport étant centré sur l'IA et l'énergie et leur interdépendance, le carbone y est davantage documenté que d'autres dimensions environnementales, telles que l'épuisement des ressources. Une évaluation complète doit être multicritère et intégrer, selon les usages et les infrastructures, la consommation d'eau, l'épuisement des ressources minérales et métalliques, l'énergie primaire, les pollutions et déchets associés à la fabrication des équipements, ainsi que les autres impacts du cycle de vie. La disponibilité encore inégale de données comparables sur ces dimensions constitue elle-même une limite actuelle de la mesure.

Une approche multicritère ne consiste toutefois pas à additionner mécaniquement des indicateurs hétérogènes. Elle doit aussi rendre visibles les transferts d'impact : une solution peut réduire les émissions de GES tout en augmentant les besoins en eau, en métaux, en foncier ou en équipements. La frugalité ne peut donc être définie par l'optimisation d'un indicateur unique ; elle suppose d'explicitier les priorités, les seuils incompressibles et les arbitrages entre impacts.

Cette lecture peut également être étendue, lorsque les données le permettent, aux impacts sociaux des chaînes de valeur : conditions d'extraction et de fabrication, travail de production, d'annotation ou de modération des données, exposition des travailleurs, ainsi qu'aux asymétries géographiques entre les lieux d'extraction des ressources, de fabrication des équipements, de calcul et de captation de la valeur.

Des outils comme CodeCarbon, EcoLogits, Alumet ou Carbontracker peuvent contribuer à estimer certains impacts liés à l'entraînement ou à l'inférence, selon le niveau d'accès à l'infrastructure et le périmètre retenu. Cependant, ils ne résolvent pas l'opacité des architectures des modèles d'IA, SaaS et cloud propriétaires.

Ce que la mesure actuelle dit	Exemple de ce qu'elle ne dit pas
Consommation électrique estimée du modèle	Localisation réelle d'exécution à l'heure de la requête
Empreinte carbone estimée par requête	Mix électrique physique du data center au moment précis
Volume de tokens consommés	Taux d'utilisation réel des serveurs (impact sur l'efficacité)
Comparaison entre modèles en conditions de test	Durée de vie réelle des équipements matériels mobilisés
Émissions scope 2 déclarées par le fournisseur	Allocation entre clients sur serveurs mutualisés
Facteur d'émission moyen location-based	Eau consommée pour le refroidissement de ce calcul précis
Indicateurs PUE du data center	Empreinte scope 3 : fabrication des GPU, serveurs, équipements réseau, câbles, batteries et équipements techniques.

Un indicateur comme le PUE doit donc être lu avec prudence. Il indique une moyenne. Un PUE annoncé en phase de conception ne dit pas nécessairement le PUE réel observé en exploitation : niveau de charge, saison, densité des racks, redondance, incidents, maintenance et évolution des équipements peuvent modifier la performance effective. Selon l'importance de l'information, la doctrine de reporting devrait distinguer les indicateurs de conception, les indicateurs contractuels et les indicateurs mesurés en exploitation.

« PUE réel, on ne sait pas ce qu'on gagne ; PUE annoncé à 1,25, je ne sais pas quel degré réel. », entretien avec un acteur de la filière énergie.

La conséquence pratique est que les organisations pilotent surtout des coûts apparents, des quotas d'usage ou des accès fonctionnels. Elles ne peuvent pas piloter l'empreinte réelle de l'inférence, car toutes les données nécessaires ne leur sont pas transmises par les fournisseurs.

La transparence ne deviendra une pratique courante que si elle entre dans les contrats, les appels d'offres, les exigences de *reporting* et les critères d'achat. Sans exigence de preuve, l'opacité reste l'architecture par défaut, ce qui peut démotiver même les plus engagés.

6.4 Location-based, market-based : la comptabilité qui ne dit pas toute la réalité physique

La distinction entre méthodes location-based et market-based est indispensable pour comprendre les bilans cloud et IA. Elle est souvent mal comprise et parfois instrumentalisée.

Location-based vs. market-based : deux lectures d'une même réalité

Location-based : l'empreinte est calculée sur la base du mix électrique moyen du réseau dans la zone géographique de consommation. En France, RTE indique une intensité carbone de la production électrique de 21,7 gCO₂eq/kWh en 2024⁸⁹. Pour l'Union européenne, l'EEA⁹⁰ documente une baisse de l'intensité GES de la production électrique en 2024, estimée à -9% par

⁸⁹ <https://analysesetdonnees.rte-france.com/bilan-electrique-2024/emissions#Introduction>

⁹⁰ <https://www.eea.europa.eu/en/analysis/indicators/greenhouse-gas-emission-intensity-of-1>

rapport à 2023.

Market-based : l'empreinte est calculée sur la base de l'électricité que l'entreprise a contractuellement choisie, via des garanties d'origine (GO), des PPA (*Power Purchase Agreements*) ou d'autres instruments. Elle reflète une stratégie d'achat, pas nécessairement la physique du réseau à l'heure de la consommation.

Le GHG Protocol⁹¹ distingue les approches *location-based* et *market-based* pour la comptabilité des émissions liées à l'électricité : la première renseigne le mix physique moyen d'un réseau, la seconde les instruments contractuels d'achat d'électricité. Les deux approches sont utiles, mais elles ne répondent pas à la même question.

Le risque apparaît lorsque le *market-based* est lu comme une description physique complète. Un contrat d'électricité bas carbone peut être pertinent pour signaler une intention stratégique, mais ne dit pas quel mix alimente physiquement un data center à une heure donnée, ni quels effets locaux la demande produit sur le réseau, ni si l'électricité disponible aurait pu servir d'autres usages décarbonants.

Pour la France, ce point est structurant. Le mix électrique français étant bien en deçà de la moyenne européenne, l'avantage français est réel en méthode *location-based*. Mais il ne dispense pas de s'interroger sur les volumes, les effets de concentration locale, les pointes de demande et les effets rebond.

« Dans nos rapports, on alloue la consommation électrique de manière localisée. On ne peut pas financer des énergies renouvelables en Inde pour compenser notre consommation française. », entretien avec un acteur de l'énergie.

6.5 Le *carbon-aware computing* : intérêt réel, limites françaises

Le *carbon-aware computing*⁹² peut être utile s'il permet de déplacer certaines charges de calcul vers des moments ou des lieux où l'électricité est moins carbonée, moins coûteuse ou moins contrainte. Dans le cas français, cette approche doit être nuancée : l'électricité moyenne est déjà largement bas carbone, ce qui réduit l'intérêt d'un arbitrage purement géographique à l'intérieur du pays. En revanche, la temporalité, les contraintes locales de réseau, les pointes, la disponibilité solaire, la chaleur récupérable et le signal prix peuvent devenir plus importants.

La cloche solaire : déplacer le calcul au bon moment

En France, certaines fenêtres de milieu de journée peuvent être favorables aux charges déplaçables, notamment lorsque production solaire, signal prix, disponibilité réseau et intensité carbone convergent. C'est le cas de la cloche solaire, généralement située autour de 10h-14h, qui peut correspondre à des prix d'électricité plus bas et à un mix plus favorable. Ce pilotage doit toutefois s'appuyer sur des données prévisionnelles réelles, car prix, carbone et contrainte réseau ne coïncident pas toujours.

⁹¹ <https://ghgprotocol.org/corporate-standard>

⁹² Le *carbon-aware computing* consiste à orienter ou moduler des calculs numériques en fonction de l'intensité carbone de l'électricité disponible. Dans sa version géographique, il s'agit de déplacer une charge de calcul vers une région où le mix électrique est moins carboné ; dans sa version temporelle, il s'agit d'exécuter certains traitements au moment où l'électricité disponible est moins carbonée ou moins contrainte.

Dans le cas français, cette approche doit être nuancée. L'électricité moyenne est déjà largement bas carbone, ce qui peut réduire l'intérêt d'un arbitrage purement géographique à l'intérieur du pays. En revanche, la temporalité, les contraintes locales de réseau, les pointes, la disponibilité des équipements, la chaleur récupérable et le signal prix peuvent devenir plus importants.

Selon un entretien avec une organisation énergétique, « *Entre 10h et 14h, l'électricité peut être 30 à 40% moins chère qu'aux heures de pointe, selon les conditions de marché. L'inertie thermique des bâtiments de data centers, dont la température cible se situe entre 19 °C et 27 °C, permet souvent de suspendre la climatisation pendant cette fenêtre sans intervenir sur les serveurs.* » Ce verbatim doit être lu comme un signal opérationnel, non comme une règle générale. Le potentiel réel dépend du type de refroidissement, du niveau de charge IT, des conditions météorologiques, des exigences de continuité de service, des marges d'exploitation du site et de la capacité à piloter le froid sans dégrader les conditions d'exploitation.

Les charges déplaçables incluent l'entraînement différé, l'inférence non-temps réel, l'automatisation programmable, certaines analyses documentaires et certaines tâches agentiques. Ce signal illustre l'intérêt potentiel d'une synchronisation entre charge IT et production solaire, encore peu exploitée.

RTE publie des prévisions d'intensité carbone du mix électrique à 72 heures. Ces prévisions permettent à des systèmes d'orchestration de planifier les charges déplaçables dans les fenêtres les plus favorables, à la fois pour l'empreinte et pour le coût.

Limite : toutes les charges IA ne sont pas déplaçables. L'inférence temps réel -*chatbot*, assistant, recommandation- n'a pas de marge de déplacement. Les agents IA autonomes pourraient créer une charge permanente difficile à moduler. En revanche, les charges batch - entraînements, analyses documentaires, traitements différés - offrent une vraie flexibilité temporelle.

Certains acteurs interviewés expérimentent cette logique : orientation des workflows vers les meilleurs mix énergétiques, API de *load balancing* entre *data centers* selon le mix instantané, priorité aux charges déplaçables pendant la cloche solaire. Ces pratiques restent marginales, mais illustrent une direction.

6.6 Le Scope 3 de l'IA : la chaîne matérielle oubliée

Le bilan environnemental de l'IA comprend une chaîne matérielle amont et aval dont les entretiens soulignent l'importance et l'invisibilité. Mais avant même de parler de ce que cette chaîne pèse, il faut aborder une question : deux acteurs, travaillant sur le même objet, peuvent afficher des chiffres qui semblent se contredire sans qu'aucun des deux ne se trompe.

Un cas concret : 39% ou 10%, qui a raison ?

Lors d'un échange, un architecte spécialisé indiquait que la part du bâtiment dans l'empreinte d'un data center était, selon lui, proche de 10%. Un autre interlocuteur répondait que cette part était plutôt de l'ordre de 40%, en renvoyant à l'ACV Data4 / APL⁹³ du DC19 de Marcoussis.

Les deux chiffres ne sont pas nécessairement contradictoires. Ils ne portent pas sur le même périmètre.

L'ACV Data4 / APL ne porte pas sur un data center complet. Elle porte sur l'infrastructure d'accueil du data center sur vingt ans, hors IT, c'est-à-dire hors serveurs, GPU, stockage, réseau IT, renouvellement de ces équipements et consommation d'énergie associée. Sur ce périmètre hors IT, la production des matériaux et équipements représente 39% de l'impact climat. Si l'on ajoute le transport sur site, le chantier et l'installation, la construction au sens large représente

⁹³ <https://www.data4group.com/ressources/data-center-mesurer-limpact-pour-mieux-agir/>

environ 48% de l'impact étudié. Les 52% restants correspondent principalement à l'exploitation hors IT, 48%, et à la maintenance hors IT, 4%.

Mais si l'on ajoutait l'IT au périmètre, le total augmenterait fortement. La part relative du bâtiment et de l'infrastructure d'accueil diminuerait donc mécaniquement. Par exemple, si l'impact IT ajouté pesait autant que tout le périmètre Data4 / APL, les 48% deviendraient environ 24% du total élargi. S'il pesait deux fois ce périmètre, ce serait environ 16%.

L'ordre de grandeur de 10% donné par l'architecte repose vraisemblablement sur un périmètre plus large, intégrant aussi l'IT, ses consommations associées, ses renouvellements rapides et une durée de vie du bâtiment supérieure à vingt ans.

Le point clé est donc simple : **avant de comparer des chiffres, il faut toujours regarder ce qui est inclus dans le périmètre.** Le 39% de Data4 / APL ne signifie pas que le bâtiment représente 39% de l'empreinte d'un data center complet. Il signifie que, dans une ACV centrée sur l'infrastructure d'accueil hors IT, la production des matériaux et équipements représente 39% de l'impact climat.

Le même phénomène, à l'échelle mondiale

Ce n'est pas un cas isolé propre à un data center français. Le même mécanisme de périmètre explique l'écart, en apparence incohérent, entre les projections mondiales d'émissions des data centers pour le milieu des années 2030 : en 2025, l'AIE avançait 300 MtCO₂ dans son scénario de politiques actuelles et 500 MtCO₂ dans un scénario de déploiement massif de l'IA, mais ces chiffres ne couvraient que l'électricité consommée, assimilable au scope 2. Le Shift Project⁹⁴, sur un périmètre élargi intégrant une part de la fabrication du matériel, avançait un scénario haut à 920 MtCO₂e à horizon 2030.

La controverse n'est donc pas : qui a le bon chiffre ? La question clé est aussi : **quel périmètre guide le mieux la décision qu'on cherche à prendre ?**

Le tableau suivant propose une lecture simple de ces périmètres, non pour figer une méthode unique, mais pour rendre visibles les couches d'impact souvent mélangées dans les chiffres disponibles.

⁹⁴ <https://theshiftproject.org/publications/intelligence-artificielle-centres-de-donnees-rapport-final/>

Décomposer les scopes pour rendre l’empreinte IA visible

Scope	Ce qu’il couvre pour l’IA	Niveau de visibilité actuel
Scope 1 (émissions directes)	Fluides frigorigènes, fuites, groupes électrogènes, tests et secours.	Faible : données rarement disponibles au niveau du service IA ou de la requête.
Scope 2 (électricité)	Consommation électrique des serveurs et équipements mobilisés pendant l’entraînement ou l’inférence.	Variable : dépend de la méthode utilisée, location-based ou market-based et du niveau de transparence du fournisseur.
Scope 3 amont - fabrication et chaîne matérielle	Fabrication et acheminement des GPU, serveurs, câbles, équipements électriques, refroidissement, batteries, matériaux du bâtiment, extraction des minerais critiques, eau ultrapure et procédés de production des puces.	Très faible : souvent absent des métriques opérationnelles utilisées pour piloter les usages IA. ⁹⁵
Scope 3 aval - durée de vie, fin de vie, réemploi	Renouvellement des serveurs, prolongation de la durée de vie, réemploi, recyclage, fin de vie des GPU, batteries, équipements électriques, reconversion ou démantèlement des sites.	Quasi nul : angle mort structurel, alors que la durée de vie des équipements conditionne fortement l’empreinte réelle.
Scope 3 élargi - effets induits des usages	Effets rebond, usages induits, automatisations supplémentaires, optimisation d’activités carbonées, augmentation des volumes d’activité. Cette catégorie ne correspond pas toujours à une comptabilité carbone directement attribuable, mais elle est indispensable pour évaluer l’impact réel d’un usage IA.	Non mesuré ou très difficile à attribuer : ce scope oblige à regarder la finalité de l’usage, pas seulement l’infrastructure qui l’héberge.

Les ordres de grandeur disponibles confirment l’importance du scope 3 amont au-delà de ce seul cas, mais ils ne doivent pas être lus comme des chiffres directement comparables. L’hypothèse Schneider Electric reprise par The Shift Project aboutit à environ 25% de l’empreinte carbone totale. L’estimation Boavizta peut atteindre jusqu’à 45% dans un contexte électrique bas carbone. L’écart s’explique vraisemblablement par les périmètres retenus, le mix électrique considéré et les hypothèses de durée de vie ou de renouvellement des équipements. Plus l’électricité d’usage est bas carbone, plus la fabrication des équipements et de l’infrastructure devient visible dans l’empreinte

⁹⁵ Sources : 1) Shift Project <https://theshiftproject.org/app/uploads/2025/11/RF-IA-UK.pdf> Le Shift indique que les estimations actuelles placent l’empreinte carbone des data centers autour de **75% pour l’usage et 25% pour l’empreinte incorporée**, sur la base de Schneider Electric 2023. Il précise aussi que Boavizta donne une répartition différente selon l’intensité carbone de l’électricité : en ordre de grandeur, **20% fabrication / 80% usage** avec un mix européen 2017 très carboné, et jusqu’à **45% fabrication / 55% usage** dans un contexte français bas carbone. Le Shift insiste surtout sur le fait que la part fabrication augmente mécaniquement lorsque l’électricité d’usage se décarbone.;

2) Schneider Electric 2023 confirme que le scope 3 peut représenter une part majeure de l’empreinte totale d’un data center : selon les hypothèses, **scope 1 : 0,2-0,5%, scope 2 : 31-61%, scope 3 : 38-69%**. Le même papier indique que prolonger la durée de vie des serveurs d’un an peut réduire le carbone incorporé cumulé d’environ **16%**, tout en rappelant qu’il faut arbitrer avec les gains d’efficacité énergétique de nouveaux équipements. <https://blog.se.com/datacenter/2023/07/11/demystifying-data-center-scope-3-carbon-with-our-findings/> ;

3) Data4 / APL <https://www.data4group.com/ressources/data-center-mesurer-limpact-pour-mieux-agir/> ajoute un point très utile : même en excluant les serveurs informatiques du périmètre de l’ACV, la **production des matériaux et équipements** représente **39%** de l’impact climat du data center DC19 sur 20 ans, derrière l’exploitation à 48%.

totale. Ces estimations convergent surtout sur un point : la chaîne matérielle peut représenter une part significative de l’empreinte, en particulier lorsque l’électricité d’usage est bas carbone.

De façon dispersée, mais présente, ce même enjeu de périmètre revient dans les entretiens. Un interlocuteur, dont la logique rejoint vraisemblablement celle de l'architecte cité en ouverture de cette section, insiste sur la dissociation entre bâtiment, équipements informatiques et obsolescence : « *Ils font des ACV de construction : 50 ans pour le bâtiment, y compris la construction. Concernant les data centers, la part de la construction par rapport à tout le matériel - serveurs... - est petite, à peine 10%. Problème pour les data centers : l'infrastructure est soumise à une obsolescence supérieure à n'importe quel ouvrage.* » Le même écart de périmètre - infrastructure d'accueil hors IT d'un côté, data center élargi intégrant l'informatique et ses renouvellements de l'autre - explique donc à la fois le 10% évoqué plus haut et le 39% de l'ACV Data4.

La durée de vie matérielle apparaît aussi comme un levier frugal concret, notamment pour les GPU : « *On sait assurer la durée de vie du GPU. On a des RTX A4500 qu'on a retrofités - durée de vie de 8 ans.* », illustration par un expert. Plusieurs entretiens relient directement ces sujets à l'exigence de preuve : « *On rate le train de la transparence. Notre objectif : aller chercher la performance et la transparence.* »

La fabrication d'un GPU de dernière génération mobilise des chaînes d'approvisionnement longues : minerais critiques, eau ultra-pure, gaz industriels, lithographie EUV. Elle consomme de l'énergie et de l'eau dans les usines de Taïwan, de Corée ou d'autres pays producteurs, et génère des déchets chimiques spécifiques. Cette empreinte de fabrication est rarement attribuée de manière transparente aux services IA, aux clients ou aux requêtes. Elle peut apparaître dans des bilans corporate agrégés, mais reste largement absente des métriques opérationnelles utilisées pour arbitrer les usages IA.

Ce que cette discussion de périmètre implique pour l'action

Faire durer les GPU, éviter les renouvellements prématurés dictés par la course à la performance, n'apparaît dans aucun bilan scope 2. C'est un argument supplémentaire pour ne pas réduire la frugalité à la seule optimisation de l'inférence. Plus largement, tant que ce scope 3 matériel reste un angle mort méthodologique, il risque de devenir un droit à compenser plutôt qu'un objet de pilotage.

Voici des axes d'action qui s'en dégagent : éviter la demande inutile avant d'optimiser ce qui reste ; réduire la puissance et la durée d'usage ; prolonger la durée de vie des équipements ; substituer par du matériel ou une architecture moins intensifs ; ne traiter en dernier recours, par compensation, que le résiduel réellement incompressible. Qualifier le périmètre est donc la condition pour que cette hiérarchie s'applique au bon dénominateur.

6.7 Ce que la traçabilité d'exécution devrait couvrir

Pour rendre l'inférence gouvernable, il faudrait rendre visible ce qui est aujourd'hui masqué.

Les entretiens font ressortir une liste de ce qu'il serait souhaitable d'avoir pour permettre un pilotage environnemental sérieux, sachant que certaines données sont difficiles à obtenir.

Les données nécessaires à une traçabilité souhaitable

Localisation d'exécution : quel data center ? Dans quelle région ? Sur quel mix électrique réel à l'heure de la requête ?

Modèle appelé : quel modèle ? Quelle version ? Quelle taille (paramètres) ? A-t-il changé depuis la dernière requête ?

Matériel mobilisé : quel type de GPU ou CPU ? Quel âge des équipements ? Quel taux d'utilisation au moment de la requête ?

Intensité énergétique : consommation effective (pas seulement estimée par des benchmarks en conditions idéales) ?

Eau : volume d'eau consommé pour le refroidissement de cette requête, ou de cette session ?

Redondance : niveau de redondance activé ? (Impact sur la consommation réelle vs capacité installée)

Allocation multi-clients : quelle part de l'infrastructure physique est allouée à ce client sur ce serveur mutualisé ?

Scope 3 amont : empreinte de fabrication amortie sur la durée de vie des équipements ?

Cette traçabilité ne signifie pas qu'il faudra exiger que chaque donnée soit mesurée directement à la requête. Certaines devront être estimées, allouées ou amorties. Mais sans exigence minimale de traçabilité, aucun pilotage sérieux n'est possible. En effet, ces données ne sont aujourd'hui disponibles que pour une fraction d'acteurs et rarement communiquées aux clients. Des opérateurs cloud publient des tableaux de bord carbone, mais avec des périmètres et des méthodes qui rendent les comparaisons difficiles.

Des outils techniques existent pour collecter une partie de ces données. CodeCarbon, EcoLogits, Alumet et d'autres outils open source permettent d'instrumenter du code pour mesurer la consommation réelle, à condition d'avoir accès à l'infrastructure. Cependant, c'est le fournisseur qui décide ce qu'il transmet. Et aujourd'hui, il transmet peu.

6.8 AI for Green vs. Green AI : une distinction nécessaire

Le chapitre 5 a introduit la distinction AI for Green / Green AI pour éviter de confondre finalité environnementale et frugalité réelle. Ce chapitre la reprend sous l'angle de la preuve : ces deux catégories ne demandent ni les mêmes indicateurs, ni les mêmes scénarios de comparaison.

Concept	Définition	Ce qu'il suppose de mesurer
AI for Green	L'IA comme outil au service de la transition écologique. Par exemple, prévision ENR, maintenance prédictive réseau, optimisation thermique, détection de fuites...	Gain environnemental net de l'usage IA (scénario avec versus sans IA), robustesse du modèle, effets rebond, durabilité dans le temps
Green AI	L'IA éco-conçue et éco-exploitée de manière frugale : modèle proportionné, architecture efficiente, infrastructure bas-carbone, matériel durable	Empreinte de l'infrastructure IA elle-même : scope 1, 2, 3 ; durée de vie matérielle ; empreinte par service rendu

La distinction est importante parce qu'elle évite deux confusions symétriques déjà mentionnées : croire qu'une IA utilisée pour l'environnement (AI for Green) est automatiquement frugale (Green AI). Elle peut être surdimensionnée, mal intégrée, peu robuste ou générer des effets rebond ; croire

qu'une IA écoconçue (Green AI) contribue automatiquement à la transition écologique. Elle peut servir un usage à faible valeur environnementale, tout en étant techniquement sobre.

Les entretiens soulignent que les usages IA les plus pertinents pour la transition énergétique, tels que prévision de consommation, maintenance prédictive, optimisation des réseaux, relèvent souvent du *machine learning* classique, pas des grands modèles génératifs. **L'AI for Green est souvent une IA non spectaculaire** : un modèle de prévision robuste, un algorithme d'optimisation, un système de détection d'anomalies. **Sa frugalité est souvent structurelle** : elle est conçue pour un usage précis, avec des données délimitées, dans un système supervisé.

« L'IA générative est moins présente dans l'IA for Green, qui relève plutôt du machine learning classique. Ce sont des modèles de prévision, d'optimisation, de classification, de computer vision, de détection d'anomalies ou de séries temporelles, souvent plus pertinents et potentiellement moins coûteux que les grands modèles génératifs. », entretien avec une organisation énergétique.

Un exemple de port européen transmis en entretien illustre cette exigence de preuve. L'interviewé le résume ainsi : « On a aussi l'optimisation du chargement-déchargement des bateaux au niveau du port. Analyse prédictive des horaires d'arrivée des bateaux de marchandises. Il y a des installations spécialisées. Les capitaineries ne savent pas prédire les heures d'arrivée, avec la marée, la houle. C'est un port fluvial, donc beaucoup de bateaux arrivent plus tôt ou plus tard. Il faut alors s'ancrer, attendre son tour, jusqu'à 12 h ou 24 h. Certaines compagnies vont alors vers d'autres ports pour décharger. Le système permet de prédire l'heure d'arrivée, ce qui augmente le flux de bateaux et améliore la gestion des équipements. Le taux d'amende pour retard a diminué de 90%. Le système se base sur un historique, le trafic, les marées. »

Ce cas montre qu'un usage IA peut produire un gain opérationnel réel, mais que ce gain doit être qualifié dans son système complet : réduction des attentes, meilleure utilisation des équipements, baisse des pénalités, mais aussi effets induits éventuels sur les flux.

Mise en garde : prôner une IA sobre devient d'autant plus important que l'IA générative, par ses effets d'échelle et sa facilité d'usage, pourrait se substituer à des approches plus économes. Le choix ne se limite pas à grands modèles versus petits modèles : il porte aussi sur l'accessibilité, l'outillage, les plateformes de regroupement, les aides au choix, le routage et les conditions d'intégration métier.

6.9 AI for Brown : ce que la mesure doit aussi couvrir

La mesure de l'empreinte IA ne peut pas s'arrêter aux usages vertueux. L'IA optimise ce qu'on lui demande d'optimiser, y compris des activités dont l'intensité carbone est élevée.

Des dirigeants du secteur pétrolier ont explicitement présenté l'IA comme un moyen de produire du pétrole et du gaz plus vite et à moindre coût, en guidant les forages, en prédisant les problèmes et en ouvrant des zones jugées trop difficiles ou trop coûteuses. Cette application de l'IA, loin d'être un cas marginal, illustre que le bilan environnemental global de l'IA dépend autant des usages servis que de l'efficacité de l'infrastructure.

Une analyse de Wood Mackenzie⁹⁶ estime que l'amélioration des taux de récupération des champs pétroliers existants, appuyée notamment sur des outils d'analyse par IA, pourrait permettre de récupérer 470 milliards à plus de 1 000 milliards de barils supplémentaires selon les hypothèses

⁹⁶ <https://www.woodmac.com/ja/horizons/ai-powered-analysis-oil-field-potential/>

retenues. Il s'agit d'un potentiel technique et économique, non d'une prévision de production effectivement réalisée. Une étude publiée en 2026 dans npj Climate Action⁹⁷ montre par ailleurs que l'amélioration de la productivité des filières énergétiques par l'IA peut modifier les prix, la production, la consommation et les substitutions entre sources d'énergie. Dans les scénarios où seuls les gains de productivité des filières fossiles sont appliqués, le modèle estime de 0,6 à 2,4 Gt CO₂ par an d'émissions de combustion supplémentaires. Lorsque des gains de productivité sont appliqués simultanément aux filières fossiles et bas-carbone, il y aura néanmoins une hausse nette des émissions, de 0,47 à 1,8 Gt CO₂ par an selon les scénarios. Ces résultats ne constituent pas une prévision énergétique à 2035 : ils comparent une situation sans gains de productivité liés à l'IA à une situation où ces gains sont introduits, sans représenter explicitement une trajectoire complète de transition énergétique ou de sortie des fossiles.

Un exemple récent illustre directement cette logique : en septembre 2026, TotalEnergies⁹⁸ et Mistral AI ont annoncé un partenariat de trois ans, représentant plus de 100 millions d'euros, pour développer des modèles d'IA dédiés notamment aux géosciences et à l'ingénierie des réservoirs de pétrole et de gaz. Le programme vise notamment à générer des scénarios d'exploration et à optimiser ou prolonger la durée de vie de projets existants. Ce cas montre que l'IA peut devenir un levier d'innovation et de productivité dans les activités fossiles elles-mêmes.

Ce point est important pour la gouvernance : un data center bas-carbone peut héberger des usages qui accélèrent des activités carbonées. L'avantage carbone de l'infrastructure ne suffit donc pas à établir le bilan environnemental de l'usage : celui-ci dépend aussi des activités que l'IA remplace, optimise, prolonge ou rend possibles, ainsi que de leurs effets induits. **La sobriété de l'infrastructure et la pertinence de l'usage sont encore trop souvent évaluées séparément.**

Les choix les plus frugaux sont ceux qui combinent une architecture sobre (logiciel, données, infrastructure, matériel) et un usage à forte valeur environnementale ou sociale démontrée. À l'inverse, une IA très optimisée mais appliquée à une activité fortement carbonée peut rester problématique.

De l'impact environnemental à l'arbitrage énergétique

Ce résultat ouvre une question qui dépasse l'empreinte de l'IA : si l'IA améliore simultanément la productivité des hydrocarbures, des renouvelables, des réseaux et demain d'autres filières énergétiques, **quelles trajectoires renforcera-t-elle réellement ?** Les prix et la disponibilité des différentes énergies compteront, mais ils ne seront pas les seuls déterminants. Les choix et dépendances géopolitiques et géostratégiques, les contraintes climatiques, la sécurité d'approvisionnement, les politiques industrielles, les infrastructures existantes et les choix de souveraineté pèseront également dans l'arbitrage.

L'AI for Brown ne se limite pas aux usages directement liés à l'extraction ou à la consommation d'énergies fossiles. Elle peut aussi améliorer l'efficacité, la compétitivité ou la capacité d'expansion de systèmes dont la trajectoire reste fortement carbonée, par exemple certaines chaînes logistiques, industrielles ou de transport. L'enjeu est donc moins d'étiqueter un usage en fonction de son secteur que d'identifier la trajectoire économique et matérielle qu'il contribue à prolonger ou à accélérer.

⁹⁷ <https://www.nature.com/articles/s44168-026-00411-0>

⁹⁸ <https://totalenergies.com/newsroom/totalenergies-annonce-un-partenariat-avec-mistral-en-vue-de-developper-des-modeles-de-frontiere-dintelligence-artificielle-dedies-a-lexploration-et-a-lingenierie-des-reservoirs/?lang=fra>

L'IA peut ainsi devenir un **accélérateur de stratégies énergétiques déjà divergentes**. D'un côté, certains pays, tels les États-Unis, peuvent renforcer une logique de mobilisation accrue des ressources fossiles. De l'autre, la Chine poursuit une stratégie d'électrification, de développement massif des capacités renouvelables et de maîtrise industrielle visant aussi à réduire certaines dépendances énergétiques extérieures. Enfin, l'**Europe**, qui dispose de moins de leviers sur l'offre mondiale, peut néanmoins jouer un **rôle de contributeur-arbitre** par ses choix de consommation, d'équipements, d'infrastructures, de normes et de technologies, susceptibles d'orienter la demande et les chaînes industrielles.

Dénoncer les usages de l'IA qui prolongent les trajectoires fossiles est nécessaire, mais ne suffit pas à constituer une stratégie. Les producteurs disposant d'un intérêt économique à utiliser l'IA pour accroître leurs ressources le feront vraisemblablement dès lors que cette technologie améliore leur compétitivité. La question pour l'Europe est donc aussi de savoir où se situent ses choix stratégiques et ses marges d'action : quelles filières accélérer, quelles infrastructures financer, quels équipements acheter, quels usages soutenir et quelles dépendances réduire. **L'enjeu devient moins de juger l'IA en elle-même que de choisir les systèmes énergétiques dont nous voulons qu'elle augmente la capacité.**

6.10 L'inférence gouvernable : voir, décider, contractualiser

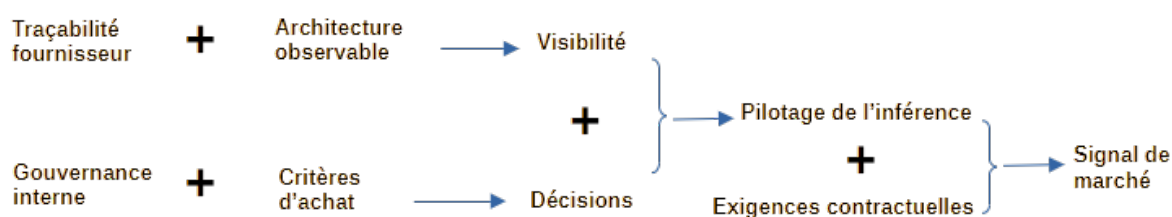
Rendre l'inférence gouvernable consiste à relier trois capacités : voir ce qui se passe, décider ce qui est acceptable, traduire ces décisions dans les architectures, les achats et les contrats. Sans visibilité, la frugalité reste aveugle ; sans gouvernance, la visibilité ne produit pas d'arbitrage ; sans critères d'achat, l'arbitrage ne produit pas de signal de marché.

Les entretiens permettent d'identifier quatre conditions nécessaires, mais non suffisantes, pour que cette gouvernabilité devienne réelle dans les organisations.

Condition	Ce qu'elle implique concrètement	Illustration d'appui
Traçabilité fournisseur	Exiger des fournisseurs cloud et des éditeurs logiciels des données d'usage : localisation, modèle, énergie, eau. Sans obligation contractuelle, cette transparence restera volontaire et partielle.	Les fournisseurs cloud et SaaS transmettent encore des données hétérogènes, avec des périmètres comptables variables - scope 3, location-based, market-based, méthodes européennes ou américaines - ce qui rend difficile la comparaison et le pilotage de l'empreinte réelle de l'inférence.
Architecture observable	Concevoir les systèmes IA pour qu'ils soient « instrumentables » : logs d'usage, métriques de consommation, attribution par cas d'usage.	À chaque couche, l'utilisateur perd de la visibilité. <i>« Le développeur ne voit pas toujours le data center. L'organisation cliente ne voit pas toujours l'infrastructure cloud. Le décideur ne voit pas toujours la chaîne matérielle et contractuelle. »</i> , illustre un interviewé venant d'une ESN.

Condition	Ce qu'elle implique concrètement	Illustration d'appui
Gouvernance interne	Définir qui décide des usages IA, qui fixe les budgets tokens, qui arbitre entre modèles, qui mesure les gains nets. Sans gouvernance, la frugalité reste déclarative.	« C'est d'abord un sujet de gouvernance et de change, après un sujet d'opérationnalisation. Je ne sais pas ce que fait l'ingénieur network, l'ingénieur FinOps, le CloudOps. Ce travail d'intégration doit être fait. Pas de dissonance cognitive. », poursuit le même interviewé.
Critères d'achat	Introduire des exigences de transparence et de frugalité dans les appels d'offres et les contrats. C'est le seul mécanisme qui crée un signal de marché et donc une incitation pour les fournisseurs à progresser.	« La transparence ne deviendra pas usage courant par volontariat fournisseur », précise un responsable du numérique responsable.

Ces conditions ont besoin les unes des autres. La traçabilité fournisseur et l'architecture observable rendent visibles les usages, les appels de modèles, les consommations et les impacts. La gouvernance interne et les critères d'achat transforment cette visibilité en décisions. Sans visibilité, aucun arbitrage n'est possible ; sans arbitrage, la frugalité reste une intention ; sans traduction contractuelle, elle ne modifie ni les architectures, ni les achats, ni les choix d'infrastructure.



Cette chaîne résume le passage d'une frugalité déclarée à une frugalité pilotable.

Voir ne suffit pourtant pas. Une entreprise peut savoir qu'elle consomme sans pouvoir agir sur cette consommation, tant que cette visibilité ne s'est pas traduite dans une gouvernance interne et dans des critères d'achat opposables aux fournisseurs. C'est le **passage d'une frugalité qui se constate à une frugalité qui se pilote**, et c'est précisément ce que peu d'organisations savent faire aujourd'hui à l'échelle, en dehors de pionniers isolés.

Le chapitre 7 prend le relais sur ce point exact : comment ces conditions de gouvernabilité, encore rares, pourraient devenir la norme plutôt que l'exception, par les architectures, les incitations, les achats et les contrats.

Chapitre 7 - Industrialiser la frugalité : ce qui ne se fera pas par bonne volonté

Le chapitre en une phrase

La frugalité ne se diffusera pas par la bonne volonté des utilisateurs ou des développeurs : elle doit entrer dans les architectures, les achats, les contrats et les budgets. Le chapitre cartographie les leviers logiciels, architecturaux et infrastructurels disponibles, les quatre formes d'effet rebond à surveiller et propose la frugalité comme avantage concurrentiel européen plutôt que comme simple contrainte.

La frugalité ne se diffusera pas parce que les utilisateurs feront moins de prompts ou parce que les développeurs seront plus vertueux. Elle se diffusera si elle entre dans les architectures, les achats, les contrats, les politiques de modèles, les budgets d'usage et les critères de performance.

Ce chapitre descend dans la chaîne : où peut-on agir, qui peut agir, et quelles limites empêchent la frugalité de devenir spontanément la norme ?

7.1 Les trois niveaux de la frugalité : usage, architecture, infrastructure

La frugalité IA ne se joue pas à un seul niveau. Les entretiens font ressortir trois niveaux d'arbitrage qui n'ont ni les mêmes acteurs, ni les mêmes outils, ni les mêmes temporalités.

Les trois niveaux : usage, architecture, infrastructure

1. **Frugalité d'usage** : faut-il vraiment utiliser l'IA pour ce problème ? Une règle métier, un moteur de recherche, une automatisation classique ou une meilleure structuration des données pourraient-ils produire le même résultat avec moins de complexité ? Ces options n'ont pas les mêmes coûts, dépendances, besoins de données, ni la même empreinte.
2. **Frugalité d'architecture** : si l'IA est utile, quelle forme doit-elle prendre ? Un logiciel standard avec IA intégrée, une IA sur mesure, un processus métier transformé ou une automatisation sans IA ? Un LLM généraliste est-il nécessaire ou un modèle plus petit suffit-il ? Faut-il du RAG, un routeur de modèles, du cache, une limitation de contexte ? C'est ici que se jouent les choix entre SLM, quantization, distillation, optimisation, fine tuning et limitation des appels inutiles.
3. **Frugalité infrastructurelle** : où et quand le calcul s'exécute-t-il ? Sur quel matériel ? Dans quel data center ? Avec quelle durée de vie des équipements ? Quel refroidissement ? Quel mix électrique ? Quelle temporalité ? Ce niveau relie directement les choix logiciels aux data centers.

Cette tripartition permet d'éviter deux impasses : 1) moraliser les utilisateurs : « faites moins ou pas de prompts » ; 2) techniciser la frugalité : « utilisez des modèles plus petits et moins gourmands ». **Ces gestes peuvent être utiles, mais sont insuffisants. Une IA réellement frugale doit être proportionnée à l'usage, structurée dans son architecture et située dans une infrastructure maîtrisée.**

7.2 Les leviers logiciels et architecturaux

Au niveau de l'architecture logicielle, plusieurs leviers sont documentés dans les entretiens. Ils ne sont pas tous équivalents en effort, en gain ou en applicabilité, mais constituent un répertoire

concret pour les organisations qui souhaitent passer de la frugalité déclarative à la frugalité opérationnelle. Le chapitre 5 a présenté ces leviers comme composants d’une architecture frugale. Ici, ils sont repris comme répertoire d’action, en particulier pour identifier où une organisation peut réellement intervenir.

Levier	Ce qu'il produit concrètement
SLM (petit modèle spécialisé)	Réduit la puissance de calcul nécessaire à l'inférence. Plus rapide, moins cher, plus explicable, déployable localement. Combine frugalité, souveraineté et performance située.
RAG (Retrieval-Augmented Generation)	Contextualise un modèle général avec une base documentaire métier sans réentraîner. Évite le coût d'un entraînement complet tout en améliorant la pertinence.
Routeur de modèles	Orienté automatiquement chaque requête vers le modèle le moins coûteux et suffisamment performant. Transforme la frugalité en décision d'architecture plutôt qu'en geste individuel.
Quantization	Réduit la précision numérique des poids du modèle (32 bits → 8 ou 4 bits), diminuant la mémoire et la puissance de calcul nécessaires à l'inférence au prix d'une légère perte de précision.
Pruning (élagage)	Supprime les paramètres du modèle qui contribuent peu à la performance, réduisant la taille et la complexité de calcul.
Distillation	Entraîne un petit modèle à reproduire le comportement d'un grand modèle (modèle enseignant/élève). Permet d'obtenir une performance proche avec un coût d'inférence bien inférieur.
Cache des réponses	Stocke les réponses aux requêtes fréquentes pour éviter de recalculer à chaque fois. Très efficace pour les usages répétitifs ou documentaires.
Limitation du contexte	Réduit la taille de la fenêtre de contexte envoyée au modèle, ce qui réduit proportionnellement une partie de la consommation. Souvent possible sans perte de qualité pour des tâches ciblées.
Frugalité de la donnée	Sélectionne, nettoie et réduit les datasets d'entraînement. Certains jeux de données sont énormes, mal triés ou bruités. Entraîner sur une donnée surdimensionnée consomme inutilement du calcul sans améliorer la performance.

Ces leviers ne sont pas indépendants. Un routeur de modèles présuppose plusieurs modèles disponibles, dont des SLM. Le RAG présuppose une base documentaire bien structurée. La quantification (*quantization*) présuppose un modèle déjà entraîné. La frugalité architecturale est un système d'arbitrages.

« Les développeurs prennent des modèles, des bibliothèques, des datasets, des morceaux de code, les testent, les versionnent, les industrialisent et les déploient, mais entrent rarement finement dans les algorithmes. L'optimisation profonde apparaît surtout lorsqu'une contrainte forte l'impose, par exemple une contrainte budgétaire ou matérielle. », précise un acteur du numérique.

La frugalité de la donnée, levier oublié

La frugalité ne se joue pas seulement dans le choix du modèle, la taille du contexte ou l’optimisation de l’inférence. Elle commence dans la donnée. Les entretiens la font émerger comme

un frein majeur à l'adoption de l'IA, au-delà du modèle, du calcul ou de l'infrastructure. Silotée, hétérogène, dégradée dans le temps, structurée dans des formats anciens, la donnée industrielle et énergétique pose un problème que le choix d'un meilleur modèle ne résout pas.

« Il y a un gros enjeu IA : tout est siloté, hétérogène, la qualité évolue dans le temps. Comment nettoyer ? Tenir à jour ? Sous-jacent : faire des choses avec l'IA. », entretien avec un acteur de la R&D.

Le travail de data engineering - nettoyage, structuration, documentation, mise à jour, gestion de la qualité dans le temps - est long, coûteux et incontournable. Il est pourtant une condition de frugalité. Entraîner ou interroger un modèle sur des données mal nettoyées, bruitées, redondantes ou surdimensionnées consomme du calcul inutilement, produit des résultats moins robustes et augmente le risque de dérive en production. La frugalité de la donnée consiste donc à sélectionner, nettoyer, documenter, réduire et maintenir les jeux de données utiles, plutôt qu'à accumuler des volumes qui donnent une illusion de richesse informationnelle.

La dégradation permanente de la qualité de la donnée revient de façon récurrente. Il ne suffit pas de construire un bon dataset une fois. La donnée se dégrade dans le temps, par obsolescence des capteurs, par évolution des comportements, par changement des systèmes sources ou par réingestion de données parasites. Un système IA intégré dans un processus opérationnel doit donc intégrer dès sa conception une stratégie de maintien de la qualité de la donnée, non comme un chantier ultérieur, mais comme une composante structurelle du système.

« IA aujourd'hui : ingestion et réingestion des données, y compris parasites. Il y a une dégradation permanente et continue de la qualité de la donnée disponible. Je crois à l'association LLM + compétences IA + bases de connaissances, pas à une IA générale. », entretien avec un acteur numérique.

7.3 Les leviers infrastructurels

La frugalité ne peut pas rester logicielle. L'infrastructure matérielle - GPU, refroidissement, durée de vie des équipements, localisation du calcul - détermine une part importante de l'empreinte réelle, souvent invisible depuis la couche logicielle.

Le chapitre 6 a établi pourquoi le scope 3 matériel reste un angle mort ; voici comment agir concrètement sur ce levier.

Faire durer les GPU : un levier de frugalité encore négligé

Les entretiens font émerger un aspect encore peu présent dans les débats sur la frugalité : la durée de vie des équipements de calcul. Un expert de la construction donne un ordre de grandeur utile : *« Les GPU de dernière génération coûtent environ 40 000 €/kW installé, contre moins de 10 000 €/kW pour une centrale nucléaire »*⁹⁹. Les deux investissements n'ont ni la même fonction, ni la même durée de vie, ni le même périmètre, mais la comparaison met néanmoins en évidence un

⁹⁹ Il s'agit d'ordres de grandeur issus d'un entretien, destinés à illustrer l'intensité capitaliste des équipements de calcul et non à comparer directement un GPU et une centrale nucléaire. NVIDIA indique un TDP maximal de 700 W pour le H100 SXM (<https://www.nvidia.com/en-us/data-center/h100/>). Le prix du B200 Blackwell a été annoncé par Jensen Huang, PDG de Nvidia, entre 30 000 et 40 000 dollars par puce (<https://www.cnbc.com/2024/03/19/nvidias-blackwell-ai-chip-will-cost-more-than-30000-ceo-says.html>). Ces données conduisent à des ordres de grandeur de plusieurs dizaines de milliers de dollars par kW au niveau de l'accélérateur seul, hors serveur, réseau, refroidissement, bâtiment et intégration <https://www.nvidia.com/en-us/data-center/h100/>. À titre de comparaison, la Cour des comptes évalue Flamanville 3 à 23,7 Md€-2023 pour 1 650 MW, soit environ 14 400 €/kW par simple division. Les périmètres, fonctions et durées de vie étant très différents, ces ratios ne constituent pas une comparaison de coût complet. (Source : (<https://www.ccomptes.fr/sites/default/files/2025-01/20250114-La-filiere-EPR%20-une-dynamique-nouvelle-des-risques-persistants.pdf>))

point souvent occulté : au-delà de l'électricité consommée, l'IA mobilise des équipements très coûteux dont le risque d'obsolescence rapide peut devenir un enjeu de frugalité.

Un acteur interviewé indique ainsi privilégier un équipement progressif, ajusté à la demande réelle, plutôt qu'un achat massif en amont susceptible de devenir rapidement obsolète. L'enjeu est double : éviter d'immobiliser du capital dans des capacités inutilisées et ne pas confondre souveraineté avec accumulation de matériel. L'arbitrage porte alors autant sur le **moment où l'on achète** que sur la quantité achetée : disposer de capacités suffisantes sans renouveler prématurément des équipements encore utilisables.

Faire durer le matériel : quatre démarches concrètes

1. **Éviter les renouvellements prématurés** : ne pas remplacer des GPU encore performants sous l'effet de la pression marketing ou de la course à la dernière génération. L'équipement obsolète d'une entreprise peut être suffisant pour une autre.
2. **Réemploi et reconditionnement** : des serveurs d'occasion ou reconditionnés peuvent couvrir certains usages à faible exigence de latence ou de performance maximale. Le marché du réemploi de GPU est embryonnaire mais réel.
3. **Retraitement et modularité** : concevoir des infrastructures physiques évolutives qui permettent de mettre à jour les composants sans remplacer l'ensemble du rack ou du bâtiment.
4. **Mutualisation** : partager des capacités de calcul entre acteurs à des moments différents, selon les usages et les pics de charge. Selon un interviewé, le data center universitaire de Barcelone illustre ce modèle : usage industriel la nuit, académique le jour.

IA pour optimiser les data centers eux-mêmes

Les entretiens signalent aussi un usage plus discret de l'IA : l'optimisation des data centers eux-mêmes. Jumeaux numériques, supervision thermique, répartition dynamique des charges dans les racks, maintenance préventive, analyse des flux aérauliques et pilotage du refroidissement peuvent contribuer à améliorer l'exploitation. Mais ce levier doit être qualifié avec prudence : il peut optimiser un site donné sans répondre aux questions de fond sur la croissance globale de la demande, l'addition d'usages ou la valeur captée. L'IA peut donc améliorer l'efficacité opérationnelle d'un data center, sans rendre automatiquement frugale la trajectoire numérique qu'il soutient.

La temporalité du calcul : la cloche solaire et le déplacement de charges

Certaines charges IA ne sont pas contraintes par une latence immédiate. Entraînements, analyses documentaires, traitements batch, synchronisations, validations - autant de tâches qui peuvent être décalées vers des fenêtres temporelles plus favorables pour le réseau et pour l'empreinte. Voir aussi l'encadré sur la cloche solaire au chapitre 6.

Ce levier est actionnable aujourd'hui pour les charges différables, mais il ne concerne pas l'inférence temps réel.

L'edge computing comme frugalité située

Déployer des capacités de calcul au plus près des usages - dans les usines, les bâtiments, les véhicules, les stations réseau - peut réduire simultanément la latence, la bande passante consommée pour les transferts de données et la dépendance à des data centers distants. L'edge n'est pas systématiquement moins consommateur en absolu - un parc de petites machines peut être moins

efficace qu'un data center optimisé. Il peut aussi créer de l'obsolescence rapide des terminaux. Dans beaucoup de cas l'*edge* peut être plus frugal, notamment quand les modèles et les données sont bien choisis.

Cependant, ce n'est pas toujours une motivation environnementale qui conduit à ce choix, c'est une contrainte de souveraineté. La frugalité située et la souveraineté convergent.

Il faut y ajouter les modèles alternatifs traités dans l'interlude stratégique.

7.4 La contrainte structurelle du « *build on top* » : le développeur comme assembleur

Dans l'IA contemporaine, le développeur est de plus en plus un assembleur : il est placé dans une chaîne où une part majeure de l'empreinte, de la souveraineté et de la dépendance est déjà inscrite dans les briques qu'il mobilise : modèles pré-entraînés, API, cloud, bibliothèques, *frameworks*, jeux de données et couches d'orchestration.

Il peut limiter les appels inutiles, réduire les contextes, mettre en cache, choisir un modèle plus adapté. Mais il ne peut pas, seul, corriger les choix amont des plateformes, des fournisseurs de modèles et des infrastructures. La frugalité doit aussi remonter vers l'architecture, les achats, les clauses fournisseurs, les choix d'API et la gouvernance d'usage.

Ce que le développeur ne voit pas

Le data center : le développeur appelle une API. Il ne voit pas le data center, les GPU, l'eau, le refroidissement, les batteries ou les générateurs qui rendent cet appel possible.

Le modèle réel : la couche d'abstraction de l'API peut masquer quel modèle est réellement appelé, quelle version, à quelle fréquence il est mis à jour et avec quelle empreinte.

L'empreinte matérielle : la fabrication des GPU, l'eau consommée pour les refroidir, les minerais critiques, la durée de vie des équipements - rien de tout cela n'est visible depuis l'interface de développement.

L'effet de la généralité : un modèle généraliste est conçu pour répondre à des millions de cas d'usage différents. Il est surdimensionné par construction pour la majorité des tâches spécifiques.

La frugalité dépend ainsi des choix faits en amont par les plateformes, les bibliothèques, les fournisseurs de modèles et les éditeurs de logiciels. Ce qui peut être demandé au développeur : choisir le bon modèle parmi ceux disponibles, limiter la taille du contexte, éviter les appels inutiles, implanter du cache, surveiller les volumes. Ce qui ne peut pas lui être demandé : réécrire les briques sur lesquelles il s'appuie.

7.5 La tension structurelle : généralité versus spécificité

La tension fondamentale qui sous-tend les limites de la frugalité peut être formulée simplement : les *business models* du logiciel poussent à la généralité ; la frugalité pousse au nécessaire et suffisant (voir aussi 5.10).

Un modèle de fondation généraliste, une suite logicielle avec IA intégrée, une API cloud générique sont conçus pour être réutilisés le plus largement possible. C'est leur logique économique : plus ils sont utilisés, plus la valeur par unité de coût de développement augmente. C'est **une architecture**

d’incitations. La frugalité suppose de réduire, contextualiser, filtrer. Ces deux logiques sont en tension permanente.

Un fournisseur de LLM généraliste n’a pas d’intérêt intrinsèque à réduire l’usage de son produit ; un éditeur qui intègre l’IA par défaut ne cherche pas à minimiser les appels¹⁰⁰. La frugalité ne se diffusera donc pas spontanément : elle doit être rendue achetable et/ou mesurable et/ou exigible.

Le tableau suivant met en regard ces logiques de marché, ce qu’elles produisent concrètement et leurs effets possibles sur la frugalité.

Logique du marché	Ce qu'elle produit	Effet sur la frugalité
Abonnement mensuel forfaitaire	Accès illimité ou quasi illimité à un service IA	Aucune incitation à réduire les usages. La facture ne varie pas
Facturation au token	Consommation visible et avec un prix unitaire	Incitation économique à l'optimisation, sans lien direct avec l'empreinte environnementale
Modèle généraliste par défaut	Couverture maximale de cas d'usage	Surdimensionnement systématique pour la majorité des tâches
Intégration IA dans les logiciels existants	Diffusion invisible	Croissance invisible de la demande sans décision explicite
Course à la performance	Renouvellement accéléré des modèles et équipements	Obsolescence prématurée, pression sur l'investissement matériel

Révéler cette tension entre diffusion maximale et frugalité maîtrisée permet aussi de révéler des risques d’effet rebond : même lorsque l’empreinte unitaire baisse, l’empreinte totale peut augmenter si les usages se multiplient, se banalisent ou s’étendent à de nouveaux processus.

7.6 Les quatre formes d'effets rebond

Nous avons déjà évoqué les effets rebond de façon transversale (voir 5.3 et 6.6) ; il convient d'en distinguer quatre formes qui doivent être particulièrement surveillées.

Quatre formes d'effets rebond Rebond volumique direct : un modèle plus rapide et moins cher génère plus de requêtes du même utilisateur. L'empreinte totale augmente même si l'empreinte unitaire baisse. Comme le résume un interviewé : <i>« La voiture a rendu le transport tellement abordable qu'elle a engendré une surconsommation. »</i> Rebond applicatif : une IA plus efficace rend possibles de nouveaux usages qui n'existaient pas et qui créent une nouvelle demande de calcul non anticipée. L'empreinte totale augmente. Rebond organisationnel : l'adoption généralisée de l'IA dans une organisation multiplie les points d'entrée dans les systèmes. Chaque cas d'usage semble modeste, mais l'accumulation peut être massive. Rebond sectoriel : l'IA optimise des processus carbonés, augmentant leur productivité et donc leur volume d'activité (forage pétrolier, logistique fossile, agriculture intensive). C’est l’AI for Brown abordée en 6.9 : l’efficacité de l’IA amplifie alors l’activité à impact négatif.

¹⁰⁰ Ce modèle d’architecture d’incitations peut connaître des évolutions à cause notamment de pénuries qui pourraient l’y contraindre

L'effet rebond organisationnel (l'accumulation de petits usages IA dans les processus internes) est à surveiller particulièrement, car il peut devenir une limite structurelle de la frugalité alors même qu'il résulte de choix d'organisation et de diffusion des usages. Les leviers logiciels, architecturaux ou infrastructurels peuvent réduire l'empreinte unitaire, mais cette amélioration ne garantit pas une baisse de l'empreinte totale si les usages se multiplient.

D'un point de vue opérationnel, nous pouvons conclure qu'une organisation ne peut pas piloter la frugalité en suivant seulement l'empreinte unitaire. Elle doit aussi suivre les volumes, les usages induits et l'empreinte totale. Une empreinte par requête qui baisse de 30% ne rend pas le système plus sobre si le nombre de requêtes triple. La frugalité réelle se mesure au total gouverné, pas seulement à l'unitaire optimisé.

7.7 La frugalité institutionnalisée : mécanismes organisationnels

De ce qui précède, nous comprenons aisément que la frugalité doit être gouvernée. Les entretiens font ressortir six mécanismes organisationnels qui, assemblés, peuvent constituer une gouvernance frugale opérationnelle.

Mécanisme	Ce qu'il produit
1. Cartographie des usages IA	Visibilité sur ce qui tourne, où, pour qui, à quelle fréquence, avec quel modèle. Condition de tout pilotage.
2. Politique de modèles	Critères de choix entre LLM, SLM, RAG, on-premise, cloud - documentés et applicables par les équipes.
3. Critères de pertinence d'usage	Grille d'arbitrage : faut-il l'IA pour ce problème ? Quel bénéfice attendu ? Quel scénario sans IA ?
4. Budgets ou quotas intelligents	Tokens, GPU, appels API - non seulement comme maîtrise de coûts, mais comme proxy de maîtrise d'empreinte.
5. Exigences de transparence fournisseur	Localisation, modèle, énergie, eau, réversibilité - intégrées dans les achats et les contrats, pas seulement dans les politiques internes (voir 6.10).
6. Capacité à relier usages et infrastructures	Former les équipes à comprendre que leurs choix logiciels ont des contreparties physiques : GPU, data center, eau, électricité.

Un **septième mécanisme** devrait compléter cette gouvernance : **le pilotage dans le temps et la fin de vie**. La **frugalité** ne se joue pas seulement au moment du choix du modèle, du fournisseur ou de l'architecture. Elle **se vérifie dans la durée** : évolution des volumes de requêtes, montée en charge des usages, renouvellement des modèles, obsolescence des GPU, durée de vie des serveurs, réemploi, *retrofit*, réversibilité fournisseur, sortie des contrats et fin d'exploitation des équipements. Sans suivi temporel, une architecture initialement frugale peut devenir coûteuse, opaque ou surdimensionnée. Le pilotage doit donc couvrir à la fois la vie des usages, la vie des modèles, la vie des équipements et la sortie des dépendances contractuelles.

Ces mécanismes constituent la condition pour éviter que l'IA ne devienne un appel permanent et invisible aux data centers. Une organisation qui passe de la frugalité déclarative à la frugalité gouvernée gère ses usages IA comme elle gère ses énergies : avec des compteurs, des budgets, des critères et des responsables.

Ce pilotage doit aussi intégrer les dimensions humaines, comportementales et sociales : formation, supervision, évolution des métiers, intensification éventuelle du travail, responsabilité des décisions et effets sur les prestataires.

7.8 La frugalité comme avantage concurrentiel émergent

Les entretiens apportent un autre éclairage : la frugalité IA peut devenir pour l'Europe ce que la puissance brute est aux *hyperscalers*, à savoir non pas une contrainte défensive, mais un terrain de différenciation.

L'Autorité de la concurrence¹⁰¹ identifie la frugalité de l'IA comme un paramètre concurrentiel émergent : des solutions plus frugales peuvent permettre à des acteurs moins dotés en infrastructures de rivaliser avec des acteurs disposant de ressources plus massives.

Ce point est important pour la stratégie européenne. L'Europe et la France ne peuvent pas fonder leur stratégie uniquement sur une course à la puissance brute de calcul face aux États-Unis et à la Chine. Elles peuvent se différencier sur la frugalité, la souveraineté, l'explicabilité, la robustesse, des attributs qui correspondent précisément aux exigences des systèmes critiques où l'Europe est forte et peut encore se renforcer : énergie, industrie, santé, défense, transport.

Dans un monde marqué par des ressources finies, des risques de pénurie et une concurrence croissante entre usages - numérique, industrie, transition écologique -, la frugalité acquiert une pertinence différenciatrice.

Exemples de ce que la frugalité peut produire comme avantage stratégique

Maîtrise des coûts : moins de dépendance aux grands modèles, c'est moins de coûts tokens, moins de risque de hausse tarifaire unilatérale par le fournisseur.

Accès au marché : des solutions plus frugales peuvent servir des clients dont l'infrastructure est limitée : industries en région, collectivités, PME, pays du Sud. Si les risques de pénurie ou les conflits d'usage se propagent, elles peuvent être aussi une réponse pour des pays ou organisations qui ne peuvent plus raisonner comme si la puissance de calcul, l'énergie, les GPU et l'eau étaient disponibles sans limite.

Conformité réglementaire : l'AI Act impose déjà une logique de documentation, de maîtrise des risques et de proportionnalité des obligations selon les catégories d'usages.

Différenciation souveraine : un SLM souverain, déployable localement, explicable, maîtrisé, peut gagner des marchés que les grands modèles de fondation ne peuvent pas adresser pour des raisons de confidentialité ou de réglementation.

¹⁰¹ Source : Autorité de la concurrence https://www.autoritedelaconcurrence.fr/sites/default/files/2026-01/Etude_IA_Energie_ADLC-LR.pdf Étude sur les enjeux concurrentiels liés à l'impact énergétique et environnemental de l'IA, 17 décembre 2025, section II.B, « Les enjeux liés à l'émergence de la frugalité comme paramètre de concurrence » (pages 23-27, 92, 100-103, 105-110). L'Autorité y indique que la frugalité et l'impact carbone peuvent devenir des paramètres de concurrence, affectant le prix, la qualité et la capacité d'innovation des acteurs.

PARTIE III

L'IA au service des systèmes
énergétiques : des gains réels,
sous conditions de preuve opérationnelle

Après avoir examiné ce que l'IA demande au système énergétique, cette partie examine ce qu'elle peut lui apporter. Le chapitre 6 a introduit la distinction entre AI for Green et Green AI. Cette partie explore un sous-ensemble précis de l'AI for Green : l'IA appliquée directement aux systèmes énergétiques - réseaux, bâtiments, *microgrids*, infrastructures, organisations qui produisent, transportent, distribuent ou consomment l'énergie. Nous l'appellerons ici « *AI for Energy* ». Elle est traitée comme une hypothèse à vérifier, non comme un acquis : quels usages fonctionnent, dans quels systèmes, avec quels gains mesurés, quelles contraintes métier et quel coût de calcul ?

Chapitre 8 - Quand l'IA devient utile : intégration métier, robustesse et résistances

Le chapitre en une phrase

L'IA industrielle exige une double robustesse, de la solution et du contexte métier. Elle montre qu'elle peut être utile quand elle est intégrée à des métiers, des systèmes physiques, des données fiables et des décisions supervisées. En revanche, elle ne confirme pas le récit d'une IA qui résoudrait mécaniquement ses propres impacts.

La réciprocité IA-énergie, posée dans l'introduction générale et développée au fil de cette partie, ne se vérifie pas dans les promesses générales de l'IA. Elle se vérifie dans sa capacité concrète à s'intégrer dans des systèmes énergétiques réels. C'est précisément le terrain de l'« *AI for Energy* ».

Pourquoi l'énergie est un test sérieux pour l'IA

L'énergie est un système physique, critique, régulé et supervisé par des métiers experts. Y appliquer l'IA ne revient donc pas à l'appliquer à un domaine parmi d'autres : l'avantage stratégique ne se joue pas seulement dans le modèle, mais dans la capacité à l'intégrer sans fragiliser le système. C'est ce qui fait de l'« *AI for Energy* » un terrain de vérité : robustesse, explicabilité, intégration opérationnelle et mesure du gain net y deviennent des conditions de crédibilité.

8.1 L'IA utile est souvent une IA qui disparaît dans l'outillage métier

Les entretiens invitent à déplacer le regard : dans les organisations énergétiques et industrielles, l'IA la plus utile n'est pas toujours celle qui se voit. Elle est souvent intégrée dans un automate, un logiciel industriel, un jumeau numérique, un module de maintenance, un système de documentation, une interface opérateur ou une chaîne d'aide à l'exploitation. Elle ne se présente pas forcément sous la forme d'un chatbot spectaculaire, mais se loge dans des chaînes de décision et d'exploitation déjà existantes. Dans les bâtiments et les usages domestiques, cette logique d'intégration peut être très concrète. Un interviewé travaillant chez un équipementier énergétique l'illustre ainsi : « *On peut relier l'IA à n'importe quel capteur de la maison via notre passerelle, connectée au cloud ou au compteur Linky, pour optimiser la consommation d'énergie.* »

Cette IA « discrète » est parfois plus structurante. Elle ne vaut pas par sa nouveauté, mais par sa capacité à s'insérer dans un processus, comprendre une contrainte physique, améliorer une décision opérationnelle et rester explicable pour les équipes. Cette lecture invite une fois de plus à ne pas confondre adoption de l'IA et visibilité de l'IA.

8.2 Réseaux électriques : optimiser sous contraintes fortes

Les réseaux électriques constituent un terrain majeur d'application de l'IA : prévision, détection d'anomalies, reconfiguration, réduction des pertes, amélioration des profils de tension, gestion de

contraintes et arbitrages d'exploitation. Mais ces usages s'inscrivent dans un système où fréquence, puissance réactive, impédance, charges inductives ou capacitatives, stabilité et règles d'exploitation imposent des exigences fortes de robustesse et de validation.

Une IA qui suggère une optimisation sur un réseau critique doit être testée, validée, explicable, traçable, supervisée et compatible avec les règles d'exploitation. *« On commence à penser l'IA pour attaquer des gros métiers de l'entreprise, mais pour certains métiers, c'est compliqué de leur demander de faire un pas de côté pour un projet IA. »*, précise une interviewée. La prudence d'acteurs, tels des opérateurs ou dispatcheurs, ne doit donc pas être interprétée comme résistance au changement. Elle peut être une exigence professionnelle légitime.

8.3 Bâtiments, chauffage, climatisation, recharge, *microgrids* : l'IA face aux usages réels

Les entretiens évoquent aussi l'optimisation énergétique des bâtiments : chauffage, climatisation, occupation variable des salles, météo, inertie thermique, fenêtres ouvertes, vacances, piscines ou équipements à forte inertie. Ces cas rappellent que l'IA agit sur des comportements, des horaires, des bâtiments imparfaits, des capteurs, des équipements existants.

« Véhicules électriques, le chargeur est basique. L'IA intervient dans la gestion de la charge et de la donnée résultante : optimiser la charge en fonction des critères du réseau, choisir le moment de charge le plus approprié par rapport à la facture et à la consommation de la maison. On peut aussi utiliser la batterie de la voiture pour alimenter la maison. Dans ce type de microgrid domestique, l'IA joue un rôle fondamental. », entretien avec un équipementier énergétique qui apporte également un chiffre à l'appui : *« Les gains dépendent beaucoup de la configuration du foyer, on peut avoir un gain de 25%. »*

Ces cas montrent que l'IA peut produire des gains dans des systèmes concrets où interviennent météo, comportements, équipements existants, contraintes réseau, usages domestiques et conditions d'exploitation. Mais ils rappellent aussi que l'IA utile dépend de données historiques, de modèles adaptés, d'une intégration opérationnelle et d'une capacité à mesurer le gain réel.

8.4 IA générative : utile, mais pas toujours centrale dans l'énergie

Les entretiens suggèrent que l'IA générative peut avoir des usages utiles dans les organisations énergétiques : documentation, assistance, interfaces opérateurs, recherche d'information, aide à la formation, explicitation de procédures, accès à des connaissances métier. Une interviewée d'une organisation énergétique évoque ainsi la *« génération d'études de raccordement par IA générative [, l'] assistance à la décision pour la conception de réseaux »* ou encore *« des modèles de prévision utilisés « depuis longtemps » »*.

Mais elle n'est pas nécessairement le cœur des usages les plus critiques en énergie.

Un autre interviewé résume cette diversité d'usages : *« Computer vision, maintenance prédictive, optimisation, GenAI spécifique pour les usages - par exemple des chatbots intégrés dans le logiciel pour aider les opérateurs. »*.

Dans les systèmes industriels, un modèle ne suffit pas. Il faut un environnement autour du modèle : méthodes, règles, supervision, évaluation, traçabilité, explicabilité, robustesse, documentation, garde-fous, reproductibilité. La différenciation européenne pourrait aussi se jouer dans cette

capacité à rendre l'IA intégrable et fiable dans des systèmes critiques, plutôt que dans la seule course au modèle généraliste le plus puissant.

Les retours recueillis convergent ainsi vers une hiérarchie assez nette des usages : l'essentiel de la valeur opérationnelle actuelle repose sur des systèmes d'optimisation, de prédiction, de détection et d'aide à la décision. L'IA générative trouve encore principalement sa place dans des fonctions tertiaires ou d'assistance. Le potentiel des systèmes agentiques reste, à ce stade, plus ouvert.

8.5 L'intégration industrielle : un verrou plus critique que le modèle seul

Comme nous venons de le voir, dans les systèmes énergétiques et industriels, le verrou n'est pas seulement le modèle d'IA. Il est souvent dans l'intégration. Une IA performante en laboratoire a peu de valeur si elle ne peut pas être reliée à des données fiables, à des équipements existants, à des logiciels industriels, à des automates, à des jumeaux numériques, à des règles de sûreté, à des procédures métier et à une supervision humaine.

L'état réel des données constitue ainsi un premier verrou : plans anciens ou incomplets, actifs re-numérisés, PLM (*product lifecycle management*) difficiles à maintenir, bases documentaires hétérogènes, données silotées, capteurs imparfaits, historiques incomplets. L'enjeu stratégique n'est pas seulement de brancher un modèle sur une base documentaire, mais de maintenir dans le temps un patrimoine de données suffisamment fiable, structuré et exploitable pour permettre l'IA.

Plusieurs entretiens formulent ce verrou très concrètement. Un acteur industriel souligne que l'IA ne peut pas compenser une donnée mal qualifiée : « *Si la donnée n'est pas fiable, l'IA ne corrige pas le problème ; elle l'accélère.* » Un autre insiste sur l'écart entre démonstrateur et exploitation : « *Le prototype peut marcher, mais le vrai sujet est de le raccorder aux systèmes existants, aux règles métier, aux contraintes de sécurité et aux personnes qui vont devoir l'utiliser.* »

Le second verrou est celui des systèmes complexes : réseaux électriques, contrôle-commande, bâtiments, procédés industriels, *microgrids*, robotique, logiciels industriels, électronique de puissance, cybersécurité et supervision humaine. Dans ces environnements, la confiance vient aussi de l'environnement qui encadre les modèles : explicabilité, traçabilité, reproductibilité, journalisation, documentation, garde-fous, contrôle des données et capacité de retour arrière. Cette exigence devient encore plus forte avec l'IA agentique, lorsque le système enchaîne des actions ou des recommandations.

Ce point rejoint l'analyse du développeur assembleur : lorsque les systèmes sont construits par assemblage de briques - modèles, API, bases documentaires, *workflows*, règles métier, interfaces et modules de supervision -, la maîtrise stratégique se déplace vers l'architecture d'ensemble.

La couche «intégration», facteur de différenciation de l'Europe

L'Europe peut construire une différenciation forte sur cette couche d'intégration : non pas seulement produire des modèles, mais savoir les insérer dans des environnements industriels fiables, gouvernés, interopérables et auditables.

Selon les entretiens, des acteurs comme Dassault Systèmes et Schneider Electric incarnent différentes formes de cette maîtrise : logiciels industriels, jumeaux numériques, automatisation, contrôle-commande, systèmes complexes, interaction entre données, énergie, machines et supervision humaine. Ce sont des signaux d'une capacité européenne possible : faire de l'intégration industrielle, de la confiance opérationnelle et de l'interopérabilité une couche stratégique.

L'entretien avec un chercheur en robotique renforce ce point : l'enjeu n'est pas de faire fonctionner une IA isolée, mais de l'inscrire dans une chaîne où interagissent perception, capteurs, robotique, décision, sécurité, environnement physique et opérateurs humains. L'interviewé le formule ainsi : « En robotique, le problème n'est pas seulement l'algorithme. C'est l'intégration dans un système réel : les capteurs, l'environnement, les contraintes physiques, la sécurité, l'interaction avec l'humain. Le modèle seul ne suffit pas. »

Ce verrou d'intégration déplace directement la question des compétences : il ne suffit pas de former des spécialistes IA, il faut former des intégrateurs capables de relier les systèmes.

Cloud de confiance, IA de confiance : l'intégration comme environnement de preuve

Les entretiens montrent aussi que l'intégration industrielle ne consiste pas seulement à brancher un modèle sur une infrastructure cloud. Elle suppose de construire un environnement de confiance autour du modèle : gouvernance, traçabilité, explicabilité, reproductibilité, contrôle des accès, localisation, sécurité et capacité d'audit. Un acteur du cloud de confiance le formule ainsi : « *L'IA agentique oblige à forcer des IA de confiance. L'IA générative doit être explicable, traçable et reproductible. Il faut créer un environnement autour du LLM.* »

Ce même entretien relie directement cette exigence aux choix d'infrastructure : « *On parle de cloud de confiance et on n'emploie pas le mot souverain, car il a été dévoyé par les hyperscalers.* » L'interviewé distingue ainsi deux offres : d'un côté, une offre cloud avec des sites français déjà SecNumCloud ; de l'autre, Bleu, coentreprise française associant Orange et Capgemini¹⁰², qui propose des services Microsoft Azure et Microsoft 365 dans un environnement opéré et hébergé en France, en cours de qualification SecNumCloud. Ce dispositif peut renforcer la maîtrise opérationnelle et juridique des données et des traitements, sans supprimer la dépendance technologique aux produits, aux mises à jour et à la feuille de route de Microsoft. Il relève donc davantage d'un cloud de confiance que d'une maîtrise souveraine de l'ensemble de la chaîne. Cette distinction montre que la souveraineté ne se résume pas au lieu d'hébergement, mais combine exploitation, gouvernance juridique, contrôle opérationnel, certification, contrat, logiciel et dépendance technologique.

Elle éclaire aussi le lien entre frugalité et obsolescence matérielle. La logique qui consiste à proportionner le calcul à la demande réelle, détaillée en 10.8, vaut aussi pour les choix d'investissement matériel. Elle rejoint l'idée que la frugalité industrielle est une discipline d'investissement, de durée de vie matérielle et de calibration de la capacité.

¹⁰² <https://www.capgemini.com/fr-fr/actualites/communiqués-de-presse/capgemini-et-orange-annoncent-le-lancement-des-activités-commerciales-de-bleu-leur-future-plateforme-de-cloud-de-confiance/> ; https://commission.europa.eu/document/download/25695177-8073-4ce3-bf81-eb816dc6b468_en?filename=Report+on+the+first+periodic+review+of+the+functioning+of+the+adequacy+decision+on+the+EU-US+Data+Privacy+Framework.pdf

Former des intégrateurs, pas seulement des spécialistes IA

Les compétences critiques ne se limitent pas aux *data scientists*, *data engineers*, *prompt engineers* ou experts *cloud*. Les entretiens pointent le besoin de profils capables de relier IA, données, énergie, data centers, réseaux, logiciels industriels, métiers, cybersécurité, environnement, réglementation et exploitation. Ces profils traduisent un besoin métier en architecture technique, comprennent les contraintes d'un réseau électrique, qualifient une donnée, évaluent un modèle, négocient avec un fournisseur cloud, suivent les impacts environnementaux, dialoguent avec les territoires et maintiennent la confiance dans des systèmes complexes. Il ne s'agit pas forcément de compétences portées par une seule personne, il peut s'agir de compétences d'un collectif entraîné à l'interaction et au raisonnement inter-métiers. Des méthodes comme le codesign¹⁰³ peuvent aider.

Ces compétences deviennent une infrastructure critique. Sans intégrateurs capables de faire dialoguer IA, énergie, réseau, cloud, sécurité, environnement et métiers, les infrastructures de calcul peuvent rester sous-utilisées, mal gouvernées ou dépendantes d'acteurs extérieurs. Les ontologies dynamiques de compétences peuvent aider à dépasser les taxonomies figées : plutôt que de classer les métiers une fois pour toutes, elles cartographient en continu les compétences, les relations entre métiers, les technologies, les infrastructures et les besoins émergents.

Des modèles d'écoles industrielles, de reconversions internes ou de formation par le terrain, évoqués dans les entretiens à partir d'exemples comme Safran¹⁰⁴ ou SNCF, peuvent être source d'inspiration, pour des logiques de professionnalisation hybride.

Le mentorat inversé numérique mis en place au Ministère des Armées¹⁰⁵ offre également un exemple utile : il inverse la relation classique de transmission pour faire remonter des compétences numériques depuis des profils plus jeunes vers des autorités plus expérimentées. À l'ère de l'IA, ce mentorat peut aussi être pensé dans l'autre sens : des métiers vers le numérique. Les experts métier peuvent aider les profils numériques à comprendre les spécificités et contraintes physiques, opérationnelles et organisationnelles qui donneront un surplus d'efficacité et d'innovation aux IA. Dans le contexte IA-énergie, ce type de dispositif illustre une idée importante : l'intégration ne dépend pas seulement de formations descendantes, mais aussi de boucles d'apprentissage entre générations, métiers, niveaux hiérarchiques et cultures techniques.

8.6 Ce qui résiste : données, confiance, responsabilité, temps organisationnel

Les entretiens convergent sur plusieurs sources de résistance : qualité des données, silos métiers, cybersécurité, explicabilité, responsabilité humaine, culture opérationnelle, coût d'intégration, dépendance éditeur, difficulté à mesurer les gains et prudence des opérateurs. Ces résistances déterminent si l'IA produit réellement de la valeur.

¹⁰³ **Codesign** : démarche de conception menée conjointement avec plusieurs parties prenantes, qui participent à la définition du besoin, à l'élaboration de la solution et à son ajustement. La clé de réussite est la possibilité d'intégrer tous les points de vue.

¹⁰⁴ <https://www.leparisien.fr/oise-60/je-navais-jamais-fait-de-chaudronnerie-chez-safran-compiegne-on-forme-les-salaries-en-reconversion-24-04-2024-LBTQ267LZVCBNAQ3AFOBULG4.php> ; <https://www.safran-group.com/fr/videos/immersion-au-coeur-ecoles-industrielles-safran-1>

¹⁰⁵ <https://www.defense.gouv.fr/ema/actualites/4e-saison-du-mentorat-inverse-numerique-est-lancee>

Un interviewé formule cette prudence sur la maintenance prédictive : « *La maintenance prédictive, ça fait vingt ans que j'en entends parler. Je n'ai pas vu de vrais cas d'usage significatifs. Peut-être qu'avec l'IA, ça va changer.* »

Un autre entretien rappelle que la couche numérique existante est souvent elle-même mal entretenue avant même de parler d'IA : « *80% des installations de GTB [gestion technique du bâtiment] sont à l'arrêt ou mal exploitées.* »

Ces constats conduisent à éviter l'expression « déployer l'IA » comme si le déploiement suffisait. Dans l'énergie, déployer signifie intégrer à des systèmes existants, valider avec les métiers, prouver les gains, documenter les limites, maintenir dans le temps, gérer les mises à jour, sécuriser les accès et conserver une supervision humaine.

Le temps organisationnel, un incontournable

Les entretiens invitent à ajouter une ressource souvent absente du débat IA-énergie : le temps organisationnel. Déployer l'IA ne consiste pas seulement à appeler un modèle ou à raccorder une infrastructure ; il faut cadrer le besoin, qualifier la donnée, tester, documenter, former, contrôler, corriger, sécuriser, superviser, mesurer les gains et maintenir les systèmes. Une organisation qui n'a pas le temps d'impliquer les métiers risque de produire des usages IA techniquement disponibles mais faiblement appropriés, peu robustes et difficiles à évaluer.

Un entretien illustre ce point : « *Le problème ici, c'est qu'on ne les sort pas de leur poste, pour ne pas perdre de temps. Mais on va leur proposer ensuite des solutions IA faites sans eux, pas forcément adaptées à leurs contraintes et on leur fait perdre du temps pour s'adapter à une solution peut-être moins performante que celle qui prévalait.* » Ce verbatim rappelle que le temps métier n'est pas un coût périphérique : il conditionne la pertinence, l'appropriation et la performance réelle de l'outil.

Le doute dynamique : une compétence stratégique

Temps organisationnel et compétence d'intégration supposent une compétence plus profonde : le doute dynamique. Il ne s'agit ni de scepticisme paralysant, ni de rejet de l'innovation. Il s'agit d'une capacité à questionner les évidences du moment, à tester les hypothèses dominantes, à recomposer l'existant et à explorer des trajectoires que les acteurs dominants n'ont pas encore rendues légitimes.

Cette compétence est essentielle pour innover, construire la souveraineté et réduire les dépendances. Une position de suiveur conduit souvent à considérer comme impossible ce qui n'a pas encore été validé par les grands acteurs américains ou anglo-saxons. Elle pousse à reprendre leurs catégories, leurs priorités, leurs arguments et parfois leurs angles morts. Or l'histoire récente de l'IA montre que certaines ruptures ne proviennent pas toujours d'une invention radicalement nouvelle, mais d'une recomposition efficace de méthodes existantes.

L'exemple de DeepSeek (IA chinoise) a rendu visible ce point : des options réputées secondaires, insuffisantes ou non compétitives - *pruning*, distillation, mixture of experts, optimisation des données, choix d'architecture, contraintes matérielles assumées - pouvaient, combinées autrement, produire des résultats significatifs. La leçon est de comprendre que l'innovation peut aussi venir d'une contrainte assumée, d'un assemblage différent, d'une optimisation méthodique et d'une capacité à contester les évidences industrielles du moment.

Le doute dynamique consiste donc à franchir le « mur de l'impossible » avant qu'il ne soit franchi par d'autres. Pour l'Europe, cette compétence est particulièrement importante. Elle peut permettre d'explorer des zones où la puissance brute n'est pas le seul critère : intégration industrielle, systèmes critiques, frugalité, explicabilité, interopérabilité, robustesse, souveraineté d'usage et qualité de la donnée. C'est précisément dans ces zones que l'Europe peut trouver une différenciation, à condition de ne pas attendre que les acteurs dominants les désignent eux-mêmes comme stratégiques.

Le doute dynamique est une compétence d'innovation : oser formuler de nouvelles hypothèses, tester des combinaisons techniques non évidentes, refuser les dépendances de langage et reconstruire des trajectoires adaptées à nos contraintes réelles.

Ce doute dynamique invite aussi à élargir nos sources d'apprentissage. L'Europe s'est longtemps formée aux écoles de management anglo-saxonnes ; elle a aussi appris, à d'autres moments, des méthodes japonaises de production, de qualité et d'amélioration continue. Elle pourrait aujourd'hui observer plus attentivement certaines approches chinoises, non pour les copier, mais pour comprendre leur capacité à combiner contraintes matérielles, vitesse d'exécution, recomposition de briques existantes, discipline industrielle et capacité de passage à l'échelle. Il ne s'agit pas d'importer un modèle culturel ou politique, mais de ne pas se priver d'un apprentissage stratégique sur les formes contemporaines d'innovation sous contrainte.

8.7 Conclusion - L'IA industrielle demande une double robustesse

L'IA industrielle demande une double robustesse : **robustesse de la solution et robustesse du contexte métier**. La première concerne la performance, la stabilité, la reproductibilité, l'explicabilité et la capacité à éviter des erreurs. La seconde concerne l'environnement dans lequel la solution agit : qualité des données, règles de sûreté, contraintes physiques, procédures métier, responsabilité humaine, cybersécurité, maintenance, réglementation et capacité d'intervention en cas d'écart.

Cette distinction est centrale dans l'énergie. Une IA peut produire une recommandation statistiquement performante et rester inutilisable si elle ne s'inscrit pas dans les règles d'exploitation, les contraintes électrotechniques, les procédures de sécurité ou la culture métier des opérateurs. À l'inverse, une IA moins spectaculaire mais mieux intégrée, traçable, supervisée et compréhensible peut produire plus de valeur dans un système critique.

L'IA dans l'énergie existe déjà. Elle montre que l'IA peut être utile lorsqu'elle est intégrée à des métiers, des systèmes physiques, des données fiables et des décisions supervisées.

Le chapitre suivant descend dans les usages où cette promesse commence à être documentée : prévision, maintenance, optimisation, contrôle qualité, bâtiments, *microgrids*, assistance aux métiers et maîtrise thermique.

Chapitre 9 - Là où l'IA produit déjà des gains mesurables

Le chapitre en une phrase

Cinq domaines documentés - prévision, maintenance prédictive, optimisation, contrôle qualité, assistance métier - montrent que l'IA utile dans l'énergie est souvent non spectaculaire, mesurable et bornée. Trois leçons transversales : cette IA est structurellement frugale par construction, le bénéfice n'est crédible que s'il est mesuré et l'intégration métier prime sur la qualité du modèle.

9.1 La méthode : partir des usages, pas des promesses

L'IA dans l'énergie souffre d'un paradoxe : les discours de rupture sont omniprésents, mais les cas d'usage matures et documentés sont beaucoup plus rares que la communication sectorielle ne le laisse supposer. Comme déjà noté en 8.6, notamment avec le verbatim sur la maintenance prédictive, l'écart entre discours et pratique documentée est un thème récurrent des entretiens.

Cette lucidité permet de distinguer ce que les organisations font réellement de ce qu'elles projettent de faire et elle évite la justification circulaire : « l'IA consomme, mais elle optimisera le système ».

Ce chapitre part donc des usages observés plutôt que des promesses. Il s'appuie sur les entretiens menés avec des acteurs de la chaîne de valeur de l'énergie pour identifier les cas où l'IA apporte déjà une valeur mesurable, les domaines où la maturité reste plus contrastée et les conditions qui rendent ces gains crédibles.

Comme déjà évoqué, les usages les plus crédibles de l'IA appliquée à l'énergie reposent souvent sur du *machine learning* classique, de la prévision, de l'optimisation, de la détection d'anomalies, des capteurs, des automates et des modèles métiers intégrés aux systèmes industriels.

Ce chapitre ne cherche pas à établir un bilan national agrégé entre la consommation des data centers (chapitres 1 et 2) et les gains documentés ici : les deux relèvent de périmètres et d'échelles différents et leur addition produirait précisément le type de comparaison trompeuse contre laquelle ce rapport met en garde ailleurs (voir 6.6). Le choix méthodologique est donc d'instruire chaque usage à son échelle, plutôt que de risquer un solde national qu'aucune donnée disponible ne permet aujourd'hui d'établir avec rigueur.

9.2 Des usages inégalement matures

Les usages documentés dans les entretiens peuvent être regroupés en cinq familles. Elles ne présentent pas le même niveau de maturité : la prévision est le domaine le plus installé ; la maintenance prédictive reste plus contrastée ; l'optimisation opérationnelle est le lieu où les gains économiques et environnementaux se rejoignent le plus clairement ; le contrôle qualité est souvent mesurable ; les outils collaboratifs relèvent davantage de l'assistance interne que du pilotage critique des systèmes énergétiques.

Domaine 1 - Prévision et anticipation

Cas d'usage	Ce que l'entretien documente	Type d'effet dominant inspiré de la grille 1.4 ¹⁰⁶
Prévision de consommation à maille fine	Deep learning sur séries temporelles multivariées.	Remplace une prévision moins fine par une méthode plus précise, sur une infrastructure de calcul déjà existante (données Linky)
Prévision de production solaire décentralisée	Gestion de l'intermittence à l'échelle locale. La Cour des comptes évoque 95% des cas ¹⁰⁷ .	Améliore un processus de gestion de l'intermittence déjà en place
Prévision des prix de marché	Optimisation des décisions d'achat/vente d'électricité. Séries temporelles, données météo, données de production	Optimise une décision déjà prise aujourd'hui par d'autres moyens
Anticipation des impacts climatiques (Windi)	Anticiper les impacts des événements climatiques (vent, tempête) sur le réseau. Identification préventive des zones vulnérables	Crée une capacité nouvelle, sans équivalent métier préexistant documenté dans l'entretien
Prévision de la demande de raccordement	Modélisation des profils de montée en charge, trajectoires probables des projets (entretien RTE)	Outils une instruction aujourd'hui largement humaine et manuelle.

Plusieurs verbatims confirment que la prévision constitue un domaine ancien et déjà opérationnel de l'IA dans l'énergie. Un interviewé rappelle que la prévision à maille fine n'était pas possible sans la disponibilité de nouvelles données : « Avec Linky, la maille de prévision basse tension est devenue possible. Avant, on n'y serait pas arrivés avec plusieurs variables et des mailles fines. » Un autre insiste sur le rôle central du réseau de distribution dans l'intégration des renouvelables : « La très grande majorité des raccordements ENR passent par le réseau de distribution. » Ces propos montrent que l'IA utile repose ici moins sur un modèle spectaculaire que sur la disponibilité de séries temporelles, de données météo, de données de production, de données de consommation et d'un besoin métier clairement défini.

Ces usages ont en commun d'être conçus pour la plupart « depuis longtemps » selon les entretiens, sur des données structurées, avec des métriques de performance claires et des bénéfices mesurables. Ils relèvent du machine learning classique - régression, classification, séries temporelles - bien avant l'ère des LLM.

Un cas d'usage plus discret, mais potentiellement structurant, concerne la production et la comparaison d'études. L'IA peut aider à générer des premières versions d'études exploratoires de raccordement, repérer des écarts entre études successives, documenter les hypothèses et faciliter l'accès à des analyses pour des profils qui ne codent pas. Sans remplacer l'expertise métier, ce type

¹⁰⁶ Grille 1.4 : substitution = amélioration ou remplacement d'un usage existant ; addition = création d'un usage ou d'un volume supplémentaire ; relocalisation = déplacement d'un usage ou d'une infrastructure ; effet d'appel = stimulation de nouveaux usages par l'offre, le coût ou l'accessibilité.

¹⁰⁷ Le réseau de distribution est au cœur de l'intégration des énergies renouvelables : la Cour des comptes indique que le raccordement des énergies renouvelables au réseau de distribution se fait dans 95% des cas. Enedis indique par ailleurs que 51 GW d'ENR étaient connectés au réseau de distribution fin 2025, produits par 1,3 million de producteurs.

Sources : https://www.ccomptes.fr/sites/default/files/2025-02/ROD-2025-205-Rapport_th--matique_r--gional_r--seaux-DE.pdf et <https://www.enedis.fr/presse/electrification-de-la-france-584-000-raccordements-et-mises-en-service-realises-par-enedis>

d'usage peut augmenter la capacité d'instruction. Il impose toutefois une condition forte : conserver la traçabilité des hypothèses, la revue humaine et la responsabilité de l'expert. Ce point sera repris dans le Domaine 5, consacré aux outils collaboratifs et à l'assistance interne.

Domaine 2 - Maintenance prédictive et gestion des actifs

La maintenance prédictive est un deuxième domaine d'usage fréquemment cité, mais de maturité plus contrastée. Elle consiste à détecter des anomalies ou prévoir des pannes avant qu'elles ne se produisent, à partir de données de surveillance réelles plutôt que d'un calendrier fixe, pour intervenir de manière ciblée plutôt que systématique ou curative.

Par exemple, la maintenance du réseau s'appuie sur les données Linky pour détecter des anomalies de tension, des comportements inhabituels ou des précurseurs de pannes. La détection des pertes d'électricité et de la fraude est mentionnée comme un usage déployé. Les jumeaux numériques de centrales et les outils de surveillance des actifs de production sont documentés comme des prolongements de la data science existante.

Un interviewé le formule ainsi : « *Data center pour l'exploitation centrale, l'optimisation de la maintenance, la surveillance des assets de production. L'IA, ce n'est pas nouveau : c'est aussi le prolongement des sciences de la donnée, des statistiques, des prévisions, de la gestion des connaissances. Prolongement de la data science : méta-modélisation, bases de données qu'on manipule, détection d'anomalies, jumeaux numériques.* »

Le secteur a amorcé un virage data depuis le déploiement de Linky, sur lequel s'appuient aujourd'hui des solutions de maintenance prédictive (voir 10.1).

Un élément clé issu des entretiens concerne le traitement d'image, ou computer vision, pour l'inspection des actifs physiques : pales éoliennes, barrages, lignes électriques. Les entretiens signalent des usages d'analyse d'image pour détecter des défauts, inspecter des barrages ou analyser des pales éoliennes. Un équipementier décrit également des cas de computer vision sur des produits de grande consommation pour détecter des défauts et réduire les rebuts, avec un bénéfice environnemental mesurable (réduction des déchets chimiques). Même lorsque l'exemple ne relève pas directement d'un actif énergétique, il illustre un mécanisme transférable : détection visuelle, réduction des défauts, limitation des rebuts et mesure du gain.

Domaine 3 - Optimisation opérationnelle

L'optimisation en temps réel ou quasi temps réel des systèmes énergétiques est le troisième domaine. Il couvre des applications allant des jumeaux numériques de centrales à la gestion des ressources énergétiques distribuées.

Les cas Schneider Electric : MicroGridAdvisor, WiserSystem, Automation Expert

Les entretiens et publications disponibles font ressortir plusieurs cas où l'optimisation opérationnelle produit un double bénéfice potentiel : économique et environnemental.

MicroGridAdvisor : Schneider Electric¹⁰⁸ documente publiquement des gains de coûts énergétiques de 10% à 30% sur ses Smart Factories / Smart Distribution Centers avec

¹⁰⁸<https://www.se.com/ww/en/work/campaign/smart-factory/> ;

<https://www.se.com/ww/en/about-us/newsroom/news/press-releases/schneider-electric-launches-ai-powered-home-energy-management-feature-for-wiser-home-66d6bac4d6b0eff3580dc113/> ; <https://blog.se.com/digital-transformation/artificial-intelligence/2025/11/14/engineering-at-scale-how-ai-is-transforming-plc-coding/>

EcoStruxure.

WiserSystem : gestion de l'énergie domestique : chauffage, véhicule électrique, production solaire, ballon d'eau chaude. L'algorithme optimise les arbitrages selon le profil de consommation, les prix, la production photovoltaïque et les capacités de stockage. Le système part de profils pré-entraînés et s'adapte au profil réel dans le temps.

Automation Expert : assistant IA pour la programmation de contrôleurs PLC, c'est-à-dire des **automates programmables industriels** qui pilotent des équipements physiques dans les usines, bâtiments ou infrastructures : moteurs, vannes, pompes, températures, vitesses de rotation, séquences de production. L'ingénieur décrit fonctionnellement ce qu'il veut : prendre les températures, lire les vitesses de rotation, ouvrir les vannes de x%. L'IA génère le code, puis l'ingénieur valide et intègre.

Ces trois produits ont comme point commun le potentiel de double bénéfice, à la fois économique et environnemental.

Domaine 4 - Contrôle qualité et sûreté

Le contrôle qualité automatisé est un domaine où les résultats sont mesurables et les délais de validation raisonnables. Plusieurs cas déployés, par exemple : détection de défauts par computer vision, contrôle de qualité sur des produits de grande consommation avec réduction des rebuts.

Dans les systèmes critiques - barrages, centrales, réseaux - les entretiens signalent aussi l'usage croissant d'algorithmes de détection d'anomalies, de méta-modélisation et de jumeaux numériques. La modélisation de procédés bas-carbone, dont le captage de CO₂ est citée comme un usage en développement interne : des algorithmes proches de ceux utilisés pour les protéines (AlphaFold) pour générer des variantes de procédés et trouver des optimums économiques et techniques.

Domaine 5 - Assistance interne, interfaces métier et systèmes embarqués

Le cinquième domaine est le plus récent et le plus lié à l'IA générative. Il couvre l'ensemble des usages internes qui améliorent la productivité des équipes sans nécessiter d'intégration dans les systèmes opérationnels critiques.

Usage	Ce que documentent les entretiens	Type d'effet dominant inspiré de la grille 1.4
Génération d'études de raccordement	Réduction du temps de production d'études exploratoires. « <i>Les gens qui ne savent pas coder vont pouvoir produire des études</i> », interviewé travaillant dans une organisation énergétique.	Remplace une production manuelle d'études
Traitement des réclamations clients	Assistance à la rédaction de réponses, classification automatique. Prolongement de la data science existante	Remplace un traitement humain existant

Usage	Ce que documentent les entretiens	Type d'effet dominant inspiré de la grille 1.4
Interface opérateur	<i>Chatbot guidé pour les opérateurs de réseau électrique : « je vois telle panne, la procédure c'est ça. 1 serveur, ~300 k€, 10 opérateurs simultanés »</i> (ordre de grandeur issu d'un entretien)	Ajoute un nouvel outil et une nouvelle infrastructure dédiée (serveur, ~300 k€), sans équivalent préexistant
Documentation automatique de code PLC	L'ingénieur modifie un contrôleur, l'IA génère automatiquement l'explicabilité du code. Utile pour la traçabilité en agro-industrie et pharmacie	Remplace une documentation manuelle
Assistance à la veille et aux offres	Réponse aux appels d'offres avec l'IA : plans d'installation électrique, métré, listes d'équipements, tarifs. « <i>Gain de productivité (est-ce un fantasme ?)</i> »	Remplace une production manuelle de réponses aux appels d'offres
Systèmes experts embarqués	Collecte de données terrain (edge), optimisation de la consommation, pilotage des véhicules électriques, home-to-grid	Création de nouveaux dispositifs déployés en périphérie, capacité de calcul supplémentaire

Ces exemples sont mobilisés comme cas permettant d'observer les conditions de crédibilité d'une IA énergétique : finalité claire, intégration système, indicateur de gain, supervision humaine et contexte d'usage défini.

Ce domaine illustre un point souligné dans plusieurs entretiens : à l'heure actuelle, l'IA générative est plus utile dans les tâches d'appui aux métiers (documentation, synthèse, assistance, génération de code) que dans les systèmes opérationnels critiques où l'explicabilité et la robustesse sont non négociables.

9.3 L'optimisation énergétique du bâtiment

Le cas de l'optimisation énergétique du bâtiment mérite d'être repris ici sous l'angle de la mesure des gains. Il illustre une logique différente des grandes organisations énergétiques : celle d'un équipementier électrique qui intègre l'IA à ses produits pour offrir à ses clients une optimisation de leur consommation énergétique.

Le cas d'usage central décrit dans les entretiens est l'optimisation du microgrid résidentiel ou tertiaire : le contrôleur collecte les données de tous les capteurs de la maison ou du bâtiment (production solaire, consommation, véhicule électrique, ballon d'eau chaude, chauffage) et l'algorithme décide en temps réel des arbitrages : home-to-grid, charge du véhicule optimisée selon le prix réseau et l'état de la batterie, vente de surplus. Des gains de 25% sur la consommation énergétique sont cités¹⁰⁹, sous réserve de configuration équipée, de données disponibles et de conditions d'usage favorables.

« L'IA peut aider dans la recherche du point de fonctionnement idéal (même par anticipation) : optimisation, prévision, analyse du passé. Dans ce type de microgrid, l'IA joue un rôle fondamental. », entretien avec un équipementier.

¹⁰⁹ Il s'agit d'ordres de grandeur cités en entretien.

Ce cas illustre une fois de plus que l'IA utile aux systèmes énergétiques n'est pas nécessairement spectaculaire ni fondée sur un grand modèle. C'est souvent un algorithme d'optimisation situé, sur des données locales, dans un contexte défini. Sa frugalité est souvent structurelle.

D'autres entretiens confirment ce point et rappellent que l'IA du bâtiment n'agit pas seule : elle dépend de capteurs, d'équipements pilotables, de données locales, de prix de l'énergie, de production photovoltaïque éventuelle, de stockage et de comportements d'usage. Par exemple, un autre équipementier souligne les conditions matérielles de cette optimisation : « *Il faut des capteurs et des devices pour que ça marche.* »

9.4 Les freins réels à l'adoption

Ces cas montrent aussi pourquoi l'IA reste difficile à déployer dans l'énergie. Quatre freins structurels reviennent fréquemment dans les entretiens.

Frein 1 - La qualité et la disponibilité de la donnée

Les systèmes d'information industriels sont hétérogènes, les données sont silotées et leur qualité peut se dégrader dans le temps. Cette difficulté est amplifiée dans les organisations énergétiques par l'ancienneté des systèmes opérationnels : les données des centrales, des réseaux et des équipements existent, mais dans des formats et des systèmes très anciens.

Ce frein confirme un point traité dans la partie consacrée à l'IA frugale (§7.2) : la donnée n'est pas seulement un préalable technique, elle conditionne aussi la sobriété, la robustesse et la crédibilité des usages IA.

Frein 2 - La résistance des métiers cœur

Les métiers opérationnels critiques - ingénieurs de centrales nucléaires, opérateurs de réseau, dispatcheurs - sont prudents sur l'intégration de l'IA dans leurs processus. Cette prudence traduit des enjeux légitimes de sécurité, de réglementation et de responsabilité opérationnelle.

« *Quand on parlait de simulations numériques, mécanique des fluides, notre ingénierie disait : je ne veux pas de ça dans mon dossier. C'est la même chose aujourd'hui avec l'IA. Ça va prendre du temps. C'est une exigence de preuve et de maîtrise, légitime dans des systèmes où une erreur peut avoir des conséquences graves* », explique un expert R&D.

Frein 3 - L'explicabilité comme condition sine qua non

Cette exigence n'est pas spécifique à l'énergie, mais elle y est particulièrement forte. Le nucléaire, les barrages, le dispatching du réseau, la gestion des actifs de production : autant de contextes où « le système ne peut pas se tromper » et où la responsabilité reste humaine. L'explicabilité est une condition de la maîtrise opérationnelle.

Frein 4 - La fragmentation de la gouvernance IA

Les projets IA sont souvent portés par des équipes R&D ou des DSI sans lien direct avec les métiers. La valeur créée reste difficile à mesurer et à attribuer. Le passage de POC à déploiement généralisé, « *irriguer les processus* », selon l'expression d'un interviewé, suppose une articulation entre équipes techniques et métiers que beaucoup d'organisations n'ont pas encore construite.

Les entretiens mentionnent aussi la difficulté de dégager du temps métier pour les projets IA, c'est le temps organisationnel dont nous avons parlé en 8.6. Les sujets prioritaires à court terme -

exploitation, maintenance, réglementation - absorbent les capacités. L'IA devient un sujet de l'équipe R&D qui peine à atteindre le cœur opérationnel.

Cette difficulté rejoint la prudence des métiers cœur : l'IA ne peut pas atteindre le cœur opérationnel si elle reste portée à côté des métiers, par des équipes qui n'ont ni leurs contraintes quotidiennes ni leur responsabilité d'exploitation.

9.5 L'IA dans la maîtrise thermique des bâtiments : une illustration ciblée

Le projet ANR AI4HP¹¹⁰, sur des pompes à chaleur « enrichies » d'un algorithme IA, fournit une illustration de l'IA utile aux systèmes thermiques.

Le principe : les pompes à chaleur sont mal optimisées en pratique. Les paramètres sont fixés une fois pour toutes par l'installateur, sans prise en compte des besoins réels ni de la variabilité des usages. L'objectif du projet était d'optimiser la production d'eau chaude sanitaire (ECS) en anticipant les besoins via une chaîne IA embarquée, sans envoyer les données dans le cloud (contrainte RGPD allemande).

Résultats mesurés - Projet ANR AI4HP

En laboratoire (EDF R&D Climatron) : gain de COP¹¹¹ de +0,92 par rapport à la référence, pour une économie d'énergie de 40% - proche du maximum théorique de 43%. Ce gain s'explique par deux mécanismes : une température de tank plus basse, qui augmente le COP et une réduction des pertes thermiques. Une fois déduit le coût de calcul de l'algorithme lui-même (50 Wh/h, environ 1% de la consommation de référence), le gain net en laboratoire reste proche de 39%.

Sur le terrain (maison individuelle, Stiebel Eltron / Fraunhofer ISE, saison de chauffe 2024-2025) : gain de COP de +5% à +9% en conditions réelles. Consommation quotidienne de 31,5 kWh avec l'algorithme, contre 34,4 kWh en référence à température ambiante comparable - soit une économie d'environ 8%. Les déviations de confort sont réduites de moitié. Ce résultat, nettement inférieur au gain de laboratoire, est la mesure la plus fiable de ce que cette solution produit en conditions réelles.

Impact carbone net : environ 1 tCO₂/an pour un immeuble de 21 logements en eau chaude sanitaire collective et environ 32,5 kgCO₂/an pour une maison individuelle.

Architecture frugale : l'algorithme tourne localement sur un simple ordinateur portable de 75 W, sans envoi de données dans le cloud. Le modèle, léger (79 000 paramètres), prédit le volume journalier d'eau chaude à 24h avec une erreur moyenne de 10%.

Le modèle est petit, local, spécifique à l'usage, explicable et mesuré dans toutes ses dimensions, y compris le coût de calcul de l'algorithme lui-même. C'est une IA utile. Les limites sont aussi

¹¹⁰Projet franco-allemand, 2021-2025, impliquant le CEA List, LPNC, Fraunhofer ISE, EDF R&D et Stiebel Eltron, Frison L. et al., « **Evaluation of AI-driven adaptive control algorithms for optimizing heat pump efficiency in building heating and hot water supply: First practical insights** », Fraunhofer ISE / CEA List / EDF R&D / Stiebel Eltron, 2025. Voir notamment les sections 2.1.1, 3.1, 3.2 et 3.3 pour le GRU 79k, la MAE, les gains COP, la consommation de l'algorithme et les gains carbone. <https://publica.fraunhofer.de/entities/publication/3d98afc2-0ddb-48ad-8d16-6b025105666a>

¹¹¹ Le COP, ou coefficient de performance, mesure l'efficacité énergétique d'une pompe à chaleur : un COP de 3 signifie que la pompe restitue environ 3 kWh de chaleur pour 1 kWh d'électricité consommée. Une hausse du COP signifie donc que le système produit davantage de chaleur pour la même consommation électrique.

La **MAE**, ou mean absolute error, désigne l'erreur absolue moyenne d'un modèle de prévision : ici, une MAE de 10% sur le volume journalier prédit à 24 h signifie que l'écart moyen entre la consommation d'eau chaude prévue et la consommation observée est d'environ 10%. Ces deux indicateurs sont importants car ils permettent de relier la performance du modèle IA à un résultat physique mesurable : chaleur utile, consommation électrique, confort et économies d'énergie

documentées dans les publications du projet : les gains terrain (5-9%) sont inférieurs aux gains en laboratoire (40%), l'absence de données compresseur haute résolution a limité l'optimisation et des campagnes étendues sur différents types de bâtiments sont nécessaires avant de généraliser.

9.6 Ce que ce panorama dit de la Partie III

Les cas documentés dans ce chapitre permettent de formuler quatre leçons transversales.

Trois leçons transversales

1. **L'IA utile dans l'énergie est souvent non spectaculaire** : elle relève du machine learning classique : prévision, optimisation, détection d'anomalies, séries temporelles. Elle est rarement conversationnelle, multimodale ou agentique. Elle est située, spécifique et mesurable.
2. Cette spécificité a une conséquence directe pour la frugalité : une IA adaptée, bornée, intégrée dans un processus métier défini, sur des données délimitées, crée des conditions plus favorables à la frugalité et à un dimensionnement proportionné. C'est précisément le lien entre la Partie II et la Partie III : l'IA la plus utile aux systèmes énergétiques est fréquemment aussi l'IA la mieux dimensionnée.
3. **Le bénéfice est d'autant plus crédible qu'il est mesuré** : les cas qui résistent à l'examen sont ceux où le gain est chiffré, net du coût du calcul, comparé à un scénario de référence et stabilisé dans le temps. Les cas où le bénéfice est « promis » sans mesure sont ceux qui fragilisent le débat.
4. **L'intégration métier est la condition principale** : ni la qualité du modèle, ni la richesse des données, ni la performance de l'infrastructure ne suffisent. C'est l'articulation entre l'IA et les métiers avec des opérateurs qui comprennent, valident et maintiennent le système, qui détermine si l'IA produit un bénéfice réel ou reste au stade de POC.

Ces usages montrent que l'IA peut produire des gains réels dans les systèmes énergétiques. Mais leur existence ne suffit pas à fonder une doctrine. Le chapitre 10 formalise sept conditions de crédibilité opérationnelle : intégration réelle, qualité de la donnée, mesure du gain net, explicabilité-supervision-acceptabilité métier, robustesse dans le temps, frugalité par conception, maintien du jugement métier et des collectifs humains.

Chapitre 10 - Passer de la promesse à la preuve

Le chapitre en une phrase

Le chapitre formalise sept conditions de crédibilité opérationnelle - intégration réelle, qualité de la donnée, mesure du gain net, explicabilité, robustesse dans le temps, frugalité par conception, maintien du jugement métier - qui permettent de distinguer un usage IA crédible d'une promesse non vérifiée.

Le chapitre 9 a dressé le panorama de ce qui fonctionne dans les organisations énergétiques. Il a aussi présenté les freins : qualité de la donnée, résistance des métiers cœur, exigence d'explicabilité, fragmentation de la gouvernance. Ce chapitre instruit les conditions qui permettent à un usage IA dans les systèmes énergétiques de devenir crédible au-delà du POC, durablement intégré dans les systèmes réels et capable de produire un bénéfice mesurable, frugal et maîtrisé dans le temps.

Ces conditions interagissent entre elles. Une IA intégrée dans les systèmes réels mais non explicable risque de ne pas être acceptée par les opérateurs. Une IA explicable mais dont le gain n'est pas mesuré ne justifie pas l'infrastructure qu'elle mobilise. Une IA frugale par conception mais dont les collectifs humains s'érodent fragilise la résilience du système.

10.1 Le passage du POC au déploiement : la première rupture

Le premier constat des entretiens est à la fois simple et systématiquement sous-estimé : il y a un abîme entre POC (Proof of Concept) et déploiement opérationnel. Beaucoup d'organisations énergétiques ont des POC IA qui fonctionnent techniquement, sans produire de bénéfices réels, parce qu'ils tournent en parallèle des processus opérationnels.

Certaines organisations décrivent aujourd'hui le passage du POC à une logique d'« irrigation » des processus, ce qui est une bascule organisationnelle et technique profonde. Irriguer suppose d'avoir résolu les questions de qualité de donnée, d'intégration dans les SI existants, d'acceptabilité métier et de gouvernance continue. C'est là que beaucoup de projets s'arrêtent ou constatent des surcoûts inattendus.

« On est dans le virage de la data. Maintenant, on intègre de plus en plus d'IA : solutions IA-big data sur la maintenance du réseau, détection des pertes d'électricité, de la fraude. On est en train de passer de POC à on irrigue plein de processus », entretien avec une organisation énergétique.

La métaphore de l'irrigation est intéressante : elle implique une infrastructure (les données structurées), des canaux (les interfaces avec les systèmes opérationnels), un débit (la fréquence et la qualité des mises à jour du modèle) et des destinataires (les métiers qui intègrent les résultats dans leurs décisions). Chacun de ces éléments peut être un point de blocage.

Intégrer un algorithme dans un système opérationnel énergétique peut être techniquement, réglementairement et organisationnellement difficile. Certains systèmes ou interfaces d'opérateurs ont été conçus dans des logiques de sécurité et de robustesse qui rendent l'ajout d'une couche IA non trivial. Chez certains acteurs, les cycles de validation d'un système de contrôle-commande sont comptés en années. Cela signifie que l'intégration demande du temps, des ressources spécifiques et une volonté organisationnelle qui dépasse les équipes R&D.

« *Prolongement de ce qu'on a fait jusque-là. On peut imaginer que l'IA s'embarque dans les systèmes. Mais il faut l'explicabilité, la reproductibilité, le respect du réglementaire* », entretien avec une organisation énergétique.

La leçon du projet AI4HP¹¹² est ici pertinente. L'écart entre les gains mesurés en laboratoire (40% d'économie d'énergie) et les gains mesurés sur le terrain (5 à 9% sur une saison de chauffe) s'explique par des conditions physiques différentes et aussi, peut-être avant tout, par les contraintes d'intégration dans des systèmes réels : capteurs imparfaits, cycles réels, comportements d'utilisateurs non standardisés, absence de certaines données haute résolution (données compresseur). Le passage en conditions réelles est toujours un révélateur : les écarts ne disqualifient pas nécessairement le projet, mais ils indiquent les conditions à réunir pour améliorer, stabiliser ou généraliser le gain.

10.2 La qualité de la donnée : condition préalable à toute preuve

La qualité de la donnée est une condition de crédibilité des gains annoncés. Un usage IA ne peut pas être considéré comme robuste si les données d'entrée sont silotées, hétérogènes, mal documentées ou dégradées dans le temps. Ce point a une double conséquence : il conditionne la performance du modèle, mais aussi la validité de la preuve produite. Un gain mesuré sur des données mal qualifiées, non maintenues ou non représentatives peut difficilement fonder une décision d'industrialisation.

La qualité de la donnée doit donc être suivie comme une variable de gouvernance : fraîcheur, complétude, traçabilité, documentation, représentativité, dérive, responsabilité de mise à jour. Sans cette discipline, le passage du POC au déploiement risque de produire des systèmes techniquement fonctionnels mais fragiles, peu explicables et difficiles à maintenir.

La donnée n'est donc pas seulement le carburant du modèle ; elle est aussi la condition de validité de la preuve.

10.3 La mesure du gain net : le test de vérité

Mesurer le gain net produit par l'IA suppose d'abord de définir ce que l'organisation cherche réellement à gagner. L'évaluation repose ensuite sur trois opérations distinctes : définir un scénario de référence sans IA ou avec une autre solution d'IA, déduire le coût du calcul associé au système lui-même, puis mesurer le bénéfice effectivement produit¹¹³.

La nature de ce gain doit être explicitée. Il peut être économique, opérationnel, temporel, énergétique, environnemental, ou porter sur la sécurité, la qualité ou la réduction des risques. Tous ces gains ne sont ni substituables ni mesurables avec les mêmes indicateurs. La preuve doit donc commencer par préciser le bénéfice attendu et l'indicateur retenu pour l'évaluer : coût évité, réduction du temps de traitement ou du time-to-market, énergie économisée, émissions évitées, baisse du taux d'erreur, amélioration de la disponibilité ou réduction d'un risque opérationnel.

Le scénario de référence est souvent la première difficulté. Est-ce le comportement antérieur de l'opérateur ? Un contrôle périodique standard ? Un algorithme règle-métier simple ? Un autre système d'optimisation ? Sans réponse à cette question, le gain annoncé est relatif à une référence non spécifiée et donc non vérifiable.

¹¹² Voir chapitre 9, section 9.5, pour la source détaillée du projet AI4HP (introduction d'un algorithme IA pour piloter des pompes à chaleur) et les résultats expérimentaux.

¹¹³ Pour plus de détails, voir la norme ITU L.1801, publiée en février 2026 <https://www.itu.int/rec/T-REC-L.1801/fr>

Pour le cas AI4HP, la référence est le système de régulation existant de la pompe à chaleur. Le gain mesuré en laboratoire est de 40%. Le coût du calcul de l'algorithme est de 1% de la consommation de la pompe à chaleur en fonctionnement normal. Un point d'attention : cette solution doit encore passer à l'échelle pour confirmer ces gains.

Un gain net n'existe qu'au regard d'un périmètre défini. Celui-ci doit préciser au minimum le système observé, l'horizon temporel, le scénario de référence, les ressources et impacts pris en compte, les effets déplacés hors du périmètre initial et les principaux effets rebond. Un gain local peut disparaître, voire s'inverser, lorsque le périmètre d'analyse est élargi.

Par exemple, optimiser la climatisation d'un bâtiment pendant une canicule peut réduire sa consommation énergétique à niveau de confort constant. Ce gain doit toutefois être distingué d'un éventuel effet d'extension de l'usage, d'une augmentation du niveau de confort attendu ou d'un déplacement de la demande électrique vers une période de forte contrainte.

Quatre questions qui aident dans la mesure du gain net

1. **Quel est le scénario de référence ?** Il doit être défini, documenté et comparable.
2. **Quel est le coût du calcul de l'algorithme ?** L'empreinte du système IA lui-même doit être déduite du bénéfice annoncé.
3. **Le gain se maintient-il dans le temps ?** Un gain mesuré en phase de test peut s'éroder avec la dégradation de la donnée, l'intégration système ou l'évolution du contexte.
4. **Y a-t-il des effets rebond ?** Une optimisation qui réduit le coût unitaire peut augmenter les volumes d'usage, annulant le gain global.

Cette rigueur protège contre les bilans incomplets, dans lesquels le gain affiché masque le coût réel de l'IA mobilisée.

10.4 L'explicabilité : une condition de maîtrise, pas seulement d'ergonomie

Dans les systèmes énergétiques, l'exigence d'explicabilité est une condition de sécurité, de responsabilité opérationnelle et de réglementation. Un opérateur de réseau, un technicien de maintenance, un ingénieur de centrale ne peut pas valider une recommandation dont il ne comprend pas la logique, parce que c'est lui qui assume la responsabilité de la décision, pas l'algorithme.

Cette réalité produit un arbitrage industriel que les entretiens formulent clairement :

« Si le modèle de performance n'est pas explicable, c'est un problème dans l'industrie. On va préférer une performance un peu moindre avec un modèle explicable », entretien avec un industriel.

C'est une hiérarchie de valeurs cohérente dans un contexte où les conséquences d'une erreur peuvent être graves.

L'explicabilité n'est pas seulement une propriété du modèle ; c'est aussi une propriété du système. Un industriel illustre le cas d'un *chatbot* d'aide aux opérateurs : *« L'opérateur décrit une panne ; le système guide la procédure. Le système ne décide pas à la place de l'opérateur, il lui donne les éléments pour décider. La population vieillit, l'écart générationnel est réel, les systèmes se complexifient, l'IA n'est utile que si elle guide sans substituer. »*

L'exigence d'explicabilité a aussi une implication pour l'architecture : elle favorise les modèles interprétables (régression, arbres de décision, modèles linéaires, séries temporelles balisées) sur les

boîtes noires profondes et les systèmes où la validation humaine est intégrée dans le processus plutôt qu'ajoutée après coup. Ce choix d'architecture est aussi, souvent, un choix de frugalité : les modèles explicables sont généralement plus petits, plus rapides et moins consommateurs.

Cette exigence devient plus forte avec l'IA agentique. Dès qu'un système ne se contente plus de répondre, mais enchaîne des actions, appelle des outils, déclenche des opérations ou produit plusieurs inférences successives, la confiance ne peut plus reposer sur le modèle seul. Elle dépend de l'environnement qui l'encadre : journalisation, supervision, reproductibilité, contrôle des données, documentation, garde-fous, tests de robustesse et capacité de retour arrière.

10.5 L'acceptabilité des métiers cœur : la condition la plus sous-estimée

La prudence des métiers opérationnels face à l'IA est souvent une information qui dit où les systèmes IA ne sont pas encore prêts à être déployés, pas parce que les opérateurs résistent au changement, mais parce que les conditions de sécurité et de responsabilité ne sont pas encore remplies. Cette même prudence a déjà été illustrée en 8.2 : elle n'est pas propre à un métier ou à un secteur particulier, elle traverse l'ensemble des fonctions critiques de l'énergie.

Cette acceptabilité ne se décrète pas. Elle se construit : par la qualité de la preuve accumulée, par la transparence sur les échecs autant que sur les succès, par l'association des opérateurs à la conception du système plutôt qu'à sa seule utilisation, par des systèmes qui augmentent leur jugement plutôt qu'ils ne le remplacent.

Plus qu'une variable de communication, l'**acceptabilité** doit être vue comme une **variable de conception**.

Elle dépend aussi de la manière dont l'IA s'insère dans les processus existants. Une amélioration qui respecte les gestes, les responsabilités et les chaînes de validation déjà maîtrisés est généralement plus facile à adopter qu'un système qui impose simultanément une nouvelle technologie et une réorganisation du travail. Cette continuité peut constituer un puissant facteur d'acceptabilité, sans pour autant signifier que les processus existants doivent toujours être préservés : certains peuvent eux-mêmes devoir être transformés.

Ce que l'acceptabilité métier implique pour la conception

- **Associer les opérateurs dès la conception** : pour définir ce que le système doit et ne doit pas faire.
- **Construire des systèmes d'augmentation, pas de remplacement** : l'IA qui guide et informe est acceptée. L'IA qui décide seule est refusée et souvent à raison.
- **S'insérer autant que possible dans les processus maîtrisés** : limiter les ruptures simultanées entre technologie, organisation et responsabilités, sauf lorsque le processus lui-même doit être transformé.
- **Documenter les échecs autant que les succès** : la confiance se construit par transparence sur les limites, tout autant que par présentation des meilleurs cas.
- **Déployer progressivement, avec des phases de validation humaine explicites** : ne pas demander aux opérateurs de faire confiance à un système qu'ils n'ont pas pu éprouver.

Cette acceptabilité n'est toutefois qu'une première couche, encore faut-il que **l'organisation conserve les compétences pour juger ce que l'IA produit à court terme et dans la durée.**

10.6 Le verrou du jugement métier : l'effet sur les compétences critiques

Les entretiens montrent que l'IA remplace des tâches et, en même temps, déplace les compétences nécessaires pour juger le résultat. Dans les environnements énergétiques et industriels, cette question est centrale : générer un plan, une documentation, un code automate, une étude exploratoire, une procédure de maintenance ou une recommandation d'exploitation n'a de valeur que si un professionnel compétent peut en vérifier la cohérence, la robustesse et les conséquences. L'enjeu n'est donc pas seulement l'automatisation, mais la conservation du jugement métier.

Ce point crée une tension intergénérationnelle. Les profils expérimentés connaissent souvent le résultat attendu et peuvent détecter l'erreur. Les profils juniors risquent d'être formés dans un environnement où une partie du raisonnement est déjà préparée par les outils. Le risque est alors de **fragiliser les trajectoires d'apprentissage qui permettent de former, demain, les experts capables de superviser, contester ou reprendre la main sur les systèmes IA.**

La doctrine d'usage de l'IA dans les organisations énergétiques devrait donc intégrer un critère important : **l'effet sur les compétences critiques**. Une **IA utile** n'est pas seulement celle qui accélère une tâche ; c'est celle qui **accélère sans rendre l'organisation incapable de comprendre, maintenir, contester ou remplacer le système qui l'assiste.**

10.7 La robustesse dans le temps : ce que l'IA doit tenir

Comme tout système d'information intégré aux opérations, un système IA doit être maintenu dans le temps. L'IA ne crée pas cette exigence, mais elle y ajoute certaines formes particulières de dérive : évolution des comportements des usagers, changement du parc de capteurs, dégradation de la qualité des données, évolution du contexte énergétique ou dérive de la performance du modèle. Un système qui fonctionne bien à son déploiement peut ainsi perdre progressivement sa pertinence ou sa robustesse.

La robustesse suppose au moins quatre exigences : surveiller en continu la performance du modèle et les indicateurs de dérive ; réentraîner ou recalibrer périodiquement le système sur des données mises à jour ; prévoir des protocoles de reprise, de fonctionnement dégradé ou de mise en sécurité en cas de défaillance ; définir explicitement les conditions dans lesquelles le système ne doit pas être utilisé. Ce dernier point est particulièrement important dans les systèmes critiques : savoir quand ne pas utiliser l'IA est aussi important que savoir quand l'utiliser.

Les alertes de 2026 sur les systèmes agentiques dépassent désormais la seule question de leur coût. Les incidents et travaux documentés par OpenAI et Anthropic montrent que des agents capables d'agir sur des outils et des environnements réels peuvent aussi accomplir des actions non prévues ou contourner certains garde-fous.¹¹⁴ La maturité d'un système agentique ne peut donc plus être appréciée uniquement par ce qu'il accomplit, mais aussi par la capacité à borner ses permissions, observer ses trajectoires, interrompre son action et reprendre la maîtrise (voir 1.2).

¹¹⁴ <https://openai.com/fr-FR/index/hugging-face-incident-and-the-road-ahead/>

10.8 Frugalité par conception : mesurer le coût de l'IA pour crédibiliser le gain

Cette condition relie directement les enseignements de la Partie II à la présente partie : un usage IA frugal dans les systèmes énergétiques ne peut pas être crédible si son empreinte propre n'est pas mesurée et proportionnée au gain qu'il produit.

La frugalité par conception implique que le coût environnemental du calcul soit traité comme une variable de design dès la phase de conception et non comme une contrainte découverte après déploiement. Cela suppose de : choisir la taille du modèle en fonction de l'usage - SLM, modèle spécialisé ou LLM ; définir la fréquence d'inférence en fonction du système optimisé ; choisir la localisation du calcul - edge, cloud, cloud de confiance ou infrastructure interne - selon les contraintes de souveraineté, de latence, de disponibilité et d'intensité carbone. Cela suppose aussi de formuler explicitement le gain net attendu avant d'engager les ressources.

AI4HP illustre aussi une frugalité par conception : modèle borné, exécution locale, fréquence d'inférence adaptée au système physique optimisé et coût de calcul intégré dès l'évaluation du gain. L'enjeu n'est pas seulement d'utiliser un modèle léger, mais de concevoir l'architecture pour que le coût du calcul reste proportionné au bénéfice attendu.

Cette logique de proportionnement vaut aussi pour l'infrastructure. Un interviewé le formule ainsi : *« On équipe le calcul en fonction de la demande. On ne veut pas acheter des cartes qui seront obsolètes dans deux ans. On y va en fonction des demandes clients. »* La frugalité par conception ne concerne donc pas seulement le modèle ou l'algorithme. Elle concerne aussi l'investissement matériel, la durée de vie des équipements et la capacité à éviter le surdimensionnement.

La frugalité par conception devient ainsi une condition de crédibilité environnementale : il ne suffit pas qu'un usage IA vise un objectif énergétique, il faut démontrer que le modèle, la fréquence d'appel, l'infrastructure mobilisée, la donnée traitée et le mode d'exploitation sont proportionnés au bénéfice attendu.

10.9 Le maintien des collectifs humains : la condition la plus négligée

Condition rarement formulée, mais majeure à long terme : l'intégration de l'IA dans les systèmes énergétiques ne doit pas éroder les collectifs humains de compétence qui garantissent la maîtrise et la résilience de ces systèmes.

Les entretiens documentent une tension réelle. Les organisations énergétiques font face simultanément à deux mouvements contraires :

- d'un côté, l'intégration de l'IA dans des processus de plus en plus complexes, ce qui peut déplacer une partie du raisonnement, de la recommandation ou de la décision vers le système ;
- de l'autre, un défi démographique réel : vieillissement des équipes opérationnelles, départ en retraite d'experts, difficulté à recruter face à la concurrence du secteur numérique.

« La population vieillit, il y a un gap générationnel. On est performant après une dizaine d'années dans le métier, les systèmes se complexifient, énergies intermittentes, consommations intermittentes, ENR, dégâts... On a créé une interface qui permet de guider les gens », entretien avec une organisation énergétique. Ce témoignage converge avec celui de l'industriel cité en section 10.4, qui souligne lui aussi le vieillissement des équipes, l'écart générationnel, la complexification des systèmes et la nécessité d'une IA qui guide l'opérateur sans se substituer à lui.

L'IA peut contribuer à résoudre ce problème : capitaliser les expertises avant le départ à la retraite, permettre à un opérateur junior de bénéficier de l'expertise d'un senior, guider les personnes dans des situations complexes.

Mais l'IA peut aussi aggraver le problème si elle est mal conçue. Un système qui prend en charge une décision opérationnelle sans que les opérateurs comprennent son fonctionnement produit une « expertise orpheline » : la compétence humaine s'atrophie et si le système tombe en panne ou que les données se dégradent, les opérateurs ne peuvent plus reprendre la main efficacement.

Cette perte ne concerne pas seulement la continuité opérationnelle. Une compétence peut rester formellement présente dans les référentiels de compétences, les procédures ou les bases de connaissances tout en cessant progressivement d'être exercée et renouvelée. L'organisation conserve alors formellement une capacité qu'elle devient moins capable de mobiliser lorsqu'une situation nouvelle, un incident ou une rupture exige de reconstruire le raisonnement.

Quelques règles pour préserver les collectifs humains de compétence

1. **Une IA utile ne doit pas éroder les compétences humaines nécessaires pour comprendre, contester et reprendre la main.**
2. **La capacité de reprise ou de mise en sécurité est une condition de déploiement** : tout système IA intégré dans un processus critique doit prévoir et tester les modalités permettant à l'organisation de reprendre la maîtrise, de fonctionner en mode dégradé ou de mettre le système en sécurité en cas de défaillance.
3. **Les équipes ne doivent pas perdre la maîtrise du raisonnement** : former les opérateurs à comprendre les limites du système IA est aussi important que les former à l'utiliser.
4. **Préserver les compétences critiques ne signifie pas seulement maintenir les compétences existantes** : chaque automatisation significative doit identifier les compétences qu'elle risque d'atrophier, celles qu'elle doit maintenir en exercice et celles qu'elle oblige à construire pour conserver demain une capacité de compréhension, de reprise et d'innovation.

10.10 Synthèse : sept conditions, un système

Pour clore ce chapitre, résumons les sept conditions de crédibilité opérationnelle. Elles sont interdépendantes et forment une base utile pour instruire les usages.

Condition	Pourquoi elle est systémique
1. Intégration dans les systèmes réels	Sans intégration réelle, les autres conditions ne peuvent pas être évaluées en situation opérationnelle
2. Qualité et maintien de la donnée	Sans donnée fiable et maintenue dans le temps, ni la mesure du gain, ni la robustesse, ni l'explicabilité ne tiennent durablement.
3. Mesure du gain net	Elle distingue une IA utile d'une IA coûteuse sans bénéfice mesurable.
4. Explicabilité, supervision et acceptabilité métier	Sans compréhension ni validation par les métiers, le système reste externe aux décisions réelles.
5. Robustesse dans le temps	Un gain initial qui s'érode avec les données, les capteurs ou les usages n'est pas un bénéfice opérationnel durable.
6. Frugalité par conception	Sans elle, une IA déployée au nom de l'énergie peut consommer plus qu'elle ne permet d'économiser.
7. Maintien et renouvellement du jugement métier et des collectifs humains	Sans compétences humaines capables de comprendre, contester, reprendre la main et développer les capacités nécessaires aux évolutions futures, la résilience du système se fragilise.

Ces conditions permettent d'évaluer un usage IA existant, d'instruire un projet en développement et de distinguer ce qui justifie l'infrastructure mobilisée de ce qui ne la justifie pas. Elles évitent que le bilan de l'IA pour les systèmes énergétiques soit réduit à une liste de cas d'usage sans preuve de crédibilité.

Reste alors la question : quels angles morts faut-il traiter pour que l'avantage énergétique français ne devienne pas un passif ?

PARTIE IV

Choisir avant de subir : les fondations d'un raisonnement durable

Chapitre 11 - Les angles morts qui peuvent transformer l'avantage français en passif

Le chapitre en une phrase

Sept angles morts restent insuffisamment traités : gouvernance des données hébergées, cycle de vie et fin d'exploitation, effet foncier, mesure de l'empreinte réelle, souveraineté par couches, expertise publique indépendante, absence d'apprentissage collectif. Sans leur traitement, la doctrine reste vulnérable à des arbitrages conduits dans l'opacité.

Ce chapitre isole ce qui reste insuffisamment traité : les questions encore invisibles, incomparables ou non opposables, qui devront être instruites avant que l'avantage français ne devienne un passif.

11.1 Premier angle mort : la gouvernance des données hébergées

Les procédures d'autorisation actuelles - permis de construire, raccordement électrique, instruction environnementale - ne permettent pas, en tant que telles, d'instruire de manière systématique ce qui tournera réellement dans un data center : quelles données seront hébergées ? Sous quelle juridiction ? Pour quels usages ? Par quels clients ? Avec quelles garanties pour la puissance publique ?

« Des data centers peuvent être construits par des promoteurs qui ne savent pas encore à qui ils les vendront. Des porteurs de projets déposent des demandes de raccordement sans que l'usage final soit stabilisé. Et une fois le data center en service, l'opérateur d'un site de colocation peut ne pas avoir accès aux informations sur ce que ses clients font tourner dans ses baies », explique un expert.

L'opacité est triple :

1. **Elle est technique** : on ne sait pas quels modèles IA, quelles charges, quelles consommations réelles.
2. **Elle est contractuelle** : le secret des affaires est invoqué à plusieurs niveaux de la chaîne.
3. **Elle est juridique** : la qualification des données hébergées, leur juridiction, leur nature ne sont pas déclarées dans les procédures d'autorisation.

« Business France nous oppose le secret des affaires. On demande plus de transparence dans les opérations data centers, des investissements troubles, quels clients, quelles données... on nous dit que c'est stratégique alors que ça concerne les collectivités. C'est la souveraineté française et européenne. », entretien avec un acteur territorial.

Ce vide crée un risque identifié dans les entretiens : que des acteurs dont les intérêts sont contraires aux intérêts français ou européens se retrouvent propriétaires ou locataires d'infrastructures critiques sur le sol français, sans que les autorités n'en aient eu connaissance au moment de l'autorisation. Le risque est celui de la dépendance opérationnelle : une infrastructure critique dont les clés sont à l'étranger est une infrastructure vulnérable.

NIS2 et SecNumCloud¹¹⁵ constituent des briques importantes de cybersécurité et de confiance pour certains usages, prestataires et environnements critiques. Mais elles ne répondent pas au même

¹¹⁵ NIS2 : directive européenne renforçant les obligations de cybersécurité et de gestion des risques pour les entités essentielles et importantes. SecNumCloud : référentiel de qualification de l'ANSSI applicable à certaines offres de cloud présentant un niveau élevé d'exigences de sécurité.

problème que l’instruction territoriale d’un data center. Elles permettent de qualifier des exigences de sécurité, de gouvernance, d’hébergement ou d’exploitation ; elles ne constituent pas, à elles seules, un mécanisme permettant de connaître en amont les investisseurs finaux, les clients, les usages, les charges de calcul et la nature des données hébergées.

Ce point est essentiel : la confiance dans un service cloud ne résout pas la transparence d’un projet d’implantation. NIS2 et SecNumCloud protègent des usages qualifiés ; elles ne permettent pas aux collectivités ni aux services instructeurs de savoir ce qui sera hébergé avant d’autoriser la construction. L’angle mort n’est donc pas seulement cyber. Il est aussi territorial, économique et souverain.

Les espaces de données : dissocier l’hébergement de la gouvernance

Les espaces de données, ou *data spaces*, ouvrent une autre voie de maîtrise. Ils permettent de dissocier davantage l’hébergement technique de la gouvernance de la donnée : les participants définissent les droits d’accès, les conditions d’usage, les règles d’interopérabilité et les modalités de partage de la valeur. Un *hyperscaler* peut alors rester fournisseur d’infrastructure sans disposer seul du pouvoir d’organiser l’accès aux données et d’en capter les usages.

Cette architecture ne supprime pas automatiquement les dépendances au cloud, aux logiciels, aux services d’identité ou aux outils d’intermédiation. Elle peut néanmoins déplacer une partie du contrôle vers une gouvernance sectorielle, territoriale ou européenne. La souveraineté de la donnée ne dépend donc pas seulement du lieu où elle est hébergée, mais aussi de qui définit les droits d’accès, autorise les réutilisations, contrôle les flux et organise la répartition de la valeur.

Ce que ce premier angle mort implique

Conditionner les autorisations : exiger un niveau minimal de transparence sur les investisseurs finaux, les clients pressentis lorsqu’ils sont connus, les catégories d’usages envisagées, les types de charges de calcul et les garanties de mise à jour.

Créer un registre national vérifiable : mettre en place un registre accessible selon des niveaux différenciés aux services de l’État, aux autorités compétentes et aux collectivités concernées, avec une version publique limitée aux informations non sensibles.

Définir ce qui relève du secret des affaires : toutes les informations sur les investisseurs et les usages ne peuvent pas être couvertes par le secret des affaires. Une définition légale du périmètre est nécessaire.

Soutenir les espaces de données sectoriels : dissocier l’hébergement de la gouvernance, organiser les droits d’accès et de réutilisation et rendre explicites les règles de partage de la valeur.

11.2 Deuxième angle mort : le cycle de vie et la fin d’exploitation

Le deuxième angle mort est moins visible et peut être très coûteux à long terme. **Que devient un data center devenu obsolète ?** Qui porte les coûts de démantèlement, de dépollution du site, de libération du raccordement électrique, de reconversion du bâtiment ? Que deviennent les GPU, les serveurs, les équipements de refroidissement ?

Les entretiens documentent une réalité déjà présente : des data centers qui ont 15 à 20 ans, qui n’ont pas été rénovés, dont les propriétaires ne savent pas comment les moderniser. Le rythme

d'obsolescence est incompatible avec les horizons de l'investissement immobilier : les cycles technologiques des serveurs et des GPU sont de 3 à 5 ans, mais les bâtiments, les raccordements et les travaux de génie civil s'inscrivent dans des horizons de 20 à 30 ans.

« Le risque d'obsolescence des data centers est avéré. On dépasse la loi de Moore et on multiplie par 10 la puissance chaque année. », entretien avec un architecte.

Ce risque d'obsolescence est aggravé par l'incertitude technologique propre à l'IA : évolution rapide des GPU, architectures spécialisées, modes de refroidissement, densités électriques, modèles plus petits ou plus spécialisés et possibles ruptures de calcul à plus long terme. L'informatique quantique est parfois évoquée dans les entretiens comme une rupture possible, mais elle doit rester traitée comme une hypothèse prospective, non comme un scénario établi. **Construire un data center de grande taille revient donc à immobiliser des actifs longs dans un environnement technologique court, mouvant et difficilement prévisible.**

La dissociation entre horizons est structurelle : le territoire hérite des infrastructures longues (bâtiments, raccordements, voirie, foncier immobilisé) ; l'opérateur renouvelle ses équipements IT à son rythme propre, dicté par la compétition technologique. Si le data center est vendu ou abandonné, c'est le territoire qui gère les externalités, le foncier bloqué, le raccordement inutilisé, le bâtiment parfois impropre à d'autres usages.

À ce stade, il n'existe pas de cadre systématique propre aux data centers imposant un plan de réversibilité, une provision financière de fin d'exploitation et une libération organisée du raccordement. Les obligations existantes peuvent traiter certains aspects, mais elles ne constituent pas une doctrine complète de fin de vie. Les autorisations d'urbanisme ne comportent pas de clause de réversibilité ou de fin de vie. Les opérateurs ne sont pas tenus de financer une provision pour démantèlement. Et les collectivités ne disposent pas d'outils pour négocier des garanties sur ce point lors de l'instruction.

Ce que ce deuxième angle mort implique

Un plan de réversibilité obligatoire : intégrer dans les cahiers des charges d'autorisation un plan de réversibilité et de fin d'exploitation : conditions de démantèlement, de dépollution, de libération du raccordement et de reconversion du site.

Une provision financière de fin d'exploitation : s'inspirer des logiques déjà appliquées à certaines infrastructures industrielles pour couvrir le démantèlement, la dépollution éventuelle, la libération du raccordement et la reconversion du site, afin que le coût ne repose pas sur les collectivités.

La priorité au recyclage urbain instruit : privilégier les sites industriels existants (brownfield) où l'obsolescence du data center peut ouvrir sur une reconversion déjà pensée, plutôt que des constructions neuves sur site vierge.

11.3 Troisième angle mort : l'impact sur les prix du foncier et l'effet d'éviction

Un troisième angle mort concerne les effets locaux de la concentration des data centers sur les marchés fonciers. Le phénomène est mentionné dans certains entretiens : des terrains simplement raccordables ou proches de capacités électriques disponibles peuvent atteindre des valorisations considérables, créant une concurrence asymétrique entre porteurs de projets industriels, acteurs

immobiliers et stratégies spéculatives. Aucune source ne quantifie l'effet d'éviction sur le logement, l'artisanat, les activités productives ou les équipements publics dans les zones de concentration françaises.

Ce vide analytique est problématique parce que l'effet d'éviction existe et est documenté hors de France. Le Guide France Urbaine / Ville de Demain (avril 2026) cite le cas de Franklin Farms en Louisiane comme exemple d'effet d'éviction foncière lié à un projet de data center : un site capté par Meta pour un data center aurait écarté un projet d'usine automobile qui aurait généré des emplois qualifiés. Est également partagé le cas Bissen au Luxembourg qui illustre l'effet d'éviction temporelle : un projet qui ne voit pas le jour 8 ans après, mais qui bloque les alternatives.

La comparaison avec les entrepôts logistiques est récurrente dans les entretiens. Des territoires ont dit oui à des projets logistiques, ont attendu des emplois et ont réalisé que les emplois étaient peu nombreux, les nuisances importantes et la fiscalité décevante. Le risque d'une répétition de ce schéma avec les data centers est explicitement mentionné dans les entretiens.

« Autre sujet : le foncier est rare dans les métropoles. La proximité des zones denses (localisation, fibre, énergie). Il y a un sujet de concurrence entre territoires. C'est important pour nous d'outiller pour encadrer les data centers. », entretien avec un acteur territorial.

L'angle mort spécifique est ici **l'absence de méthodologie reconnue pour mesurer le coût d'opportunité foncier lors de l'instruction des projets**. Une collectivité qui accepte un data center sur un terrain de 10 hectares en zone périurbaine ne dispose pas d'outils standard pour évaluer : qu'aurait-on pu y construire d'autre ? Combien d'emplois directs de plus ? Quel impact fiscal net ? Quelle densité d'usages ? Cette absence de méthodologie avantage les porteurs de projets qui peuvent formuler leurs promesses sans outil de contre-expertise.

La Métropole d'Aix-Marseille-Provence a développé un simulateur fiscal pour objectiver les retombées selon différentes hypothèses de qualification. C'est une réponse partielle, mais elle montre que la demande existe et que les outils peuvent être construits. Le problème est leur rareté : très peu de collectivités disposent de tels outils et aucun standard national ne permet de les comparer.

Ce point rejoint l'analyse fiscale développée en 3.5 : l'enjeu n'est pas seulement de savoir ce qu'un data center rapporte, mais de comparer ce rendement fiscal à ce que le même foncier, la même puissance électrique et les mêmes infrastructures auraient permis de produire autrement. La fiscalité doit donc être intégrée à une lecture de coût d'opportunité territorial et non isolée comme promesse de retombées.

11.4 Quatrième angle mort : la mesure de l'empreinte réelle des data centers

Le quatrième angle mort concerne la mesure. Il s'agit de l'empreinte des data centers eux-mêmes, dans leur réalité physique : électricité consommée, eau utilisée, chaleur produite/captée, fluides frigorigènes, groupes électrogènes, durée de vie du matériel, empreinte de la construction.

La mesure environnementale ne peut pas se limiter aux indicateurs annoncés en phase projet. Les entretiens signalent un écart possible entre performance déclarée, performance de conception et performance réelle en exploitation. La doctrine de reporting devrait donc distinguer les indicateurs de conception, les indicateurs contractuels et les indicateurs mesurés en exploitation. Elle devrait aussi intégrer les angles morts aujourd'hui peu comparables : WUE, fuites de fluides frigorigènes,

tests des groupes électrogènes, chaleur effectivement valorisée, bruit, consommation d'eau réelle et impacts cumulés à l'échelle d'un bassin territorial.

Cette mesure n'est pas disponible de manière standardisée, vérifiable et comparable. Les opérateurs publient des chiffres -PUE, part d'énergie renouvelable, WUE- mais avec des périmètres et des méthodes différents qui rendent les comparaisons impossibles. Les collectivités qui instruisent des projets ne disposent pas de bases de données permettant de comparer les engagements annoncés avec les performances réelles d'autres sites du même opérateur ou de sites comparables.

La standardisation de la mesure est nécessaire, mais elle ne doit pas devenir un nouveau verrou de marché. Si les exigences de *reporting* environnemental ne sont accessibles qu'aux grands fournisseurs capables de produire des métriques propriétaires complexes, elles peuvent renforcer les acteurs dominants au lieu d'ouvrir le marché. La doctrine doit donc viser des métriques comparables, auditable, proportionnées et interopérables, utilisables aussi par les acteurs européens, les opérateurs de taille intermédiaire et les acheteurs publics. Une mesure robuste doit être praticable par les acteurs de toutes tailles, pas seulement par les *hyperscalers* qui disposent des ressources pour influencer les cadres méthodologiques et produire des *reportings* complexes.

« On manque d'expertises locales neutres et indépendantes. On a une obsession sur l'eau et la chaleur fatale. Il faut une transversale. Il y aura peu de réalisations si on n'y pense pas. On manque d'expertises, on manque de vision systémique, on manque de vision politique. », entretien avec un expert de l'énergie.

Les angles morts spécifiques de la mesure sont connus. Par exemple, le WUE est rarement mesuré ou rarement publié avec un niveau de détail comparable par les acteurs et n'est pas intégré dans les critères d'évaluation des autorisations. Les fuites de fluides frigorigènes¹¹⁶ (HFC à très fort potentiel de réchauffement global) sont une source d'émissions directes souvent négligée.

La question de la standardisation est aussi une question concurrentielle, que l'Autorité de la concurrence¹¹⁷ souligne explicitement.

Ce que ce quatrième angle mort implique

Métriques sectorielles adaptées : PUE, WUE, bilan acoustique, chaleur fatale valorisée ou non : des métriques qui parlent à ce secteur, pas la transposition inadaptée du décret tertiaire (kWh/m²/an).

Obligations de déclaration : consommations réelles d'eau et d'électricité, PUE et WUE réellement observés, fuites de fluides frigorigènes, tests des groupes électrogènes : des bilans annuels vérifiables, au-delà des engagements initiaux.

Accès public et vérifiabilité : les données environnementales essentielles des data centers autorisés doivent être accessibles aux collectivités et aux citoyens sous une forme compréhensible et non sensible, avec un accès plus détaillé pour les services compétents.

¹¹⁶ <https://www.ville-demain.com/guide-du-data-center-durable-et-acceptable-partie-2-limpact-environnemental/>

¹¹⁷ Source : Autorité de la concurrence, *Étude sur les enjeux concurrentiels liés à l'impact énergétique et environnemental de l'IA*, 17 décembre 2025, section II.B, « Les enjeux liés à l'émergence de la frugalité comme paramètre de concurrence », §92 et §100-103. https://www.autoritedelaconcurrence.fr/sites/default/files/2026-01/Etude_IA_Energie_ADLC-LR.pdf

La mesure environnementale contribue à la capacité publique à comparer, négocier, autoriser, contrôler et, le cas échéant, refuser. Le même raisonnement vaut pour la souveraineté : sans qualification des couches critiques, la doctrine reste prisonnière du seul critère de localisation.

11.5 Cinquième angle mort : la souveraineté par couches ou ce qu'il devient critique de ne pas maîtriser

Le cinquième angle mort est le plus stratégique. Le chapitre 4 a montré que localisation, souveraineté et valeur captée ne coïncident pas. Il propose une grille d'instruction des dépendances d'un data center IA fondée sur trois questions : maîtrise, vulnérabilité et substituabilité. Ces trois questions qualifient un profil de dépendance ; elles ne déterminent pas, à elles seules, sa criticité. Celle-ci dépend également de l'usage concerné, des conséquences de la perte et de la capacité à maintenir, substituer ou reconstruire la fonction dans un délai acceptable.

Les annonces franco-allemandes faites à VivaTech en juin 2026 donnent une boussole politique utile. La France et l'Allemagne ont publié une position commune sur la souveraineté numérique européenne et annoncé la relance d'un Forum franco-allemand pour l'avenir, avec deux premières réalisations attendues : un catalogue commun de solutions numériques souveraines et un cadre d'évaluation des dépendances numériques critiques de l'Europe. Cette orientation est importante : elle reconnaît que la souveraineté numérique ne se réduit pas à la localisation des données, mais suppose d'identifier les dépendances, les capacités industrielles, les conditions d'interopérabilité, les infrastructures critiques et les marges de substitution.

Les annonces intervenues durant l'été 2026 confirment que cette instruction ne peut pas se limiter aux nationalités ou aux lieux d'implantation. Un acteur français peut développer une capacité européenne tout en s'appuyant sur une plateforme américaine pour sa distribution. Un opérateur historique français peut partager la gouvernance de ses infrastructures avec un investisseur mondial. Une capacité européenne de calcul peut rester dépendante de puces extra-européennes. La maîtrise d'une couche peut donc progresser tandis qu'une autre dépendance se renforce.

Une politique de souveraineté ne doit ni rejeter automatiquement ces configurations hybrides, ni leur attribuer globalement le caractère de souveraineté. Elle doit préciser ce qu'elles rendent possible, ce qu'elles placent sous contrôle européen, les dépendances qu'elles maintiennent et les conditions dans lesquelles elles pourraient être substituées.

Nous proposons d'appliquer cette boussole à un test de réalité opérationnel : couche par couche, que maîtrise-t-on vraiment ? Que peut-on auditer ? Que peut-on substituer ? Que devient-il critique de ne pas maîtriser ? La souveraineté ne peut donc pas être traitée comme une propriété générale d'un acteur ou d'un lieu. Elle doit être **instruite comme une chaîne de dépendances systémiques : modèle, données, calcul, cloud, puces, câbles, logiciels, plateformes de distribution, poste de travail, interfaces, intégration, accès aux clients, exploitation, support, mises à jour, contrats et compétences.**

Une souveraineté relative à un usage, un risque et un périmètre

Une solution n'est pas souveraine dans l'absolu. Elle peut être suffisamment maîtrisée pour un usage donné, dans un périmètre donné et face à un scénario de dépendance explicite. La question « cette IA est-elle souveraine ? » doit donc être remplacée par trois questions : **souveraine pour quel usage, face à quel risque et à quelle échelle ?**

L'échelle pertinente fait elle-même partie de l'instruction. Une capacité¹¹⁸ ne doit pas être recherchée systématiquement au niveau national si les ressources nécessaires, les chaînes industrielles ou les interdépendances dépassent cette échelle. L'échelon européen devient pertinent lorsqu'il permet de construire une capacité de décision, de négociation ou de substitution que les États membres ne pourraient assurer efficacement seuls. La souveraineté consiste donc aussi à déterminer à quelle échelle une dépendance peut réellement être maîtrisée.

L'échelle doit aussi être instruite dans l'autre sens. Une capacité considérée comme stratégique au niveau national ou européen peut mobiliser localement du foncier, de l'électricité, de l'eau, des infrastructures et des choix d'aménagement pendant plusieurs décennies. Le renforcement d'une capacité souveraine à une échelle ne doit pas conduire à une perte non instruite de capacité de décision à une autre. La souveraineté par couches doit aussi permettre de vérifier qui conserve la capacité de comprendre, négocier, conditionner, réexaminer ou infléchir la trajectoire du projet. La souveraineté ne se juge donc pas seulement par les dépendances que l'on réduit ou maîtrise, mais aussi par celles que l'on crée et par les capacités de décision que l'on préserve à chaque échelle.

Pour les **activités régaliennes**, la défense, certaines infrastructures critiques ou les données publiques les plus sensibles, la maîtrise peut exiger une capacité de fonctionnement déconnecté, une administration sous contrôle européen, une auditabilité complète, une maîtrise des mises à jour et une possibilité réelle de substitution.

Pour les **administrations**, les hôpitaux, les collectivités, la recherche publique ou certains services essentiels, la commande publique peut soutenir des acteurs européens et organiser leur apprentissage. Cette préférence ne doit toutefois pas reposer sur la seule nationalité du fournisseur. Elle doit porter sur des garanties vérifiables : données, modèles, infrastructure d'exécution, administration, interfaces, portabilité et conditions de sortie.

Pour une **entreprise opérant dans plusieurs régions du monde**, l'objectif ne sera pas toujours d'imposer un fournisseur français ou européen à l'ensemble de ses filiales. Ses besoins incluent la couverture internationale, l'intégration aux outils mondiaux, le support local et la conformité à plusieurs juridictions. Sa stratégie de souveraineté peut alors privilégier la maîtrise des données et des clés, la portabilité, l'interopérabilité, le choix du lieu d'exécution, une architecture multi-modèles et la possibilité de changer de plateforme.

Cette différenciation évite de faire de la souveraineté une exigence déclarative uniforme, impossible à appliquer, et permet de concentrer l'effort de maîtrise sur les usages et les couches dont la perte aurait les conséquences les plus graves.

Le tableau suivant ne reproduit pas les huit couches de la grille d'instruction présentée au chapitre 4. Il les prolonge à l'échelle d'une stratégie française et européenne, en identifiant les capacités à construire ou à renforcer.

¹¹⁸ Cette formulation s'inspire de la réflexion de Jean Tirole et Karine Van der Straeten, dans *Économie du désordre mondial*, sur la bonne échelle d'action en matière de souveraineté.

Couche	Situation actuelle	Exemple de ce qui manque
Modèles de fondation	Mistral figure parmi les acteurs européens les plus visibles sur les modèles de fondation ; d'autres acteurs européens, plus spécialisés ou sectoriels, existent mais ne disposent pas du même effet d'échelle que les grands fournisseurs américains. La dépendance persiste sur les capacités de calcul, les données, les capitaux, les infrastructures d'entraînement et les canaux de distribution.	Commande publique orientée vers des solutions européennes lorsque c'est pertinent ; accès garanti à du calcul compétitif ; capacité de test, d'évaluation et de déploiement dans les administrations et secteurs critiques ; stratégie de distribution et d'intégration dans les postes de travail ; culture des organisations pour s'affranchir des outils américains.
Puces IA	TSMC domine la fabrication avancée des puces IA ; ASML, acteur européen, est le fournisseur critique des machines EUV.	Stratégie européenne des semi-conducteurs en cours, notamment dans le cadre du Chips Act, mais encore insuffisante pour sécuriser la chaîne complète des puces IA avancées. Développer une stratégie différenciée selon les briques : accélérateurs IA, processeurs HPC, microcontrôleurs, architectures ouvertes, packaging avancé et composants embarqués. L'Europe peut notamment s'appuyer sur SiPearl pour les processeurs HPC et sur STMicroelectronics pour les microcontrôleurs, l'IA embarquée, les capteurs, la connectivité et l'électronique de puissance. Les architectures ouvertes telles que RISC-V peuvent réduire certaines dépendances de propriété intellectuelle, sans remplacer les capacités de fabrication, les outils de conception, les mémoires et les accélérateurs nécessaires à l'IA avancée.
Infrastructures cloud	OVHcloud, Scaleway, Outscale, Orange Cloud Avenue, Bleu, NumSpot et d'autres offres européennes ou françaises structurent une alternative partielle, mais restent confrontés à la profondeur de services, à l'écosystème logiciel, aux habitudes d'achat et à la puissance commerciale des <i>hyperscalers</i> américains.	Commande publique structurante, trajectoire de migration réaliste, interopérabilité, réversibilité, financement de la montée en capacité, doctrine claire entre cloud de confiance, cloud souverain, cloud critique et cloud substituable.
Câbles sous-marins	Place croissante des grandes plateformes numériques dans le financement, la propriété ou l'usage capacitaire des câbles sous-marins. Incidents en mer Baltique 2023-2024.	Politique de diversification et protection des routes de câbles
Logiciels opérationnels	Dépendances fortes à des éditeurs et fournisseurs extra-européens dans de nombreuses organisations publiques, comme privées. Des incidents récents en France ont illustré concrètement cette vulnérabilité : « en décembre 2024, PSL a	Suite DINUM en déploiement incomplet. Vision et choix open source pertinents et intégrés

Couche	Situation actuelle	Exemple de ce qui manque
	<i>basculé vers Microsoft 365 ; en janvier 2025, Polytechnique a rencontré des blocages lors d'une opération similaire. Ces cas signalent que la dépendance logicielle n'est pas seulement un risque théorique de souveraineté, mais un problème opérationnel déjà vécu dans des établissements publics d'enseignement supérieur français», un chercheur.</i>	
Interfaces de travail	Suites bureautiques, ERP, CRM, outils métiers et copilotes deviennent les portes d'entrée silencieuses de l'IA	Capacité à auditer, désactiver, substituer ou router les fonctions IA intégrées
Intégration IA, énergie et industrie	Les compétences sont dispersées entre énergie, numérique, data, achats, urbanisme et environnement	Formation d'intégrateurs capables d'arbitrer usage, infrastructure, réseau, impact et souveraineté
Expertise indépendante	Expertise data centers et IA encore dispersée, avec une forte dépendance à des référentiels et cabinets anglo-saxons.	Pôle national d'expertise publique indépendant

La souveraineté par couches ne doit pas s'arrêter aux briques visibles : modèle, cloud, puces, câbles. Elle doit inclure les interfaces de travail et les compétences d'intégration. Les interfaces déterminent les usages réels et les intégrateurs la capacité à faire dialoguer énergie, IA, données, métiers, réseau, sécurité et impacts. Sans ces deux couches, la souveraineté reste une architecture dépourvue de capacité d'exécution.

« *Le balancier est parti loin vers le cloud. Il est en train de revenir rapidement* », confie un industriel. Ce retour se fait parce que la dépendance devient concrètement problématique pour des organisations qui ont besoin de garanties sur la disponibilité, la confidentialité et la maîtrise de leur système d'information.

La disponibilité déclarée d'une technologie ne garantit pas non plus un accès maîtrisé. Le pouvoir peut aussi s'exercer sur les conditions de circulation : calendrier d'accès, prix, niveaux de service, fonctionnalités disponibles, ordre de priorité des clients, maintenance, mises à jour ou possibilité de suspendre un service. Une capacité peut rester officiellement accessible tout en étant fournie plus tard, plus cher ou dans une version dégradée. La grille de souveraineté doit donc évaluer non seulement l'existence d'une alternative, mais les conditions concrètes dans lesquelles elle peut être mobilisée. Le contrôle des flux peut également consister à les ralentir, à les renchérir ou à en hiérarchiser l'accès.

La souveraineté doit enfin être réévaluée dans le temps. Un montage acceptable au moment de sa création peut changer de nature si le capital évolue, si un canal de distribution devient incontournable, si les conditions tarifaires ou contractuelles se modifient, si un modèle devient indisponible ou si une infrastructure mutualisée devient captive d'un client dominant. **La souveraineté n'est donc pas un label attribué une fois pour toutes, mais une capacité à maintenir et à réexaminer périodiquement.**

De la maîtrise présente à la capacité dynamique de souveraineté

La grille maîtrise-vulnérabilité-substituabilité donne une photographie structurée des dépendances à un moment donné. Mais la souveraineté ne peut être réduite à cette photographie. Une organisation peut aujourd'hui disposer d'un niveau élevé de maîtrise tout en étant engagée dans une trajectoire qui érode ses compétences, réduit ses alternatives ou augmente progressivement son coût de sortie. À l'inverse, elle peut encore dépendre fortement d'un fournisseur tout en développant les compétences, les architectures et les alternatives qui augmenteront demain sa capacité de choix.

La souveraineté repose donc aussi sur une capacité dynamique : la faculté de maintenir, renouveler, recomposer et reconstruire les capacités qui permettent à l'organisation de **continuer à choisir et à agir lorsque son environnement évolue**.

L'entreprise « déchirée » face à l'IA : accélérer sans se déposséder

La sociodynamique développée par Jean-Christian Fauvet¹¹⁹ à partir des années 1970 ne réduit pas les acteurs à une opposition binaire entre « pour » et « contre ». Elle analyse leur position à partir de deux dimensions qui peuvent coexister : leur **synergie** avec un projet (l'énergie qu'ils mobilisent pour le faire réussir) et leur **antagonisme** (l'énergie qu'ils mobilisent pour le contester ou s'y opposer).

Dans cette cartographie apparaît notamment la figure du « **déchiré** » : un acteur qui présente simultanément une forte implication en faveur du projet et un antagonisme important. Il n'est donc ni simplement opposant, ni hésitant : il porte simultanément des raisons fortes d'avancer et de résister.

Nous proposons ici une transposition du comportement individuel ou collectif à **l'organisation elle-même face à l'IA**. Cette transposition n'appartient pas, à notre connaissance, au cadre initial de Fauvet. Il s'agit d'une proposition analytique de ce rapport. L'entreprise peut en effet avoir simultanément beaucoup à gagner et beaucoup à perdre dans l'automatisation. Elle doit exploiter l'IA pour gagner en performance, apprendre et innover ; mais elle sait également que cette transformation peut déplacer hors d'elle une partie de son jugement, de ses compétences, de son patrimoine ou de ses capacités d'innovation.

La question stratégique n'est donc pas de choisir entre adoption et protection, mais de savoir **comment transformer ce déchirement en capacité d'arbitrage** : entrer dans la transformation tout en conservant la possibilité de comprendre ses effets, de modifier ses choix et de corriger sa trajectoire.

Quatre capacités souveraines à préserver dans la transformation

Ces capacités ne constituent pas de nouvelles couches de souveraineté. Elles décrivent ce que l'organisation doit rester capable de faire à travers les dépendances identifiées par les couches et qualifiées par MVS.

La souveraineté d'action désigne la possibilité de continuer à agir lorsque les conditions extérieures changent : retrait d'un fournisseur, modification d'un modèle ou d'une API, rupture d'accès, hausse de prix, panne d'infrastructure. Elle mobilise la continuité, la substituabilité, la sortie et la capacité de reconstruction déjà présentes dans la grille du rapport.

¹¹⁹ Jean-Christian Fauvet, *L'Élan sociodynamique*, Éditions d'Organisation, 2004

La souveraineté de décision désigne la faculté de comprendre les choix qui structurent l'action, de les contester, de les arbitrer et de modifier les règles qui les produisent. Une organisation peut conserver juridiquement la décision finale tout en perdant progressivement les compétences nécessaires pour reconstruire ou contester le raisonnement auquel elle se fie. Le mécanisme d'« expertise orpheline » décrit au chapitre 10 fournit un exemple concret de cette dissociation.

La souveraineté patrimoniale porte sur les actifs distinctifs de l'organisation : données, connaissances, méthodes, raisonnements, règles métier, modèles, ontologies, expériences accumulées, savoir-faire et relations entre ces éléments. L'enjeu n'est pas seulement de les posséder, mais de maîtriser les conditions dans lesquelles ils sont codifiés, exploités, transférés et transformés par les systèmes d'IA.

La souveraineté d'innovation désigne enfin la faculté de produire les compétences, méthodes et alternatives qui n'existent pas encore. Une organisation peut préserver ce qu'elle maîtrise aujourd'hui tout en réduisant progressivement sa capacité à explorer de nouvelles trajectoires. Cette dimension dépasse donc la conservation du patrimoine : elle interroge les conditions de son renouvellement.

La souveraineté consiste donc aussi à préserver quatre facultés : agir, décider, protéger et renouveler son patrimoine distinctif, et continuer à innover.

Un capital de souveraineté qui peut s'accumuler ou s'éroder

Nous proposons d'appeler **capital de souveraineté** l'ensemble des ressources et aptitudes mobilisables qui permettent à une organisation ou à un collectif de conserver ses souverainetés d'action, de décision, patrimoniale et d'innovation. Ce capital comprend notamment des compétences réellement exercées, des alternatives éprouvées, une capacité de reprise et de reconstruction, des architectures réversibles, un pouvoir de négociation, des actifs critiques maîtrisés et des capacités d'apprentissage. Il ne constitue pas un score global de souveraineté : ces composantes doivent rester visibles séparément.

Le **capital de souveraineté** décrit ce que l'organisation ou le collectif a construit et peut mobiliser ; sa **capacité dynamique de souveraineté** désigne sa faculté de renouveler, recomposer et accroître ce capital.

À l'échelle d'un pays, cette dynamique peut consister à mobiliser des capacités industrielles déjà acquises pour construire les couches suivantes : la stratégie sud-coréenne illustre cette logique en s'appuyant sur ses positions dans les semi-conducteurs et les mémoires pour développer simultanément des accélérateurs IA, des modèles, du calcul, des data centers, de la robotique et des compétences (voir 4.6).

Une organisation peut disposer d'un capital important et pourtant être engagée dans sa dégradation : conserver des experts sans en former de nouveaux ; disposer d'une procédure de reprise sans jamais l'éprouver ; posséder une architecture substituable sans tester d'alternative ; capitaliser une connaissance métier sans entretenir les mécanismes qui permettent de la renouveler, ni d'explorer de nouvelles trajectoires. **Le capital s'enkyste lorsqu'il n'est plus exercé, confronté ou renouvelé.**

Le paradoxe du *Brain* de l'entreprise : capitaliser pour ne pas perdre, automatiser sans se déposséder

Le chapitre 5 a présenté le « Brain de l'entreprise » comme un vocabulaire encore émergent, associé à l'idée de rendre les connaissances, données et outils de l'organisation mobilisables par des systèmes d'IA. Sans présumer de la stabilisation du terme, cette architecture fournit ici un cas particulièrement révélateur des tensions de souveraineté.

L'organisation veut constituer son Brain parce qu'elle veut tirer parti du potentiel de l'IA, capitaliser ce qu'elle sait, diffuser l'expertise, accélérer la décision et éviter que le savoir disparaisse avec les personnes. Le chapitre 10 a montré que l'IA peut effectivement permettre de capitaliser l'expertise de professionnels expérimentés et de la rendre accessible à des opérateurs moins expérimentés.

Mais plus elle encode son fonctionnement dans ce Brain, plus elle doit vérifier qu'elle ne transforme pas progressivement **le patrimoine vivant qui produit la connaissance en patrimoine figé qui restitue la connaissance passée.**

Ce patrimoine codifié devient par ailleurs plus facilement mobilisable par la machine. Cette disponibilité accroît l'enjeu de maîtrise des conditions dans lesquelles il peut être interrogé, combiné, transféré, enrichi ou réutilisé.

La question de l'innovation devient alors centrale : **le patrimoine distinctif de l'organisation - technologique, cognitif, métier ou stratégique - protège ce que l'organisation possède, la dynamique d'innovation protège sa capacité à produire ce qu'elle ne possède pas encore.**

Une organisation peut encore posséder ses compétences actuelles, ses procédures, ses bases de connaissances et ses experts tout en réduisant progressivement sa faculté à faire émerger de nouvelles alternatives, de nouvelles compétences, de nouvelles méthodes.

Le phénomène rejoint directement l'**expertise orpheline** : le système continue de produire une décision, une recommandation ou une action alors que les personnes savent de moins en moins l'expliquer, la contester ou la corriger.

Le risque du Brain est donc au moins double : dépendre des technologies qui le rendent opérable et devenir dépendant de la représentation passée de l'organisation elle-même.

Un Brain souverain devrait mémoriser ce que l'organisation sait tout en contribuant à maintenir les conditions par lesquelles elle continue d'apprendre, de contester, de reformuler et d'inventer.

La maîtrise, la vulnérabilité et la substituabilité permettent donc de qualifier les dépendances présentes. La capacité dynamique oblige à regarder également **leur trajectoire** : les décisions prises aujourd'hui augmentent-elles ou réduisent-elles les possibilités d'action, de décision, de préservation et de renouvellement du patrimoine distinctif, et d'innovation de demain ?

Une dépendance n'est dès lors pas seulement à qualifier par ce qu'elle est, mais aussi par **ce qu'elle fait devenir à l'organisation.**

La question de souveraineté devient ainsi moins : « que maîtrisons-nous aujourd'hui ? » que : « que devons-nous continuer à pouvoir maîtriser, quelles nouvelles maîtrises devons-nous acquérir, et quelles capacités devons-nous entretenir ou construire pour y parvenir ? »

Cette lecture replace la question initiale « qu'est-il critique de ne pas maîtriser ? », dans une dynamique temporelle au lieu de la figer dans l'instant présent. Une grille maîtrise-vulnérabilité-substituabilité, complétée par l'importance de l'usage et les conséquences de la perte, permet de distinguer les dépendances qui peuvent être acceptées, celles qui doivent rester auditables et substituables et celles qui exigent une capacité de maîtrise plus directe.

Plusieurs entretiens expriment une inquiétude similaire : les investissements et implantations en France ne garantissent pas nécessairement une maîtrise française ou européenne lorsque les capitaux, les opérateurs, les éditeurs, les modèles et les interfaces restent largement extra-européens.

« Investisseurs, opérateurs, éditeurs : beaucoup d'acteurs structurants restent américains. Les data centers peuvent être en France avec des capitaux américains », entretien avec un opérateur.

11.6 Sixième angle mort : l'expertise publique indépendante

Les cinq angles morts précédents nécessitent tous une capacité d'expertise publique indépendante suffisante. Et les entretiens se rejoignent sur ce point : cette capacité manque en France et en Europe, aussi bien au niveau national que local.

L'asymétrie est structurelle. D'un côté, des opérateurs industriels disposant d'équipes techniques spécialistes, d'avocats spécialisés, de directions publiques aguerries et d'une capacité à présenter leurs projets sous le meilleur angle. De l'autre, des collectivités et des services de l'État qui manquent des compétences nécessaires pour évaluer sérieusement les bilans environnementaux, qualifier les usages, comparer les offres, négocier les contreparties et anticiper les évolutions technologiques.

Cette asymétrie favorise les entreprises qui savent comment présenter leurs projets et pénalise les décideurs qui ne disposent pas de points de comparaison indépendants. Elle est aussi une condition de dépendance : une collectivité qui n'a pas les outils pour instruire un projet de data center risque d'accepter ou de refuser à l'aveugle.

« Problème : l'aspect friable de l'expertise data center en France et en Europe (public comme privé). Les grands acteurs des data centers sont surtout anglo-saxons. », entretien avec un expert de l'énergie.

Cette lacune se manifeste à plusieurs niveaux. Au niveau national, des compétences partielles sur les data centers existent, mais, à notre connaissance, aucune n'a un mandat systématique, des ressources dédiées et une indépendance suffisante pour jouer le rôle d'analyse neutre à destination des décideurs publics. Au niveau local, la Métropole d'Aix-Marseille-Provence fait figure d'exception et a développé des outils propres (simulateur fiscal, géomatique, délibération métropolitaine). La grande majorité des collectivités n'a pas ces moyens.

La question qui persiste est celle du mandat : qui doit produire des analyses techniques neutres, avec quelles ressources, quelle indépendance et pour quels décideurs ? L'absence de décision sur ce point est elle-même une décision : celle de laisser les arbitrages se faire dans l'asymétrie. Cette lacune d'expertise devient alors une lacune stratégique.

11.7 Septième angle mort transversal : l'absence d'apprentissage collectif

Chaque collectivité, chaque gestionnaire de réseau, chaque service de l'État instruit aujourd'hui un projet de data center sans accès systématique à ce que les autres territoires ont déjà appris sur des

projets comparables. Cette situation produit une renégociation à l'aveugle : les mêmes questions de raccordement, de chaleur fatale, de consommation d'eau, de fiscalité, d'emplois, de réversibilité ou de valeur locale sont reposées projet après projet, sans capitalisation structurée.

Un retour d'expérience national utile devrait couvrir des données vérifiables et comparables : valeur créée, captée et partagée localement ; consommation réelle d'électricité au regard de la puissance contractualisée ; montée en charge effective ; flexibilité mobilisée ; chaleur réellement valorisée ; consommation réelle d'eau ; type et origine de l'eau mobilisée, WUE réel, technologie de refroidissement et niveau de pression hydrique locale ; incidents ; fiscalité nette ; emplois réels ; contraintes de raccordement ; effets territoriaux ; conditions de sortie en fin d'exploitation.

Ce retour d'expérience ne devrait pas être conçu comme une simple base documentaire, mais comme une infrastructure d'apprentissage public. Son objectif ne serait pas de rendre publics des secrets industriels, mais de sortir chaque territoire de la négociation isolée : comparer les projets dans le temps, capitaliser les conditions réellement obtenues, identifier les écarts entre promesses et réalisations et documenter les effets concrets des projets accueillis.

Ce qui vaut pour une organisation vaut à l'échelle collective : une capacité stratégique qui n'est pas exercée, confrontée et enrichie se dégrade. L'apprentissage collectif devient ainsi l'une des conditions de maintien dans le temps du capital de souveraineté à l'échelle collective.

Les données les plus sensibles - contreparties négociées, incidents, montants, clauses contractuelles - pourraient être anonymisées par défaut, avec des niveaux d'accès différenciés : données non agrégées pour l'État et les autorités compétentes ; données comparatives agrégées et anonymisées pour les collectivités ; données nationales consolidées pour les chercheurs ; version non sensible pour l'information des citoyens.

Un tel dispositif constituerait le socle d'un apprentissage collectif : non seulement mesurer, mais apprendre ; non seulement déclarer, mais comparer ; non seulement autoriser, mais réviser les doctrines à partir des retours d'expérience. Les collectivités qui négocient un nouveau projet pourraient s'appuyer sur les conditions déjà obtenues ailleurs. Les gestionnaires de réseau pourraient affiner leurs hypothèses de raccordement avec des données de réalisation plutôt qu'avec les seules puissances contractualisées. L'État et les chercheurs pourraient évaluer l'écart entre puissance réservée, puissance effectivement appelée et valeur réellement partagée.

Sans ce dispositif, l'apprentissage reste individuel et localisé, alors que la décision d'accueillir un projet engage des ressources communes - réseau électrique, foncier, eau, capacités publiques d'instruction - qui dépassent le périmètre de chaque territoire. L'absence d'apprentissage collectif reconstruit ainsi, à chaque négociation, l'asymétrie entre opérateurs et puissance publique.

11.8 Ce que ces angles morts ont en commun

Les outils disponibles : code de l'urbanisme, procédures d'autorisation environnementale, règles d'accès au réseau, cadre fiscal local, ont été conçus pour d'autres objets, dans un autre contexte. Ils ne sont pas forcément adaptés à un **objet qui est à la fois un immeuble, une infrastructure de réseau, un actif financier, un enjeu de souveraineté, un enjeu territorial/local et un problème environnemental en même temps**.

Synthèse : ce que les sept angles morts exigent ensemble

Gouvernance des données hébergées : conditionner les autorisations à la transparence sur investisseurs, clients et usages

Cycle de vie et fin d'exploitation : plan de réversibilité obligatoire, provision financière, priorité au brownfield instruit

Effet foncier et coût d'opportunité : méthodologie standard pour objectiver ce que le data center ne laisse pas faire sur le même terrain

Mesure de l'empreinte réelle : obligations de déclaration vérifiables, métriques sectorielles adaptées, accès public

Souveraineté par couches et dans le temps : qualifier les dépendances par MVS (grille maîtrise / vulnérabilité / substituabilité), identifier les capacités critiques à préserver et suivre si les choix effectués renforcent ou érodent le capital de souveraineté.

Expertise publique indépendante : décider qui produit les analyses neutres, avec quel mandat et quelles ressources

Absence d'apprentissage collectif : infrastructure d'apprentissage public qui s'appuie sur un registre national capitalisant les données de réalisation projet après projet, accessible avec différenciation aux collectivités, gestionnaires de réseau, chercheurs et État.

Sans traitement de ces angles morts, la doctrine risque de rester partielle et vulnérable à des arbitrages conduits dans l'opacité, au détriment des territoires, des usages concurrents et de la stratégie européenne.

Ces angles morts appellent un changement de posture : **passer de l'accueil opportuniste à l'accueil conditionnel**. C'est l'objet du chapitre suivant : transformer ces angles morts en critères d'instruction, en contreparties et en doctrine d'accueil.

Chapitre 12 - Accueillir sans subir : instrumenter la décision IA-énergie

Le chapitre en une phrase

Plutôt qu'une liste de mesures juxtaposées, le chapitre propose un schéma unique - sur le modèle d'un P&ID industriel - organisant sept recommandations en un même circuit de décision : qualifier la demande, orienter vers les systèmes utiles, instruire la souveraineté par couches, construire une salle de contrôle organisationnelle, faire de la puissance publique un levier de marché, outiller les collectivités, faire de chaque projet de data center un projet stratégique multi-acteurs et multi-échelles.

Pourquoi une liste ne suffit pas ?

Un rapport qui s'achève sur une liste de recommandations risque de trahir le travail qui l'a précédé. Onze chapitres ont montré que la demande IA, les infrastructures, les territoires, les organisations et le système énergétique ne sont pas des sujets séparés qu'on pourrait traiter l'un après l'autre. Ce sont des éléments d'un même système, qui se répondent en permanence. Clore ce travail par une énumération de mesures juxtaposées reviendrait à défaire, dans la dernière page, ce que toutes les précédentes ont patiemment construit.

Ce chapitre ne propose donc pas une liste isolée de recommandations, mais un schéma qui permet de comprendre où chacune doit agir.

Il faut imaginer le type de dessin qu'on trouve à l'entrée d'une installation industrielle. Il montre par où les choses entrent, où elles sont vérifiées, où elles peuvent être orientées, ralenties ou arrêtées et quels signaux permettent aux opérateurs de reprendre la main en cas de problème. Dans l'industrie, ce type de schéma s'appelle un P&ID, pour tuyauterie et instrumentation.

Cette image métaphorique correspond à ce que la doctrine doit produire : il ne suffit pas de savoir qu'un système existe, il faut savoir où le mesurer, où l'orienter, où le conditionner, où l'arrêter, où le réviser et qui est capable de le piloter.

Le système à instrumenter

Une demande de calcul entre dans le système, encore brute, non qualifiée, portée par l'offre technologique, les besoins métier réels, les ambitions territoriales, les stratégies industrielles. Cette demande traverse une série de points de contrôle avant de se matérialiser dans une infrastructure physique, qui prélève à son tour de l'énergie, du foncier, de l'eau, de l'attention publique et qui doit, en retour, restituer une part de ce qu'elle a prélevé au système qui l'a rendue possible.

Rien dans ce circuit n'est à sens unique : chaque section renvoie de l'information vers les sections précédentes, pour que la doctrine se corrige elle-même plutôt que de se figer au jour de son déploiement.

Six fonctions structurent le circuit de décision. Une septième fonction, la salle de contrôle organisationnelle, agit de manière transversale pour relier ces étapes, organiser l'apprentissage et maintenir la capacité d'arbitrage. Le schéma représente ainsi le fonctionnement de la doctrine ; les recommandations qui suivent précisent les leviers permettant de la mettre en œuvre.

Le circuit de décision IA-énergie

Qualifier, orienter, maîtriser, conditionner, instruire, contractualiser – puis mesurer et réviser.



Figure 4 - Conception de l'autrice, image réalisée par une IA

Sept recommandations pour mettre en œuvre la doctrine

Le circuit précédent décrit comment instruire une décision. Sa mise en œuvre repose sur sept recommandations qui ne reprennent pas mécaniquement chacune des étapes du circuit : certaines portent sur le processus lui-même, d'autres sur les capacités organisationnelles, l'action de la puissance publique, le rôle des collectivités ou les conditions d'accueil des data centers.

Ce schéma n'a de valeur que s'il se traduit en gestes précis :

1. **Qualifier toute demande IA structurante** avant qu'elle ne devienne un achat, une architecture, un contrat ou une infrastructure.
2. **Orienter l'effort vers les systèmes utiles** : énergie, industrie, infrastructures critiques, territoires, services publics, transition écologique et usages à gain net démontrable.
3. **Instruire la souveraineté comme une capacité** de maîtrise, de substitution et de sortie, et non comme une simple localisation.
4. **Construire une salle de contrôle organisationnelle** : compétences hybrides, temps d'apprentissage protégé, intégrateurs IA-systèmes critiques, prévention de l'expertise orpheline, capacité de dissidence et arbitrage continu.
5. **Faire de la puissance publique un levier de marché**, en rendant opposables la sobriété, la transparence, la réversibilité, la preuve de gain net et les clauses de réexamen.
6. **Donner aux collectivités les moyens d'instruire les effets territoriaux** avant d'autoriser : foncier, eau, chaleur, raccordement, valeur locale, acceptabilité, réversibilité.
7. **Faire de chaque projet de data center un projet stratégique** multi-acteurs et multi-échelles, à instruire dans ses dimensions territoriale, économique, sociale, écologique, temporelle et matérielle.

Ces sept recommandations ne sont pas hiérarchisées entre elles. Elles décrivent des dimensions complémentaires d'un même dispositif, mais n'agissent pas toutes au même niveau. Certaines relèvent directement de l'organisation et de ses choix d'usage, d'architecture ou de gouvernance ; d'autres supposent l'intervention de la puissance publique, des collectivités ou des opérateurs d'infrastructure. Leur mise en œuvre doit donc être adaptée au rôle et au périmètre d'action de chaque acteur. Elles restent néanmoins interdépendantes : aucune ne peut produire durablement ses effets si les autres dimensions du système demeurent ignorées.

Chacune de ces sept recommandations est approfondie dans un fascicule pratique dédié, incluant des outils d'instruction directement mobilisables, dont une grille d'instruction non compensatoire de la souveraineté par couches.

La France ne doit pas seulement accueillir l'IA. Elle doit choisir l'IA qu'elle veut rendre possible, instruire les infrastructures qui la supportent, qualifier les usages qui la tirent, contractualiser les contreparties et préserver sa capacité de sortie.

Quel portage à l'échelle nationale ?

La salle de contrôle organisationnelle décrite dans la quatrième recommandation ne répond pas, à elle seule, à la question du pilotage public national. La coordination de cette doctrine ne peut relever d'un seul ministère sectoriel. Les arbitrages concernent simultanément le numérique, l'énergie, l'industrie, l'environnement, la recherche, la sécurité, l'aménagement du territoire, les services publics et la commande publique. Les confier à l'un de ces périmètres seulement conduirait à privilégier une dimension du système au détriment des autres.

Cette doctrine appelle donc un portage interministériel, assorti d'une responsabilité politique clairement attribuée, d'une capacité d'arbitrage et d'un appui technique permanent. Son rôle ne serait pas de se substituer aux ministères, aux gestionnaires de réseau ou aux collectivités, mais d'organiser la cohérence entre leurs décisions, de consolider les données, de suivre les engagements, de capitaliser les retours d'expérience et de rendre visibles les conflits entre objectifs.

Cette coordination doit aussi rendre explicite la chaîne de responsabilité. Pour chaque décision structurante, il devrait être possible d'identifier qui décide, qui finance, qui exécute et qui répond des résultats ou de l'échec. Lorsque ces responsabilités sont dispersées ou mal attribuées, la capacité collective de décision peut rester formelle alors même que les moyens existent.

On pourrait objecter qu'une doctrine plus exigeante risque de faire perdre à la France des projets qui iraient chercher ailleurs des conditions plus permissives. Ce risque est réel, mais une doctrine sélective viserait précisément à privilégier les projets dont la maturité, la valeur locale et la transparence sont les mieux établies (voir 1.2, six pièges). Une allocation sélective permettrait de qualifier l'attractivité française, sans la réduire.

Les encadrés suivants approfondissent quatre des sept recommandations et montrent comment les décliner dans leur registre propre : la souveraineté, l'organisation, les data centers et les territoires.

Traduction souveraineté : cinq questions pour instruire une dépendance et sa trajectoire¹²⁰

- 1. Que maîtrisons-nous vraiment ?** Modèles, données, cloud, API, interfaces, capitaux, câbles, puces, compétences, valeur captée.
- 2. Où sommes-nous vulnérables ?** Prix, accès, conditions contractuelles, juridiction,

¹²⁰ Ces questions reprennent la grille maîtrise / vulnérabilité / substituabilité présentée au chapitre 4 et détaillée en 4.5

disponibilité, fournisseur, chaîne matérielle, dépendance métier.

3. **Que pouvons-nous substituer ?** Modèle, fournisseur, architecture, outil, interface, compétence, procédure ou capacité de reprise.
4. **Quelle trajectoire créons-nous ?** Nos choix actuels renforcent-ils ou érodent-ils, dans le temps, nos capacités d'action, de décision, de préservation et de renouvellement du patrimoine distinctif, et d'innovation ?
5. **Quel capital de souveraineté devons-nous entretenir ou construire ?** Quelles compétences, alternatives éprouvées, capacités de reprise, architectures réversibles, marges de négociation et capacités d'apprentissage doivent être maintenues ou développées ?

Tout choix structurant d'IA devrait être instruit à l'entrée, puis réexaminé lors des changements significatifs de fournisseur, de modèle, d'architecture, de périmètre d'automatisation ou de conditions contractuelles. Cette revue doit porter à la fois sur le profil MVS et sur l'évolution des capacités d'action, de décision, de préservation et de renouvellement du patrimoine distinctif, et d'innovation.

Une décision est évaluée à la fois par la dépendance qu'elle crée aujourd'hui et par les options qu'elle ferme ou qu'elle ouvre demain.

Traduction organisationnelle : six gestes de frugalité pragmatique

1. **Qualifier** les usages avant de dimensionner.
2. **Voir l'IA visible et l'IA cachée.**
3. **Objectiver le coût complet**, les impacts et les dépendances.
4. **Borner les usages sensibles**, lourds ou agentiques.
5. **Contractualiser la réversibilité**, l'auditabilité et les exigences fournisseurs.
6. **Réviser ou arrêter** ce qui ne tient pas sa promesse.

Traduction data centers : huit contreparties à instruire

Accueillir un data center comme un **contrat stratégique**, c'est rendre explicite une exigence : à chaque ressource rare mobilisée doit correspondre une contrepartie vérifiable.

1. **Puissance contre maturité**
Ne pas réserver une capacité sans preuve de projet prêt.
2. **Raccordement contre transparence**
Savoir quels usages, quels clients et quels profils de charge seront servis.
3. **Foncier contre valeur territoriale**
Mesurer fiscalité nette, emplois réels, coût d'opportunité, valeur laissée localement et capacité du territoire à peser sur la trajectoire du projet.
4. **Chaleur contre contribution aux réseaux**
Contractualiser la chaleur effectivement récupérable, utile et suivie.
5. **Eau contre sobriété**
Mesurer la consommation réelle d'eau, distinguer l'eau directe de l'empreinte hydrique indirecte liée à la production d'électricité, puis rapporter ces données au niveau de pression hydrique locale.

6. **Souveraineté affichée contre maîtrise par couches**¹²¹

Vérifier capitaux, opérateurs, modèles, données, juridictions, équipements et valeur captée.

7. **Investissement contre réversibilité**

Prévoir la seconde vie, sortie, restitution du foncier, capacités raccordées et responsabilités.

8. **Capacité de calcul contre usages qualifiés**

Distinguer entraînement, inférence, cloud généraliste, usages critiques, services publics ou usages de confort.

Traduction territoriale : trois postures face aux infrastructures IA

1. **Éviter le cas par cas**

Ne pas attendre l'arrivée des projets pour découvrir les critères d'instruction.

2. **Orienter**

Définir en amont les conditions d'accueil : foncier, eau, chaleur, raccordement, valeur locale, acceptabilité, réversibilité.

3. **Expérimenter**

Réserver ou affecter des capacités de calcul comme des actifs territoriaux au service de la recherche, des hôpitaux, des PME, des services publics, de l'industrie locale ou de la transition écologique.

Accueillir sans subir, ce n'est pas ralentir l'IA : c'est équiper le pays pour voir ce qu'elle mobilise, choisir ce qu'elle sert, conditionner ce qu'elle exige et reprendre la main lorsque le système dérive.

¹²¹ Voir encadré souveraineté et particulièrement la question 1.

Conclusion générale

Gouverner la réciprocité entre IA et énergie

Ce rapport est parti d'une relation encore trop souvent pensée à moitié. D'un côté, l'intelligence artificielle est décrite comme une nouvelle demande énergétique, matérialisée par les data centers, les GPU, l'inférence, les raccordements et les chaînes industrielles du calcul. De l'autre, l'énergie est présentée comme un domaine que l'IA pourrait optimiser, automatiser ou rendre plus sobre. L'enjeu de ce rapport a été de tenir ces deux mouvements ensemble : l'IA transforme la demande d'énergie, tandis que l'IA peut aussi transformer les systèmes énergétiques eux-mêmes.

La question n'est donc pas seulement de savoir si la France doit accueillir l'IA, raccorder des data centers ou encourager des usages numériques plus sobres. Elle est de savoir **comment gouverner une interdépendance nouvelle entre calcul, énergie, infrastructures, territoires, organisations, souveraineté et transition écologique**. Cette interdépendance peut devenir une capacité stratégique si elle est instruite, arbitrée et pilotée. Elle peut aussi devenir une dépendance supplémentaire si elle est laissée aux seules logiques d'annonce, de marché, de confort logiciel ou d'urgence technologique.

Le premier enseignement du rapport est que l'IA n'est pas immatérielle. Elle engage des ressources physiques : électricité, eau, foncier, chaleur, câbles, serveurs, GPU, métaux, bâtiments, groupes de secours, compétences et capacités industrielles. Le data center rend cette matérialité visible, mais il ne l'épuise pas. L'inférence intégrée dans les logiciels, les API, les suites bureautiques, les agents et les services SaaS diffuse aussi une demande de calcul souvent invisible pour les organisations qui l'utilisent. L'empreinte de l'IA ne se situe donc pas seulement dans les grands sites visibles. Elle se loge aussi dans les usages ordinaires, les architectures logicielles, les contrats fournisseurs et les chaînes d'inférence.

Le deuxième enseignement est que l'énergie n'est pas seulement une contrainte pour l'IA. Elle devient l'un de ses terrains d'usage les plus stratégiques. Prévision, optimisation, maintenance, pilotage, gestion des flexibilités, bâtiments intelligents, pompes à chaleur, supervision industrielle, intégration des renouvelables : les organisations énergétiques montrent que l'IA peut contribuer à mieux utiliser certaines ressources, à sécuriser certains processus et à améliorer certains arbitrages. Mais cette promesse doit rester instruite. Une IA au service de l'énergie ne devient pas automatiquement frugale, utile ou bénéfique. Elle doit démontrer ce qu'elle économise, ce qu'elle consomme, ce qu'elle transforme, ce qu'elle rend plus robuste et ce qu'elle rend plus dépendant.

Le troisième enseignement est que la demande IA n'est pas homogène. Entraînement, inférence, fine-tuning, *edge* industriel, cloud généraliste, usage critique, usage de divertissement ou de confort, automatisation agentique ou optimisation énergétique ne posent pas les mêmes questions. Tous ne justifient pas les mêmes infrastructures, les mêmes localisations, les mêmes niveaux de puissance, les mêmes exigences de continuité ni les mêmes contreparties territoriales. Qualifier l'usage devient donc une condition préalable à toute décision sérieuse. Sans qualification des usages, les infrastructures risquent d'être dimensionnées par défaut pour une demande indifférenciée.

Le quatrième enseignement est que la frugalité ne commence pas au moment d'optimiser un modèle. Elle commence au moment de qualifier un besoin. Faut-il vraiment de l'IA pour cet usage ? Le bénéfice attendu est-il démontré ? Existe-t-il un scénario sans IA ? Le modèle est-il proportionné ? L'infrastructure est-elle maîtrisée ? Les effets rebond sont-ils suivis ? Les impacts humains et sociaux sont-ils instruits ? La frugalité devient alors une discipline d'arbitrage entre

usages, coûts, souveraineté, robustesse, sécurité, dépendances, impacts et valeur créée. Elle ne consiste pas à faire moins en toute circonstance, mais à consommer juste pour un usage dont la valeur peut être démontrée.

Le cinquième enseignement est que la souveraineté ne se décrète pas par localisation. Héberger en France ne signifie pas maîtriser en France. La souveraineté se joue dans les capitaux, les opérateurs, les clients, les modèles, les données, les puces, les câbles, les logiciels, les API, les interfaces, les compétences et la capacité à substituer une solution à une autre. Cette lecture par couches évite deux impasses : croire qu'il faudrait tout maîtriser, ce qui est irréaliste, ou considérer que la localisation physique suffit, ce qui est insuffisant. La question stratégique devient : quelles couches critiques doivent être maîtrisées, lesquelles peuvent rester dépendantes, lesquelles doivent être rendues substituables, et lesquelles créent une vulnérabilité si elles disparaissent ou changent de conditions d'accès ?

Le sixième enseignement est que les territoires ne peuvent pas être de simples lieux d'accueil. Ils sont les lieux où se matérialisent les ressources mobilisées et les externalités produites : foncier, eau, chaleur, raccordement, voirie, fiscalité, bruit, groupes de secours, emplois, compétences, acceptabilité et réversibilité. Pour un opérateur mondial, un site peut être une option parmi d'autres. Pour une collectivité, c'est un engagement de long terme. Cette asymétrie impose une ingénierie publique indépendante, des critères d'accueil explicites, une négociation documentée et des contreparties vérifiables.

Le septième enseignement est que les organisations doivent apprendre à gouverner l'IA comme une infrastructure d'usage, et non comme une simple couche logicielle. Un copilote, un agent, un modèle de langage, un routeur d'inférence ou une suite logicielle enrichie à l'IA appellent des infrastructures, des données, de l'énergie, des contrats, des coûts, des dépendances et des responsabilités. Les organisations doivent donc rendre visibles les usages, suivre les coûts complets, contractualiser les exigences fournisseurs, organiser la réversibilité, former les métiers et mesurer les bénéfices dans la durée. Sans cela, l'IA risque d'être reçue plus qu'elle n'est choisie.

Ces enseignements convergent vers une même exigence : **gouverner la relation IA-énergie comme un système**. Cette gouvernance suppose de relier ce qui est encore trop souvent séparé : les usages et les infrastructures, les raccordements et les clients finaux, les promesses de chaleur et les réseaux existants, les gains d'efficacité et les effets rebond, la localisation et la maîtrise, la performance logicielle et les ressources matérielles, l'innovation et la réversibilité. Le rapport ne plaide donc pas pour une couche supplémentaire de complexité administrative. Il plaide pour une capacité d'instruction proportionnée à la complexité réelle de l'objet.

Cette doctrine appelle un **changement de posture publique**. Il ne s'agit pas seulement d'accélérer ou de ralentir. Il s'agit de qualifier, sélectionner, conditionner, contractualiser, suivre et réviser. L'accueil des data centers doit devenir un contrat stratégique explicite. Les usages IA doivent être hiérarchisés selon leur valeur, leurs coûts complets et leurs effets induits. Les organisations énergétiques doivent pouvoir expérimenter l'IA sans perdre la maîtrise de leurs systèmes critiques. Les collectivités doivent disposer des moyens de négocier. Les acheteurs publics et privés doivent intégrer la frugalité, la transparence, la réversibilité et la souveraineté par couches dans leurs critères de choix.

La France et l'Europe ne gagneront pas nécessairement en singeant les hyperscalers. Elles ne disposent pas des mêmes volumes de capitaux, des mêmes plateformes mondiales, des mêmes

marchés captifs ni des mêmes positions industrielles dans la chaîne complète du calcul. Leur stratégie peut être différente : **devenir critiques sur les couches qui conditionnent la décision, l'intégration, la confiance, la robustesse, la frugalité, l'explicabilité, la souveraineté opérationnelle et les usages industriels.** Dans un monde de ressources finies, ces attributs peuvent devenir des avantages stratégiques, pas seulement des contraintes.

Gouverner la réciprocité entre IA et énergie, c'est refuser deux simplifications. La première consisterait à voir l'IA seulement comme une charge énergétique supplémentaire. La seconde consisterait à voir l'IA seulement comme un levier d'optimisation de l'énergie. Elle est les deux à la fois : une demande nouvelle et un outil potentiel de transformation. C'est précisément cette double nature qui impose une doctrine.

La France et l'Europe disposent d'une fenêtre stratégique : qualifier les usages avant de dimensionner les infrastructures, rendre l'inférence visible, mesurer les gains nets, conditionner les ressources rares, construire des capacités souveraines par couches, renforcer l'ingénierie publique et apprendre collectivement des usages réels. L'enjeu n'est pas de choisir entre développement de l'IA et transition énergétique. Il est de faire de leur relation un objet de décision, de maîtrise et de responsabilité.

La question finale est : **quels usages IA la France veut-elle rendre possibles, avec quelles ressources, dans quels territoires, sous quelles garanties et au bénéfice de quelle stratégie industrielle, énergétique et démocratique ?**

Glossaire

N.B. : autant que possible, les définitions s'appuient sur des références normatives. Afin de ne pas alourdir la lecture, celles-ci ne sont pas systématiquement mentionnées.

Edge / edge computing

Mode de traitement dans lequel tout ou partie du calcul est réalisé à proximité de la source des données, des équipements ou des utilisateurs, plutôt que dans une infrastructure centrale distante.

Effet Jevons

Situation dans laquelle un gain d'efficacité peut entraîner une hausse des volumes telle que la consommation totale peut rester stable ou augmenter. Dans l'IA, un service plus rapide, moins coûteux ou plus accessible peut susciter suffisamment d'usages supplémentaires pour annuler tout ou partie du gain initial.

D'après William Stanley Jevons, The Coal Question, Macmillan, 1865 ; voir également Missemer, 2012, et Alcott, 2005.

Effet rebond

Hausse de consommation pouvant résulter d'un gain d'efficacité, d'une baisse de coût ou d'une plus grande facilité d'usage. Il peut être direct lorsque l'usage rendu plus efficace augmente lui-même ou indirect lorsque les économies réalisées sont réaffectées à d'autres biens ou services. L'ITU-T distingue aussi des rebonds liés au temps, à l'espace et à l'échelle de l'économie.

Entraînement / training

Processus permettant de déterminer ou d'ajuster les paramètres d'un modèle d'apprentissage automatique à partir de données et d'un algorithme d'apprentissage.

Fine-tuning / ajustement fin

Entraînement complémentaire d'un modèle pré-entraîné afin de l'adapter à une tâche, un domaine ou un jeu de données plus spécifique.

GPU - Graphics Processing Unit

Processeur conçu pour effectuer de nombreux calculs en parallèle et largement utilisé pour accélérer l'entraînement et l'exécution de modèles d'IA.

Inférence

Processus par lequel un système d'IA produit une conclusion ou un résultat à partir d'éléments disponibles. Pour un modèle entraîné, cela correspond notamment à son exécution sur de nouvelles données afin de produire une prédiction, une recommandation, une classification ou un contenu.

Latence

Délai entre l'envoi d'une information ou d'une requête et la réception du résultat attendu. Une faible latence est notamment recherchée pour les usages interactifs ou certains systèmes industriels en temps contraint.

Modèle de fondation / foundation model

Modèle d'IA pouvant être utilisé directement ou adapté à une grande variété de tâches dans un ou plusieurs domaines.

PUE - Power Usage Effectiveness

Indicateur d'efficacité énergétique d'un data center, calculé comme le rapport entre l'énergie totale consommée par le site et l'énergie consommée par ses équipements informatiques. Plus le PUE se rapproche de 1, plus la part des consommations auxiliaires est faible.

RAG - Retrieval-Augmented Generation / génération augmentée par récupération

Architecture dans laquelle un modèle génératif récupère des informations dans une source externe afin de les utiliser comme contexte pour produire sa réponse.

Robustesse

Capacité d'un système d'IA à conserver des performances comparables lorsqu'il est confronté à de nouvelles données ou à des conditions d'exploitation représentatives. Dans ce rapport, le terme désigne la robustesse du système d'IA, et non la robustesse au sens plus large des systèmes organisationnels ou socio-écologiques.

Routeur de modèles

Composant qui choisit, parmi plusieurs modèles disponibles, celui auquel adresser une requête selon des critères tels que la tâche, le coût, la performance, la latence ou la disponibilité.

SLM - Small Language Model

Modèle de langage de taille relativement réduite par rapport aux grands modèles de langage, souvent privilégié pour des tâches spécialisées, des contraintes de coût, de latence ou d'exécution locale. Il n'existe pas de seuil universel de nombre de paramètres définissant un SLM.

Token

Unité de contenu traitée comme significative par un modèle d'IA. Selon le modèle, un token peut correspondre à un mot, une partie de mot, un signe ou une autre unité de contenu.

WUE - Water Usage Effectiveness

Indicateur d'efficacité hydrique d'exploitation d'un data center, rapportant la consommation d'eau du site à l'énergie consommée par ses équipements informatiques. Il doit être interprété avec le type d'eau utilisé, la technologie de refroidissement et le contexte hydrique local.

