AFNOR SPEC 2314

JUIN 2024

Contacter :
AFNOR – Norm'Info
11, rue Francis de Pressensé
93571 La Plaine Saint-Denis Cedex
Tél : 01 41 62 76 44
Fax : 01 49 17 92 02
E-mail : norminfo@afnor.org

AFNOR SPEC

Juin 2024

Référentiel général pour l'IA frugale

Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

www.afnor.org

MINISTÈRE DE LA TRANSITION ÉCOLOGIQUE ET DE LA COHÉSION DES TERRITOIRES
Liberté
Égalité
Fraternité

ECO LAB
Innovation - Écologie - Territoires
Commissariat Général au Développement Durable

afnor

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et
réduire l'impact environnemental de l'IA

Sommaire

Remerciements

Le Ministère de la Transition écologique et de la Cohésion des Territoires et l'AFNOR remercient tout particulièrement les pilotes des groupes de travail qui ont fourni un travail remarquable dans l'élaboration de ce document de référence : Juliette Fropier (Ecolab / Ministère de la transition écologique), Pierre Monget (Hub France IA), Gilles Ribeaucourt (Muvraline), Pierre Riou (ACIMEO), Ana Semedo (ILExpansions), Denis Trystram (Université Grenoble Alpes, EcoInfo).

Ils tiennent également à adresser leurs remerciements à tous les contributeurs à ce groupe de travail, et notamment à quelques contributeurs particulièrement impliqués dans les travaux :

Ludovic Arga (Orange), Mathieu Aubry (École des Ponts ParisTech), Rémi Barrère (Thales), Vincent Baudevin (Maestis), Sylvain Baudoin (The Shift Project), Pierre-Loïc Bayart (MIAI Grenoble Alpes), Yannick Benetreau (La Poste), Jean-François Berrée (CEA-List), Eddie Bonnal, Alexis Bonnerat (AB Frame), Thomas Brilland (Ademe), Aurélie Bugeau (LaBRI), Romain Carbou (Orange), Grégory Cazanave (IRT AESE), Nathalie Charbonniaud (Orange), Victor Charpenay (Mines de Saint-Étienne), Karin Dassas (CNRS), Thomas de Latour (Ademe), Fanny Deleuze (Reworld Media), Grégoire Demarest (ARCEP), Alice Drahon (Scalian), Hammouda Elbez (CITC), Marc Fiammante, Emmanuelle Frenoux (Université Paris-Saclay), Gabriel Grandamy (Tech4us), Agathe Granier (Ekitia), Clara Grojean (AFNUM), Gaël Guennebaud (INRIA), Tristan Hamard (ARCOM), Line Hinderer (MTECT), Thibaud Hugard (APC Climat), Rémy Ibarcq, Jacques Kluska (Schneider Electric), Loïc Landrieu (ENPC ParisTech), Marine Le Gall, Claude Le Pape (Schneider Electric), Clément Le Roux (NAIA), Laurent Lefevre (ENS Lyon), Marc Léobet (Mens Data), Théophile Lenoir (Université de Milan), Enoal Lepoittevin (EyeSnap), Anne-Laure Ligozat (ENSIIE, EcoInfo), Francesca Martini (Groupe La Poste), Georges Matissart (Département Seine-et-Marne), Eric Mattman (Hub France IA), Nicolas Museux (Thales), Tom Nico (ARCEP), Sarah Oury (SopraSteria), Nicolas Palix (Université Grenoble Rhône-Alpes), Jérémy Pastouret (Les Enovateurs), Maxime Peralta (CEA), Franck Pramotton (The Shift Project), Constant Razel (EXXA), Emilie Regnier-Lody (adeq), Antoine Rigouleau (Orange), Anthea Serafin (Ekitia), Anne Tozzolino (Groupe La Poste), Laurent Tripier (bziiit), Luc Truntzler (Hub France IA), Noémie Vaudry-Blaise (Capgemini), Claire Verdier (Openstudio), Mae Yener (Hymaia).

Le Ministère de la Transition écologique et de la cohésion des Territoires et l'AFNOR remercient les organisations Hub France IA, EcoInfo, Ademe et ARCEP, qui ont compris la nécessité impérieuse de ce travail et ont joint leur expertise à l'effort de rédaction de cette Spec.

Les copilotes et les contributeurs remercient très chaleureusement Juliette Fropier (Ministère de la Transition écologique et de la Cohésion des Territoires) pour l'animation de cette AFNOR Spec, ainsi que Anna Médan (AFNOR) pour l'organisation et le soutien tout au long du projet.

Avant-propos

Le présent document a été développé par un groupe de travail ouvert et reflète à ce titre l'accord de personnes et organisations ayant participé à son élaboration. AFNOR a mis à disposition des auteurs son savoir-faire en ingénierie normative afin de coordonner les travaux d'élaboration et éditer le document. En conséquence, le contenu de ce document n'engage que ses auteurs et ne saurait être considéré comme constituant le droit applicable. En effet, AFNOR n'étant ni habilitée à délivrer du conseil juridique ni législateur, AFNOR ne saurait être tenue responsable de l'utilisation qui est faite de ce document, notamment concernant la règlementation éventuellement citée dont la bonne application relève exclusivement de la responsabilité de chacun.

L'AFNOR SPEC :

— est un document technique développé et approuvé dans le cadre d'un processus transparent et ouvert ;

— représente l'approbation de ce seul groupe de travail sur le texte final et ne doit pas être présentée comme une norme française ou comme équivalente à une norme française.

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

Responsable du projet

MINISTÈRE DE LA TRANSITION ÉCOLOGIQUE ET DE LA COHÉSION DES TERRITOIRES
Liberté Égalité Fraternité

ECO LAB
Innovation - Écologie - Territoires
Commissariat Général au Développement Durable

Ecolab, le laboratoire de l'innovation au service de la transition écologique est situé au sein du Commissariat Général du Développement Durable (CGDD). Le CGDD, acteur interministériel et direction transversale du ministère en charge de l'environnement, éclaire et alimente, par la production de données et d'analyses, l'action du ministère. Il propose une vision d'ensemble des enjeux environnementaux.

Ce document a été produit dans le cadre de la Stratégie Nationale pour l'Intelligence Artificielle. Cette stratégie a pour ambition de préserver et consolider la souveraineté économique, technologique et politique de la France et de mettre l'IA au service de l'économie et de la société. Elle est pilotée par le coordinateur national pour l'Intelligence Artificielle.

GOUVERNEMENT
Liberté Égalité Fraternité
Stratégie nationale pour l'intelligence artificielle

Avec la participation financière de la direction du Numérique Responsable et le pôle Data/IA du Groupe La Poste

LA POSTE GROUPE

Principaux contributeurs

RÉPUBLIQUE FRANÇAISE
Liberté Égalité Fraternité

ADEME
AGENCE DE LA TRANSITION ÉCOLOGIQUE

EcoInfo
POUR UNE INFORMATIQUE ÉCO-RESPONSABLE

HUB FRANCE IA

AFNOR SPEC

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

CONTRIBUTEURS

ACIMEO

I L Expansions

muvraline

UGA Université Grenoble Alpes

adeq
Votre RSE
des idées aux actes

AFNUM

APCC

Boavizta

bziiit

Capgemini

cea

CGI

cnrs
dépasser les frontières

École des Ponts ParisTech

ekitia
Data et IA en pleine confiance

emvista

ensiie
Paris | Évry-Courcouronnes

ENTROPISME

esri
get it right

exxa

EYESNAP

AFNOR SPEC 2314

Référentiel général pour l'IA frugale

Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

Contexte et enjeux

Le présent document vise à fournir un premier référentiel opérationnel sur l'impact environnemental d'un système et d'un service d'Intelligence Artificielle (IA) et les moyens de promouvoir des services frugaux d'IA pour toutes les organisations, publiques comme privées. Sa rédaction se veut volontairement simple et accessible pour être utilisable par un grand nombre d'acteurs. Ce guide n'a en aucun cas l'ambition de se substituer aux obligations légales ou réglementaires existantes et à venir. En revanche, il a celle d'aider à élaborer de futures normes européennes et internationales. Il tient compte du Règlement européen pour l'Intelligence Artificielle [3].

L'objectif de ce document est de partager un guide des définitions, méthodes et bonnes pratiques utiles pour l'évaluation et la réduction de l'impact environnemental d'un système d'IA et des services numériques qui l'utiliseraient. Ce document s'adresse à tous les acteurs des services numériques et à ceux cherchant à mettre en place un service recourant à l'IA. Il contribue aux objectifs de responsabilité sociale et environnementale des acteurs économiques, et sensibilise les individus utilisateurs du service. Enfin, ce document est une base méthodologique pour l'intégration de critères environnementaux dans les achats de services incluant un système d'IA, par exemple par les acheteurs publics.

Les enjeux sociaux et éthiques liés au développement de l'IA sont exclus du périmètre de ce document, malgré leur intrication parfois élevée avec les enjeux environnementaux. La frugalité est discutée ici sous l'angle environnemental et le document ne prend pas parti sur l'utilité des services dans les différents secteurs comme ceux de la santé ou de la sécurité. Par ailleurs, sur d'autres enjeux fondamentaux liés aux systèmes d'IA, comme la sécurité, nous renvoyons aux documents publiés par les autorités compétentes [1)].

Ce projet a été élaboré de janvier 2024 à juin 2024, à partir des contributions successives de plus d'une centaine de membres issus des organisations participantes. Il a été lancé à l'initiative du Ministère de la Transition Écologique et de la Cohésion des Territoires, plus particulièrement des équipes de l'Ecolab du Commissariat Général au Développement Durable.

1) Recommandations de sécurité pour un système d'IA générative, publié le 29 avril 2024, https://cyber.gouv.fr/publications/recommandations-de-securite-pour-un-systeme-dia-generative

Bibliographie des ressources socles

[1] ISO/IEC 22989:2022, Technologies de l'information — Intelligence artificielle — Concepts et terminologie relatifs à l'intelligence artificielle

[2] Référentiel méthodologique d'évaluation environnementale des services numériques, BUREAU VERITAS, APL - datacenter, DDemain, ADEME, 2021

[3] RÈGLEMENT DU PARLEMENT EUROPÉEN ET DU CONSEIL établissant des règles harmonisées concernant l'intelligence artificielle et modifiant les règlements (CE) n° 300/2008, (UE) n° 167/2013, (UE) n° 168/2013, (UE) 2018/858, (UE) 2018/1139 et (UE) 2019/2144 et les directives 2014/90/UE, (UE) 2016/797 et (UE) 2020/1828 (règlement sur l'Intelligence Artificielle)

[4] ITU-T L.1410 – Méthodologie pour les évaluations environnementales du cycle de vie des biens, réseaux et services des technologies de l'information et la communication

[5] ITU-T L.1450 – Méthodologies d'évaluation de l'impact environnemental du secteur des technologies de l'information et de la communication

[6] ITU-T L.1480 - Méthodologie d'évaluation des incidences de l'utilisation de solutions fondées sur les technologies de l'information et de la communication (TIC) sur les émissions de gaz à effet de serre d'autres secteurs.

[7] ISO 14040:2006, Management environnemental — Analyse du cycle de vie — Principes et cadre

[8] ISO/IEC 5338:2023, Technologies de l'information — Intelligence artificielle — Processus de cycle de vie des systèmes d'IA

[9] ISO/IEC CD TR 20226, Information technology — Artificial intelligence — Environmental sustainability aspects of AI systems

[10] Référentiel général de l'écoconception des services numériques, Arcep, Arcom, ADEME, DINUM, CNIL, Inria, 2024

[11] AFNOR SPEC 2201, Écoconception des services numériques, avril 2022

1 Terminologie

Afin de mettre en oeuvre ce référentiel général pour l'IA frugale, ce chapitre vise à rappeler des définitions socles sur l'IA et les impacts environnementaux (1.1 à 1.7) et de poser quelques définitions autour de l'IA frugale (1.7 et 1.8).

1.1 Définitions de l'Intelligence Artificielle (IA) et des données

1.1.1 Intelligence Artificielle (discipline)

Recherche et développement de mécanismes et d'applications de systèmes d'IA (d'après la norme ISO/IEC 22989:2022 [1])**.**

NOTE En tant que discipline scientifique, l'IA comprend plusieurs approches et techniques, telles que l'apprentissage automatique (dont l'apprentissage profond et l'apprentissage par renforcement sont des exemples spécifiques), le raisonnement automatique (qui comprend la planification, l'ordonnancement, la représentation et le raisonnement des connaissances, la recherche et l'optimisation), et la robotique (qui comprend le contrôle, la perception, les capteurs et les actionneurs, ainsi que l'intégration de toutes les autres techniques dans les systèmes cyber-physiques) [2)].

1.1.2 Système d'Intelligence Artificielle (IA)

Système technique qui génère des sorties telles que du contenu, des prévisions, des recommandations ou des décisions pour un ensemble d'objectifs définis par l'humain (d'après la norme ISO/IEC 22989:2022 [1]).

NOTE 1 Le système d'IA est un système automatisé conçu pour fonctionner à différents niveaux d'autonomie, qui peut faire preuve d'une capacité d'adaptation après son déploiement et qui, pour des objectifs explicites ou implicites, déduit, à partir des données d'entrée qu'il reçoit, la manière de générer des résultats tels que des prédictions, du contenu, des recommandations ou des décisions qui peuvent influencer les environnements physiques ou virtuels (d'après le Règlement IA [3]).

NOTE 2 Un système d'IA peut être étudié selon trois segments : les algorithmes, le matériel et les données permettant le développement de ce système. Les systèmes d'IA à usage général ont la capacité de répondre à diverses finalités, et donc à plusieurs destinations. Les systèmes d'IA spécialisés ont un usage spécifique, donc une seule destination.

1.1.3 Destination

Utilisation à laquelle un système d'IA est destiné par le fournisseur interne ou externe, y compris le contexte et les conditions spécifiques d'utilisation, telles que précisées dans les informations communiquées par le fournisseur dans la notice d'utilisation, les éventuelles indications publicitaires ou de vente et les déclarations, ainsi que dans la documentation technique (d'après le Règlement IA [3]).

1.1.4 Service numérique

Activité se caractérisant par la réalisation d'une prestation ou la mise à disposition d'une information mobilisant un ensemble d'équipements, infrastructures numériques et d'autres services numériques pour capter, faire circuler, traiter, analyser, restituer et stocker des données (extrait du Référentiel d'évaluation environnementale des services numériques de l'Ademe [2]).

---

2) Extrait des Lignes directrices éthiques pour une Intelligence Artificielle digne de confiance, Groupe d'experts de haut niveau de la Commission européenne sur l'Intelligence Artificielle, Avril 2019.

1.1.5 Service d'Intelligence Artificielle (IA)

Service numérique incluant un système d'IA en principal ou en accessoire. Le système d'IA inclus dans le service peut avoir été développé en interne ou par un tiers.

NOTE Dans le reste de la Spec, pour le terme « service d'IA », on entend aussi bien un service augmenté par IA qu'un service reposant majoritairement sur l'IA.

1.1.6 Produit d'Intelligence Artificielle (IA)

Élément matériel incluant un système d'IA, que ce dernier soit développé en interne ou par un tiers.

NOTE Le système d'IA peut fonctionner en interne de l'élément matériel (IA embarquée) ou échanger des données avec un système d'IA hébergé dans des centres de données distincts de l'élément matériel.

Exemples de produits et services d'IA : génération augmentée par la recherche de documents, filtre de *spam*, génération d'images, de voix ou de vidéo, modélisation et simulations de scénarios (détection de fuite dans les réseaux d'eau, consommation d'énergie, urbanisme), etc.

1.1.7 Performance d'un Système d'Intelligence Artificielle (IA)

Capacité d'un système d'IA à remplir sa destination (issu du Règlement IA [3]).

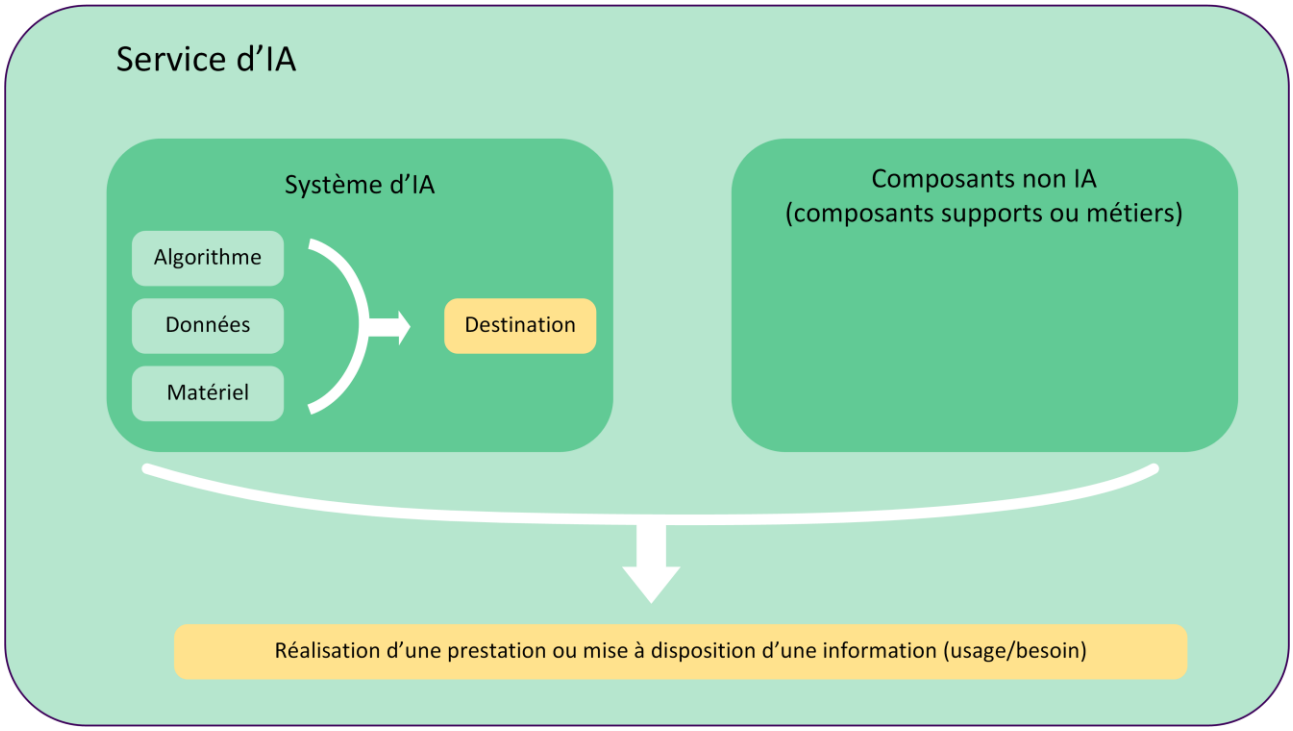

Figure 1 — Concept de service et de système d'IA

1.1.8 Données d'entraînement

Données utilisées pour entraîner un système d'IA en ajustant ses paramètres entraînables.

1.1.9 Données de validation

Données utilisées pour fournir une évaluation du système d'IA entraîné et pour régler ses paramètres non entraînables ainsi que son processus d'apprentissage, afin, notamment, d'éviter tout sous-ajustement ou sur-ajustement.

1.1.10 Données de test

Données utilisées pour fournir une évaluation indépendante du système d'IA, afin de confirmer la performance de ce système avant sa mise sur le marché ou sa mise en service.

1.1.11 Données d'entrée

Données fournies à un système d'IA ou directement acquises par celui-ci et à partir desquelles il produit un résultat.

1.1.12 Données de sortie

Données produites par un système d'IA.

1.2 Définitions des parties prenantes de l'IA

1.2.1 Fournisseur de plateforme d'IA

Organisme ou entité qui fournit des services qui permettent à d'autres parties prenantes de produire des services ou des produits d'IA (d'après la norme ISO/IEC 22989:2022 [1]).

NOTE On comprend ici les acteurs de l'infrastructure matérielle et logicielle, y compris ceux qui réalisent le stockage des données utilisées par l'IA et fournissent les ressources pour traiter, analyser et distribuer les données.

1.2.2 Fournisseur de produit ou de service d'IA

Organisme ou entité qui fournit des produits ou services d'IA qui sont directement utilisables par un client ou qui sont à intégrer dans un système qui utilise l'IA conjointement avec des composants non IA (d'après la norme ISO/IEC 22989:2022 [1]).

NOTE Cette définition est similaire à la définition de *fournisseur* du Règlement IA [3] (organisme qui développe ou fait développer un système d'IA et le met sur le marché ou met le système d'IA en service).

1.2.3 Producteur d'IA

Organisme ou entité qui conçoit, qui développe, qui soumet à l'essai et qui déploie des produits ou des services qui utilisent un ou plusieurs systèmes d'IA (d'après la norme ISO/IEC 22989:2022 [1]).

1.2.4 Client d'IA

Organisme ou entité qui utilise un produit ou un service d'IA, soit directement soit en le fournissant à des utilisateurs d'IA (d'après la norme ISO/IEC 22989:2022 [1]).

NOTE Cette définition est similaire à la définition de *déployeur* dans le Règlement IA [3] (organisme utilisant sous sa propre autorité un système d'IA, sauf lorsque ce système est utilisé dans le cadre d'une activité personnelle à caractère non professionnel).

1.2.5 Partenaire d'IA

Organisme ou entité qui fournit des services dans le contexte de l'IA, par exemple :

- Intégrateur de systèmes d'IA : organisme ou entité qui s'occupe de l'intégration de composants d'IA dans des systèmes plus étendus, qui éventuellement comprennent également des composants non IA.
- Fournisseurs de données : organisme ou entité qui s'occupe de fournir les données utilisées par des produits ou des services d'IA.

  NOTE On entend ici les acteurs qui créent des données et collectent la donnée (éventuellement avec des capteurs). Le fournisseur de données peut être interne ou externe à l'organisation. Les données peuvent être ou non partagées entre plusieurs services d'IA et/ou pour plusieurs finalités.
- Auditeur d'IA : organisme ou entité qui s'occupe de réaliser l'audit des organismes qui produisent, fournissent ou utilisent des systèmes d'IA, afin d'évaluer la conformité aux normes, aux politiques ou aux exigences légales.
- Évaluateur d'IA : organisme ou entité qui évalue les performances d'un ou de plusieurs systèmes d'IA.

1.2.6 Sujet d'IA

Organisme ou entité qui est affecté(e) par un système, un service ou un produit d'IA (d'après la norme ISO/IEC 22989:2022 [1]).

1.2.7 Autorité compétente

Organisme ou entité habilité à agir dans le domaine des systèmes, des services ou des produits d'IA (d'après la norme ISO/IEC 22989:2022 [1]).

- Décideur : organisme ou entité habilité à définir des politiques dans un domaine international, régional, national ou industriel qui peuvent avoir un impact sur un système, un service ou un produit d'IA.
- Régulateur : organisme ou entité habilité à définir, mettre en œuvre et faire respecter les exigences légales prévues dans les politiques définies par les décideurs politiques.

1.2.8 Organisation de la société civile

Structure organisationnelle dont les membres servent l'intérêt général au moyen d'un processus démocratique et jouent un rôle de médiation entre les pouvoirs publics et les citoyens [3)].

NOTE Une organisation de la société civile peut réunir des partenaires sociaux, être une organisation non gouvernementale, ou une association par exemple.

1.3 Ressources numériques

1.3.1 Ressource de calcul

Élément permettant la réalisation d'opérations de calcul qui peut aller du microcontrôleur jusqu'au *cluster* de calcul équipé de processeurs (exemples : CPU, GPU, LPU, TPU, etc.).

1.3.2 Ressource de collecte de données

Élément servant à l'acquisition des données (exemple : capteurs, caméras, etc.).

1.3.3 Ressource de stockage

Élément permettant de stocker des données.

1.3.4 Ressource réseau

Élément permettant l'échange de données.

1.3.5 Terminal utilisateur

Élément permettant l'accès pour un utilisateur aux données et à la programmation des ressources.

1.4 Définitions liées à l'impact environnemental

1.4.1 Impact environnemental

Impact incluant les effets positifs et négatifs sur l'environnement (traduit de ITU L.1410 [4]).

1.4.2 Cycle de vie

Phases consécutives et liées d'un système de produits, de l'acquisition des matières premières ou de la génération des ressources naturelles à l'élimination finale (d'après la norme ISO 14040:2006 [7]).

[3)] Avis du Comité économique et social sur « Le rôle et la contribution de la société civile organisée dans la construction européenne », Journal officiel n° C 329 du 17/11/1999 p. 0030.

1.4.3 Analyse du cycle de vie (ACV)

Compilation et évaluation des intrants, des extrants et des impacts environnementaux potentiels d'un système de produits au cours de son cycle de vie (d'après la norme ISO 14040:2006 [7]).

1.4.4 Unité fonctionnelle (UF)

Performance [4] quantifiée d'un système de produits destinée à être utilisée comme unité de référence dans une analyse du cycle de vie (d'après la norme ISO 14040:2006 [7]).

NOTE 1 L'unité fonctionnelle (UF) est l'unité de mesure utilisée pour évaluer le service rendu par le produit. De la même manière que pour comparer le prix de deux fruits, un consommateur rapporte les prix au kilo ; pour comparer les impacts environnementaux de deux produits, on ramènera les impacts à une unité de mesure commune.

NOTE 2 Dans une évaluation des effets de second ordre (voir 1.4.9), le scénario où un service d'IA est utilisé et le scénario de référence (où le service d'IA n'est pas utilisé) sont considérés comme des systèmes de produits et l'unité fonctionnelle doit être choisie de manière à la rendre applicable aux deux.

NOTE 3 L'unité fonctionnelle définit le scénario de référence et les caractéristiques de performances délivrées par le scénario où le service d'IA est utilisé. L'unité fonctionnelle doit avoir une fonction et une unité quantifiable mesurant la performance de cette fonction (traduit de ITU-T L.1480 [6]).

1.4.5 Inventaire du cycle de vie

Phase de l'analyse du cycle de vie impliquant la compilation et la quantification des intrants et des extrants, pour un système de produits donné au cours de son cycle de vie (d'après la norme ISO 14040:2006 [7]).

1.4.6 Catégorie d'impact

Classe représentant les points environnementaux étudiés à laquelle les résultats de l'inventaire du cycle de vie peuvent être affectés (d'après la norme ISO 14040:2006 [7]).

1.4.7 Indicateur de catégorie d'impact (ou indicateur environnemental)

Représentation quantifiable d'une catégorie d'impact (d'après la norme ISO 14040:2006 [7]).

1.4.8 Étape du cycle de vie

Une étape parmi plusieurs étapes consécutives et liées d'un système de produits (traduit de ITU-T L.1410 [4]).

1.4.9 Effet de second ordre

Impact indirect créé par l'utilisation et l'application des technologies d'information et de communication (TIC), qui inclut des modifications des impacts environnementaux dus à l'utilisation des TIC qui pourraient être positives ou négatives (traduit de ITU-T L.1480 [6]).

---

[4] Ici, le terme de performance repris directement de l'ISO 14040:2006 comme une unité d'utilisation, et n'est pas entendue de la même manière que la définition de « performance » indiquée plus haut et extraite du Règlement européen IA.

1.4.10 Effet d'ordre supérieur

Effet indirect (y compris, mais sans s'y limiter, les effets rebond) autre que les effets de premier et de second ordre, se produisant à travers des changements dans les modes de consommation, les modes de vie et les systèmes de valeurs (traduit de ITU-T L.1480 [6]).

NOTE 1 Il s'agit d'effets « en cascade », comme par exemple le fait que la possibilité de commander « en un clic » provoque plus d'achats, ou que les algorithmes de recommandation provoquent une augmentation de consommation (de produits, de vidéos, de musique, etc.).

NOTE 2 Les effets rebond, d'induction définis ci-dessous sont des effets de second ordre ou d'ordre supérieur. Les effets de substitution et d'optimisation de la destination du système d'IA sont également des effets d'ordre secondaire ou d'ordre supérieur.

1.4.11 Effet rebond

Augmentation de la consommation due à des interventions d'efficacité environnementale qui peuvent se produire par le biais d'une réduction de prix, de consommation d'énergie ou d'un autre mécanisme incluant des réponses comportementales (c'est-à-dire qu'un produit efficace est moins cher ou plus pratique d'une autre manière et est donc consommé dans une plus grande mesure) (traduit de ITU-T L.1480 [6]).

NOTE Les effets rebonds peuvent être distingués en :

- Effets rebond directs : Un effet rebond où une efficacité accrue, une réduction des coûts associée et/ou la commodité d'un produit ou d'un service entraînent son utilisation accrue parce qu'il est moins cher ou autrement plus pratique.
- Effets rebond indirects : Un type d'effet rebond dans lequel les économies réalisées grâce à la réduction des coûts d'efficacité permettent de consacrer davantage de revenus à d'autres produits et services.
- Effets rebond macroéconomiques et transformations sociétales : Effet rebond où une plus grande efficacité stimule la productivité économique globale, ce qui se traduit par davantage de croissance économique et de consommation au niveau macroéconomique (traduit de ITU-T L.1480 [6]).

1.4.12 Effet d'induction

Réduction ou augmentation des émissions qui se produisent en dehors du cycle de vie ou de la chaîne de valeur d'une solution mais résultant de l'utilisation de cette solution (traduit de ITU-T L.1480 [6]).

1.5 Cycle de vie [5)] du système d'IA

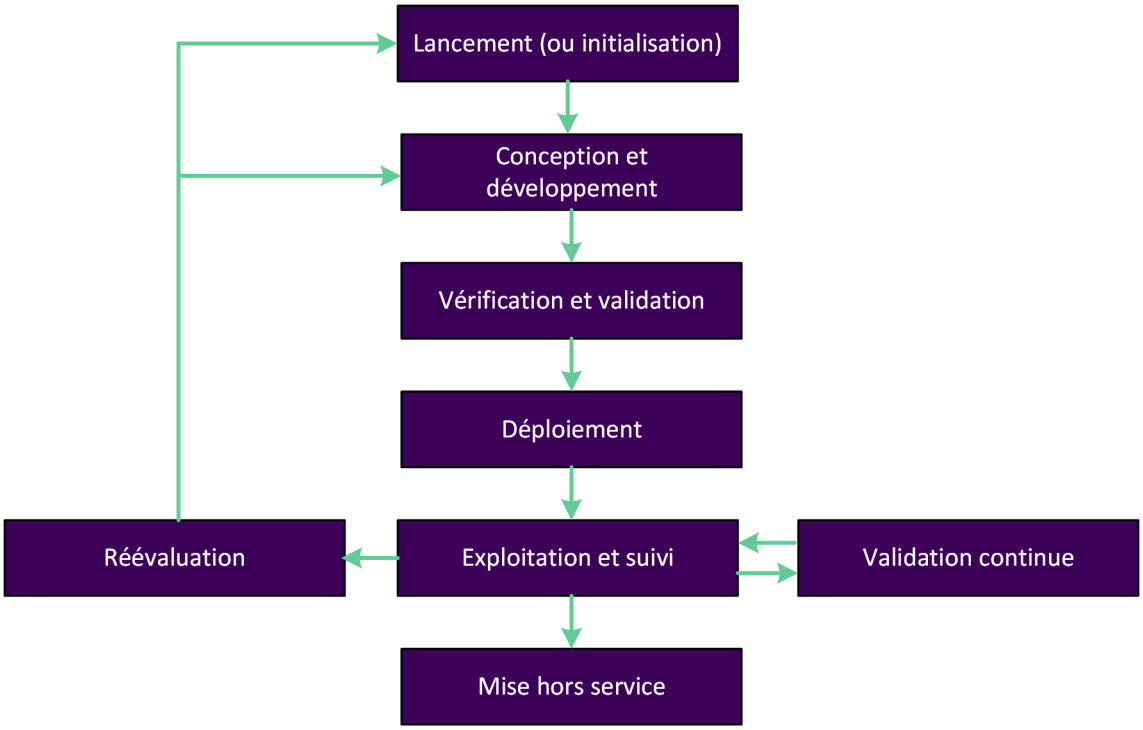

Figure 2 — Cycle de vie d'un système d'IA

NOTE L'entraînement et les inférences sont deux termes couramment utilisés par les parties prenantes de l'IA. L'entraînement consiste à créer et sélectionner des algorithmes ou des combinaisons d'algorithmes, pendant l'étape de conception et de développement (voir 1.5.2). Le ré-entraînement d'algorithmes, par la distillation ou la compression, fait également partie de la phase de conception et développement. L'inférence consiste à utiliser ces algorithmes pendant les étapes de vérification et validation (voir 1.5.3), puis, en cas de succès, pendant les étapes d'exploitation et de suivi (voir 1.5.5).

1.5.1 Lancement (ou Initialisation)

Une ou plusieurs parties prenantes décident de transformer une idée en un système concret, avec la décision de passer à l'étape de conception et développement (d'après la norme ISO/IEC 22989:2022 [1] et le rapport technique ISO/IEC 20226 [9]).

1.5.2 Conception et développement

Étape de création du système d'IA, qui résulte en un système d'IA prêt pour l'étape de vérification et validation (d'après la norme ISO/IEC 22989:2022 [1] et le rapport technique ISO/IEC 20226 [9]).

1.5.3 Vérification et validation

Vérification que le système d'IA issu de l'étape de conception et de développement fonctionne conformément aux exigences et atteint les objectifs définis lors du lancement (d'après la norme ISO/IEC 22989:2022 [1] et le rapport technique ISO/IEC 20226 [9]).

---

5) Ici, le terme de « cycle de vie » vise à expliciter les différentes étapes de la mise en oeuvre d'un système d'IA pour décrire les actions de réduction des impacts environnementaux d'un système d'IA. Il diffère du terme générique utilisé pour une analyse en cycle de vie de système de produits (voir 1.4.3), de l'extraction de matières premières à la fin de vie.

1.5.4 Déploiement

Le système d'IA est installé, déployé ou configuré pour fonctionner dans un environnement cible (d'après la norme ISO/IEC 22989:2022 [1] et le rapport technique ISO/IEC 20226 [9]).

1.5.5 Exploitation et suivi

Le système d'IA fonctionne et est généralement disponible pour une utilisation (d'après la norme ISO/IEC 22989:2022 [1] et le rapport technique ISO/IEC 20226 [9]).

1.5.6 Validation continue

Si le système d'IA utilise l'apprentissage continu, l'étape d'exploitation et de suivi est prolongée par une étape supplémentaire de validation continue. Un entraînement incrémental a lieu de façon continue tandis que le système fonctionne en production (d'après la norme ISO/IEC 22989:2022 [1] et le rapport technique ISO/IEC 20226 [9]).

1.5.7 Réévaluation

Après l'étape d'exploitation et de suivi, en fonction des résultats du travail du système d'IA, il peut apparaître nécessaire de procéder à une réévaluation (d'après la norme ISO/IEC 22989:2022 [1] et le rapport technique ISO/IEC 20226 [9]).

1.5.8 Mise hors service

Le système d'IA peut devenir obsolète dans la mesure où les réparations et les mises à jour ne suffisent pas à satisfaire aux nouvelles exigences, auquel cas il peut être mis hors service (d'après la norme ISO/IEC 22989:2022 [1] et le rapport technique ISO/IEC 20226 [9]).

1.6 Cycle de vie des données dans le cadre des systèmes d'IA

Pour objectiver l'impact environnemental du cycle de vie des données, les différentes étapes suivantes peuvent être prises en compte. Celles-ci ne suivent pas forcément un ordre successif précis [6)].

1.6.1 Spécification

Définition des données à produire pour répondre au besoin défini, ainsi que des critères de qualité requis pour l'usage considéré.

1.6.2 Collecte/acquisition

Moyens et méthodes de production des données tenant compte des critères de qualité définis, avec adjonction de métadonnées.

---

6) De la même manière que pour le cycle de vie du système d'IA, on cherche ici à décrire les différents processus auxquels sont soumis les données dans le cadre du système d'IA. Le terme « cycle de vie » est utilisé différemment du terme générique défini en 1.4.3, qui est pour des méthodes d'analyse en cycle de vie de systèmes de produits.

1.6.3 Vérification

Contrôle des données produites par rapport au besoin et aux critères de qualité préalablement définis.

1.6.4 Pré-traitement

Traitements d'adaptation des données (nettoyage des données selon les modes de collecte, changement de format, étiquetage pour l'apprentissage supervisé...) avant leur prise en compte dans le système d'IA, par exemple augmentation [7)].

1.6.5 Stockage

Recueil et conservation des données sur une ressource de stockage.

1.6.6 Transfert

Déplacement de données d'un système à un autre ou d'un réseau à un autre.

1.6.7 Mise à jour

Intégration de nouvelles données en reprenant le schéma collecte-prétraitement-intégration et vérification.

1.6.8 Archivage

Archivage intermédiaire, accessible aisément ou archivage définitif pour une longue durée.

1.6.9 Destruction

Suppression de l'intégralité des copies/archives des données. Cela peut être volontaire ou accidentel. La destruction des données stockées ou archivées suit généralement l'arrêt du service. Elle peut être fréquente dans le cas de données intermédiaires, temporaires ou générées.

1.7 Cycle de vie des ressources numériques

Pour des ressources numériques, les étapes du cycle de vie qui doivent être évaluées sont (d'après ITU L.1410 [4]) :

- Extraction des matières premières
- Production
- Utilisation
- Fin de vie

---

7) Un système d'IA peut nécessiter d'augmenter les données (par exemple découpe ou transfert de style sur des images). Cette augmentation peut être faite (i) avant l'entraînement d'un modèle, amenant à la création d'une nouvelle donnée stockée de manière pérenne, ou (ii) pendant l'entraînement, auquel cas ces augmentations seront faites plusieurs fois et aucune nouvelle donnée n'est créée de manière pérenne.

1.8 Système efficient et service frugal d'IA

La notion de frugalité est très large et sera traitée ici dans le contexte spécifique de l'IA. La frugalité peut être rapprochée de la notion de sobriété, pour laquelle « il n'existe pas de définition précise et consensuelle », mais celle-ci « assemble un continuum de démarches qui promeuvent – à différents degrés et à différentes échelles – une modération de la production et de la consommation de ressources énergétiques et matérielles, par une transformation des modes de vie au-delà de la recherche d'efficacité » [8)]. Le concept d'innovation frugale a par ailleurs fait l'objet d'un axe de recherche et l'article de Basu *et al.* le définit de la manière suivante : « L'innovation frugale est un processus d'innovation en matière de conception dans lequel les besoins et le contexte des citoyens des pays en développement sont donnés en premier afin de développer des services et des produits appropriés, adaptables, abordables et accessibles pour les marchés émergents. » [9)]

Les notions de frugalité et d'efficience doivent s'articuler autour des distinctions suivantes :

	Définition	Notions connexes	Raisonnement	Approche	Précisions
Efficience	Aptitude à optimiser les moyens alloués pour atteindre un résultat défini	Efficacité, optimisation	En relatif/par unité d'usage Le besoin prime : optimisation d'une solution jugée celle répondant le mieux au besoin	Recherche d'un optimum local ou d'un compromis sur un niveau de résultat fortement contraint	Prise en compte des effets de premier ordre pour les minimiser Prise en compte des parties prenantes de l'IA
Frugalité	Aptitude à se contenter d'un niveau de résultat jugé suffisant en redéfinissant les usages et les besoins	Sobriété (ou *Sufficiency* [10)] en anglais)	En global La contrainte sur les ressources prime : recherche de la solution utilisant le moins de ressources possible et apportant une réponse satisfaisante au besoin	Recherche d'un optimum global ou d'un compromis large sur un niveau de résultat, ce qui nécessite d'élargir ou d'assouplir le besoin	Prise en compte des effets de premier ordre et de second ordre pour minimiser les impacts environnementaux négatifs Prise en compte de tous les acteurs au-delà des seules parties prenantes de l'IA

8) Florian Cézard (AGATTE), Marie Mourad. 2019. Panorama sur la notion de sobriété – définitions, mises en œuvre, enjeux.

9) Traduit de Basu *et al.*, 2013.

10) A set of measures and daily practices that avoid demand for energy, materials, land and water while delivering human well-being for all within planetary boundaries. (Intergovernmental Panel on Climate Change, Climate Change 2022: Mitigation of Climate Change – Working Group III Contribution to the Sixth Assessment Report, 2022).

1.8.1 Système efficient d'Intelligence Artificielle (IA)

Système d'IA dont le fonctionnement prend en compte les mécanismes d'optimisation connus de l'état de l'art pour en réduire les besoins en ressources matérielles et énergétiques et les émissions associées, tout en conservant une performance suffisante. Ces mécanismes doivent être considérés à toutes les étapes du cycle de vie, par exemple :

- la quantité de données utilisées ;
- la taille et la complexité d'exécution des algorithmes (nombre de paramètres, nombre d'opérations, etc.) ;
- la consommation en ressources numériques et énergétiques (ex. localisation du centre de données).

1.8.2 Service frugal d'Intelligence Artificielle (IA)

La frugalité d'un service d'IA vise à réduire globalement les besoins en ressources matérielles et énergétiques et les impacts environnementaux associés via une redéfinition des usages ou des exigences de performance (voir 1.1.7), ou encore via une réorientation des besoins du producteur du système d'IA (amont) au fournisseur du service considéré. Un service frugal d'IA est donc un service pour lequel :

- la nécessité de recourir à un système d'IA plutôt qu'à une autre solution moins consommatrice pour répondre au même objectif a été démontrée ;
- de bonnes pratiques (*voir Chapitre « Bonnes pratiques »*) sont adoptées par le producteur, le fournisseur et le client d'IA pour diminuer les impacts environnementaux du service utilisant un algorithme d'IA ;
- les usages et les besoins visent à rester dans les limites planétaires [11)] et ont été préalablement questionnés.

NOTE 1 Le service respecte par ailleurs le principe du DNSH (*Do No Significant Harm*) européen. À ce titre, il ne doit pas porter préjudice de manière substantielle aux six objectifs environnementaux reconnus par la taxonomie verte : l'atténuation du changement climatique, l'adaptation au changement climatique, l'utilisation durable et la protection des ressources aquatiques et maritimes, la transition vers une économie circulaire, la prévention et le contrôle de la pollution, et la protection et la restauration de la biodiversité et des écosystèmes. Le fournisseur du service d'IA s'applique également à mettre en œuvre la démarche « Éviter, Réduire, Compenser » [12)].

NOTE 2 Le fournisseur d'IA doit également anticiper les usages imprévus de son système d'IA, qui viendraient impacter la frugalité du service.

---

11) Le concept des limites planétaires, proposé en 2009, révisé en 2015 (Steffen *et al.*), puis en 2023 (Richardson *et al.*), vise à définir un « espace de fonctionnement sûr pour l'humanité » qui repose sur l'évolution de neuf phénomènes complexes et interconnectés : le changement climatique, l'érosion de la biodiversité, la perturbation des cycles biogéochimiques de l'azote et du phosphore, le changement d'usage des sols, l'utilisation de l'eau douce, l'acidification des océans, l'appauvrissement de l'ozone stratosphérique, l'augmentation des aérosols dans l'atmosphère, l'introduction d'entités nouvelles dans la biosphère. Pour étudier l'évolution de ces phénomènes, une ou plusieurs « variables de contrôle » sont définies à l'échelle globale, voire régionale. Un « seuil » critique est fixé pour chacune de ces variables avec une « zone d'incertitude » constituée de deux valeurs : une valeur basse (« frontière planétaire ») et une valeur haute (« limite planétaire »). La frontière représente la zone de danger qui précède la limite au-delà de laquelle les écosystèmes pourraient basculer dans un état inconnu et probablement défavorable aux humains. (CGDD, La France face aux neuf limites planétaires, octobre 2023).

12) Ressources gouvernementales sur la séquence « Éviter, Réduire, compenser » : https://www.notre-environnement.gouv.fr/themes/evaluation/article/eviter-reduire-compenser-erc-en-quoi-consiste-cette-demarche

NOTE 3 La frugalité doit être considérée dans tous les aspects d'un service :

- dans les objectifs (Les objectifs sont-ils compatibles avec le respect des limites planétaires ?) ;
- dans la conduite du projet (par exemple questionnement des bénéfices du projet, définition d'un équilibre entre performance et coût environnemental, définition d'un budget en termes de ressources pour les différentes phases du projet, définition des algorithmes à évaluer et comparer en prenant en compte leurs coûts environnementaux de développement et d'utilisation) ;
- dans la phase de développement (par exemple, suivi du coût des diverses expériences, limitation du coût des expériences et du coût environnemental, limitations strictes des expériences les plus coûteuses, évaluation a priori des coûts et bénéfices potentiels des expériences) ;
- dans la phase d'utilisation (par exemple, évaluation du coût en ressources d'une utilisation, définition et contrôle des utilisations cibles, maintien de l'utilisation optimale des ressources de calcul) ;
- dans le choix des infrastructures (par exemple, évaluation du type d'infrastructure adaptée en prenant en compte la consommation de ressources, en particulier électrique, eau, métaux rares, en incluant le matériel informatique) ;
- dans les données utilisées, leur prétraitement et leur stockage.

1.8.3 Frugalité des usages d'un service d'IA

Réduction globale des besoins en ressources matérielles et énergétiques en questionnant l'usage avant développement et en maîtrisant l'usage qui est fait en aval du service (par exemple, en quantité).

NOTE Cet aspect dépasse le cadre du fournisseur de service, mais ce dernier peut s'en saisir *via* le choix du modèle économique et des mesures pour se prémunir d'éventuels effets rebonds.

1.8.4 Service frugal d'IA à bilan positif pour une catégorie d'impact à préciser

Service d'IA pour lequel les impacts positifs des usages [pour cette catégorie d'impact] sont supérieurs aux impacts négatifs du système [pour cette même catégorie] sur l'ensemble du cycle de vie du système d'IA et des composants non IA du service.

NOTE 1 L'évaluation des impacts du cycle de vie du service s'appuieront sur la méthodologie du Chapitre « Référentiel méthodologique d'évaluation environnementale ». Les méthodologies pour évaluer les apports du service sont spécifiques à chaque service.

NOTE 2 Notons que cette définition concerne une ou plusieurs catégories d'impact, considérées de façon indépendante, mais ne doit pas faire oublier le spectre d'impacts potentiels ou faire abstraction des autres catégories d'impact non traitées. Cette notion ne doit pas non plus inciter au transfert d'impact (c'est-à-dire, par exemple, à consommer plus d'eau pour réduire la consommation d'électricité du refroidissement d'un centre de données).

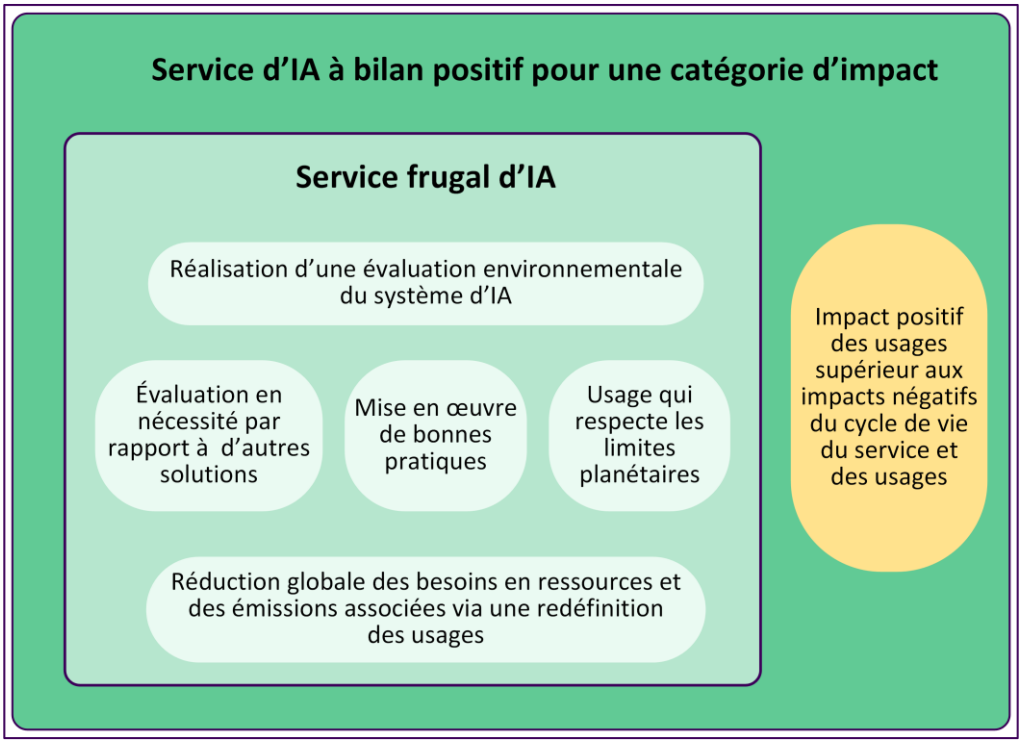

Figure 3 — Concepts de service frugal d'IA et de service à bilan positif sur une catégorie d'impact

2 Référentiel méthodologique d'évaluation environnementale

Ce chapitre vise à proposer une méthodologie de l'évaluation environnementale des systèmes d'IA. Les publics cibles de la présente partie recouvrent les intervenants dans la conception et la mise à disposition des services et systèmes d'IA à savoir :

- les producteurs d'IA, fournisseurs de plateformes, produits ou services d'IA, les partenaires d'IA ;
- les clients d'IA : organisme ou entité qui utilise un produit ou un service d'IA, soit directement soit en le fournissant à des utilisateurs d'IA ;
- les autorités compétentes.

L'objectif est de mettre à disposition des publics cibles les méthodes et les indicateurs pertinents permettant d'estimer ou de mesurer l'impact environnemental des services d'IA qu'ils conçoivent, commercialisent ou utilisent et d'effectuer les arbitrages en connaissance de cause.

Évaluer qualitativement ou quantitativement les impacts environnementaux d'un système d'IA et de son utilisation implique d'interroger les usages et les besoins, en lien avec les limites planétaires et d'évaluer l'efficacité du système. Cette évaluation ne peut être que relative (entre deux systèmes) et non absolue.

La méthodologie proposée s'appuie sur des normes et documents existants pour les services numériques :

- Prise en compte quantitative des impacts de premier ordre par une méthodologie basée sur l'Analyse du Cycle de Vie (ACV), s'appuyant sur les normes ISO, recommandation ITU-T L.1410 [4] et les règles par catégories de produit (usuellement RCP, pour Référentiel par Catégorie de Produit, ou PCR, pour *Product Category Rules*) ;
- RCP Services numériques [2] ;
- Prise en compte qualitative des impacts d'ordres supérieurs basée sur la recommandation ITU-T L.1480 [6].

On s'appuiera sur les définitions fondamentales issues de ces référentiels, en commençant par celle de l'unité fonctionnelle (voir 1.4.4), base des méthodes de calcul d'empreintes environnementales.

2.1 Impacts directs dus au cycle de vie des équipements (premier ordre)

2.1.1 Unité fonctionnelle

La définition d'une unité fonctionnelle doit répondre au questionnement suivant :

- la(les) fonction(s) assurée(s)/le(s) service(s) rendu(s) : « quoi ? ». On distinguera la fonction principale des fonctions secondaires ;
- l'ampleur de la fonction ou du service : « combien ? ». On précisera la fréquence, la durée ou la quantité d'utilisation ;
- le niveau de qualité souhaité : « comment ? » ;
- la durée (de vie) du produit : « combien de temps ? ».

Nous proposons de formaliser les unités fonctionnelles (UF) comme suit.

UF d'un Système d'IA : « Mettre à disposition le système sur une année pour x requêtes ». Cette période d'un an doit correspondre à une période d'inférence stabilisée (inférence constante ou de croissance connue).

La sortie sera de la forme :

$$ax + b + cy$$

a = coût environnemental d'une inférence

x = le nombre de requêtes sur un an

b = coût environnemental des entraînements et autres coûts fixes

c = coût environnemental d'un ré-entraînement

y = le nombre de ré-entraînements sur un an (peut être nul)

UF d'un Service d'IA : « Fournir le service sur une année à l'ensemble des utilisateurs ».

Cette UF repose sur l'UF précédente d'un Système d'IA, et prend en compte les serveurs supplémentaires (par exemple pour l'hébergement Web), les transferts réseaux, ainsi que les terminaux utilisateurs nécessaires à la fourniture du service, en incluant à la fois les composants IA et non IA.

Un schéma fonctionnel devra préciser le fonctionnement du service d'IA. L'Annexe 2 donne un exemple de schéma fonctionnel pour le service d'IA Stable Diffusion.

2.1.2 Périmètre et frontières du système

Le périmètre pris en compte sera le **périmètre global**, c'est-à-dire couvrant l'ensemble des éléments mobilisés pour permettre de délivrer le service numérique, qu'ils soient maîtrisés ou non par l'opérateur du service. Il inclura par exemple l'utilisation des réseaux ou d'heures de calcul sur des centres de données. Le niveau d'analyse correspond au **niveau 1** du RCP Services numériques, c'est-à-dire au service numérique dans sa globalité, incluant les tiers suivants :

- **Terminaux** utilisateurs, qui permettent l'exploitation, la réception et la consultation des contenus (smartphones, ordinateurs, téléviseurs, autres…) ainsi que le calcul en inférence en particulier dans le cas d'IA embarquée (brique 1 – ou *tier 1* - dans le RCP services numériques).
- **Réseaux**, qui permettent la transmission des données sur les infrastructures réseau vers les terminaux utilisateurs (brique 2 – ou *tier 2* - dans le RCP services numériques).
- **Centres de données**, qui permettent l'hébergement et le traitement des données numériques (serveurs, baies de stockage, équipement réseau…) (brique 3 – ou *tier 3* - dans le RCP services numériques).

L'**échelle** prise en compte est celle d'un **service** c'est-à-dire que l'unité fonctionnelle doit couvrir l'ensemble des éléments du périmètre permettant de délivrer le service à tous les utilisateurs.

Les étapes du **cycle de vie de chaque équipement** de chaque brique (terminal, réseau, centre de données/serveur) à prendre en compte sont les suivantes (suivant la décomposition du RCP Services numériques [2]) :

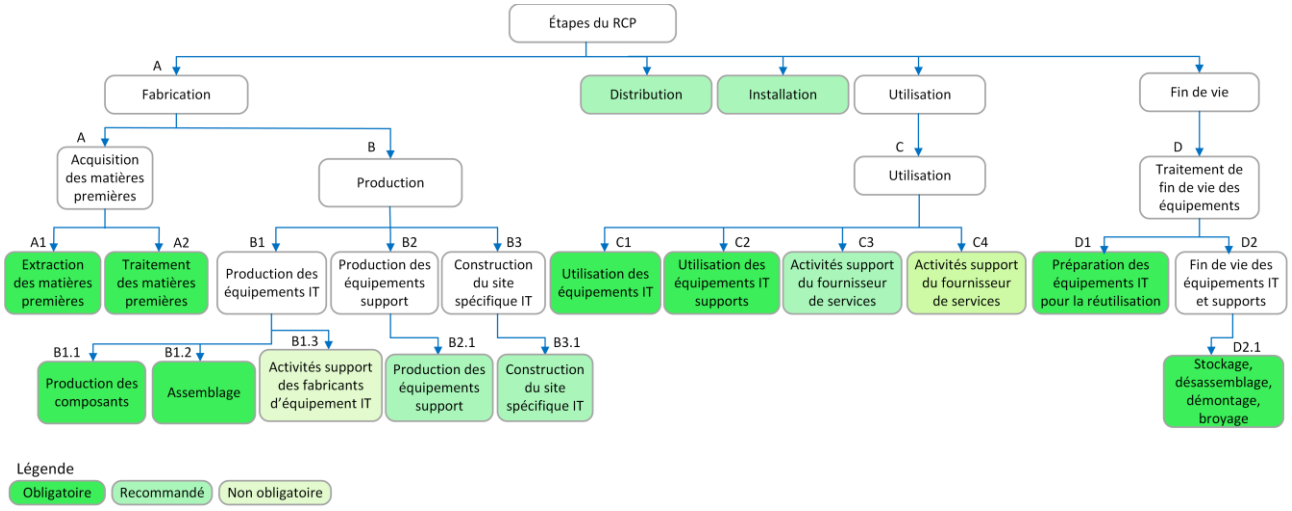

Figure 4 — Décomposition des étapes du cycle de vie de chaque équipement

Les équipements utilisés pour le service d'IA n'étant pas nécessairement en fin de vie à la fin de leur usage pour le service d'IA, il pourra être utile de préciser comment se passera la fin d'usage pour le service d'IA : (1) réemploi au sein de l'entreprise, (2) revente pour le réemploi, (3) revente pour recyclage, (4) filière de recyclage agréée (préciser laquelle), (5) don pour le réemploi (préciser), (6) autre (préciser).

En termes d'activités, le cycle de vie du système d'IA complet devra être pris en compte, suivant la décomposition définie au point 1.5 : lancement, conception et développement, vérification et validation, déploiement, exploitation et suivi, réévaluation, mise hors service.

Il faudra veiller à bien compter chaque entraînement (y compris ceux n'ayant pas abouti).

Il est également important de vérifier que le cycle de vie des données détaillé au point 1.6 est bien pris en compte, qu'elles soient créées spécifiquement pour le système d'IA ou réutilisées par ce système : spécification, collecte/acquisition, vérification, pré-traitement.

Cela implique de connaître précisément le cycle de vie des données d'entraînement et de ré-entraînement. Dans le cas d'un apprentissage continu, les nouvelles données devront être prises en compte.

2.1.3 Règles d'allocation

Selon les recommandations du RCP Services numériques **:**

- **l'approche équipement** sera utilisée pour le **périmètre maîtrisé** (centre de données/serveur et terminaux en général) : le service numérique est considéré comme une somme d'usages de chaque équipement, chaque usage étant défini à travers une règle d'allocation par rapport aux impacts totaux de l'équipement ;

- et l'**approche système** pour le **périmètre non maîtrisé** (réseau) : les équipements sont regroupés en un système physique ou virtuel au niveau duquel les impacts environnementaux ont été déterminés.

L'**allocation des phases de fabrication, distribution, installation et fin de vie** se feront, comme recommandé par le RCP Services numériques, par ordre de priorité :

- sur un **critère physique**, par exemple volume de données consommées sur volume total de données ;
- sur un **critère de temps**, par exemple durée d'utilisation sur durée de vie ;
- sur d'**autres critères** si nécessaires.

L'**allocation de la phase d'usage** se fera, comme recommandé par le RCP Services numériques, en priorité sur un critère physique de **calcul de la consommation** d'énergie et autres consommations et émissions directes du service numérique.

Allocation selon le tiers considéré

Tiers	Phases du cycle de vie de l'équipement	Type d'allocation	Calcul de l'allocation
Terminaux utilisateurs	Toutes	Temporelle	Durée d'utilisation * taux d'utilisation [13)] / (durée d'utilisation totale du terminal [14)])
Réseau	Toutes	Physique	Volume de données transmises / volume de données total sur la durée de vie
Centre de données, serveurs	Toutes	Temporelle	Durée d'utilisation * taux d'utilisation / durée de vie
	Usage	Physique pour les serveurs	Allocation à 100 % de la consommation dynamique due à l'exécution des programmes, ainsi que la consommation statique
		Physique pour les équipements supports	Allocation descendante (*top*-down) *via* un facteur correctif permettant de calculer le surcoût dû à la consommation des autres équipements du centre de données (par défaut = PUE, *Power Usage Effectiveness* [15)])

13) Le taux d'utilisation correspond au pourcentage du terminal ou du serveur utilisé par le service d'IA. Par exemple, si on utilise un smartphone pour faire des requêtes vers un grand modèle de langage et qu'à côté on a d'autres applications ouvertes, le service d'IA ne consomme qu'un faible pourcentage de la consommation totale du smartphone. Si on accède à la mesure, on peut par exemple considérer un taux d'utilisation de 0,2. Par défaut, si la mesure n'est pas possible, le taux d'utilisation est fixé à 1.

14) La durée d'utilisation totale correspond à la durée pendant laquelle le terminal ou serveur est utilisé activement pendant sa durée de vie (en prenant en compte ses durées de vie en amont et en aval dans le cas d'équipements réutilisés). Par exemple s'il est utilisé 7 h par jour pendant 5 ans, sa durée d'utilisation totale sera de : 7 * 365 * 5 h.

15) L'objectif est de calculer l'empreinte environnementale due aux équipements du centre de données autres que les serveurs, et notamment la climatisation. Pour cela, nous proposons d'utiliser un facteur correctif correspondant à multiplier la consommation des serveurs par un nombre > 1. Par défaut ce facteur peut être le PUE, mais le PUE a pour dénominateur la consommation totale des équipements informatiques et non leur consommation dynamique, donc si le ratio consommation totale/consommation dynamique des équipements informatiques est connu, il est préférable d'utiliser ce dernier.

L'intégration des sous-modèles se fera *via* une clé d'allocation au *prorata* du nombre d'inférences (avec la meilleure estimation possible du nombre total d'inférences sur la durée de vie du modèle, sur le mode de choix a priori de la durée de vie en ACV).

L'utilisation de données existantes s'opérera *via* une clé d'allocation au *prorata* d'une estimation du nombre de téléchargements (avec la meilleure estimation du nombre total de téléchargements sur la durée de vie de la base de données).

2.1.4 Collecte des données d'inventaire

Dans la mesure du possible, des données primaires seront utilisées.

Ainsi, la mesure de la consommation électrique réelle d'un programme sera privilégiée à l'utilisation d'une valeur fixée (TDP [16]) par exemple). Cette mesure pourra être menée à l'aide d'infrastructures de mesures matérielles (wattmètres, PDU intelligents...) ou de wattmètres logiciels [17]). Si les mesures effectuées ne permettent pas de prendre en compte les scénarios d'apprentissage et d'inférence réels (en nombre d'équipements ou d'ensembles complets de données par exemple), des méthodologies de projection de consommation doivent être décrites et utilisées.

La prise en compte de données d'inventaire concernant les infrastructures matérielles (serveurs, cartes graphiques, équipements réseaux) doit être menée avec soin. Si les données les plus proches concernent des équipements génériques, les incertitudes doivent être indiquées. Comme dans toute ACV, la qualité des données doit être explicitement mentionnée en termes de représentativité géographique, technologique et temporelle.

Lorsque l'accès à des données primaires n'est pas possible, par exemple pour des raisons de droits d'accès aux fichiers nécessaires sur des infrastructures partagées, des données secondaires comme le TDP pourront être utilisées.

Dans le cas où l'accès aux données (car manquantes ou réservées par le propriétaire) est rendu complexe, il faut accepter de travailler avec la meilleure estimation possible en déclarant l'incertitude.

2.1.5 Choix des méthodologies de calcul d'impact environnemental

Le calcul d'impact dépend fortement du cadre méthodologique sous-jacent. La référence ici est le cadre du RCP services numériques de l'Ademe [2] auquel on ajoute la composante de « l'usage de l'eau » et la prise en compte de la consommation électrique, particulièrement notable pour l'entraînement et l'inférence. Les facteurs d'impact environnemental devront être justifiés. En particulier, pour les facteurs d'émissions liés à la consommation d'électricité, les facteurs de localisation ou de multi-localisation (régions géographiques de localisation du service et de ses composants) seront privilégiés aux facteurs *market-based* (du fournisseur électrique du service).

---

16) Le Thermal Design Power (TDP ou Puissance de dissipation thermique) représente la puissance maximale que le système de refroidissement doit être capable de gérer pour maintenir un composant électronique (processeur, carte graphique…) à une température opérationnelle adéquate pendant son utilisation maximale prévue.

17) Mathilde Jay, Vladimir Ostapenco, Laurent Lefèvre, Denis Trystram, Anne-Cécile Orgerie, and Benjamin Fichel « An experimental comparison of software-based power meters: focus on CPU and GPU". CCGrid 2023 - 23rd IEEE/ACM international symposium on cluster, cloud and internet computing, May 2023, Bangalore, India. https://inria.hal.science/hal-04030223v2

2.1.6 Indicateurs environnementaux

Les indicateurs environnementaux EF 3.0 retenus par le RCP Services numériques [2] sont les suivants :

- Épuisement des ressources naturelles ;
- Changement climatique ;
- Acidification ;
- Émissions de particules fines ;
- Rayonnements ionisants.

Le RCP services numériques propose également l'indicateur de flux « consommation d'énergie primaire ». Il est également pertinent de considérer la consommation électrique (en énergie finale) notamment pour la phase d'usage.

En pratique, plusieurs facteurs peuvent rendre le calcul de certains indicateurs infaisable en un temps raisonnable ou non pertinent : accès à la donnée, qualité de la donnée (incertitude, obsolescence). Nous proposons une priorisation des indicateurs environnementaux à renseigner (détails et unités décrits dans le tableau ci-dessous).

Priorité Haute

- Changement climatique
- Épuisement des ressources naturelles minéraux et métaux (ADP)

Priorité Moyenne

- Consommation et prélèvement de la ressource en eau
- Acidification des océans
- Émissions de particules fines
- Rayonnements ionisants

Priorité Basse

- Le reste (Toxicité humaine liée au cancer, Toxicité humaine non liée au cancer et Écotoxicité aquatique)

Les priorités ont été choisies comme suit :

- Les indicateurs du RCP pour lesquels il existe des données publiques sont indiqués en priorité haute, auxquels nous proposons d'associer les consommations d'énergie finale et primaire ;
- Les autres indicateurs du RCP sont en priorité moyenne avec les indicateurs sur l'eau que nous commentons par la suite ;
- Les indicateurs non retenus comme obligatoires mais recommandés par le RCP sont en priorité plus basse.

Les indicateurs sur l'eau ont été ajoutés car des études ont récemment mis en avant des prélèvements et consommations potentiellement importants liés à l'IA. Les prélèvements correspondent à la quantité d'eau prélevée dans le milieu naturel puis rejetée après utilisation. La consommation d'eau (ou prélèvements nets) correspond à la partie de l'eau prélevée et non restituée aux milieux aquatiques. Les valeurs rapportées devraient, autant que possible, prendre en compte tout le cycle de vie des équipements informatiques et de l'infrastructure numérique. A minima, nous recommandons de calculer :

- Les prélèvements ou consommations de la phase d'usage, sur site. Il s'agit de la part de prélèvement et consommation d'eau du centre de données qui peut être attribuée au service d'IA. Le calcul repose sur le produit de l'intensité eau [18)], mensuelle du centre de données, et la consommation électrique des serveurs utilisés pendant ce même mois par le service.

- Si les données sont disponibles, les calculs précédents seront aussi effectués pour les prélèvements d'eau en indiquant leur provenance (par exemple, eaux pluviales, eaux grises, eaux superficielles, eaux souterraines).

- Les consommations d'eau de la phase d'usage, hors site. Il s'agit de l'eau consommée pour la production de l'énergie nécessaire au service. Cette consommation est le produit de l'intensité moyenne en eau nationale pour la production d'électricité [19)] et la consommation d'électricité des serveurs et équipements supports du service.

18) Par défaut le WUE (Water Usage Effectiveness), en L/kWh.

19) Il convient d'utiliser l'intensité eau du pays où est consommée l'énergie pour le fonctionnement du service. Pour la France, l'intensité eau peut être trouvée dans les documents d'enregistrement universel d'EDF (exemple pour 2022).

Tableau détaillé des indicateurs environnementaux (priorité 1-haute, 2-moyenne et 3-basse)

Catégorie d'impact de l'EF	Indicateur de catégorie d'impact	Unité	Modèle de caractérisation	Priorité
Changement climatique, total [20)]	Forçage radiatif exprimé en potentiel de réchauffement planétaire (PRP100)	kg CO_2eq.	Modèle de base sur 100 ans élaboré par le GIEC (basé sur GIEC 2013)	1
Épuisement des ressources [21)], minéraux et métaux	Épuisement des ressources abiotiques (dernières réserves ADP)	kg Sb eq.	CML 2002 (Guinée et al., 2002) et van Oers et al., 2002.	1
Consommation d'énergie primaire	Indicateur non ACV	kWh	Indicateur non ACV	1
Consommation d'énergie finale (consommation électrique)	Indicateur non ACV	kWh	Indicateur non ACV	1
Particules	Impact sur la santé humaine	Incidence des maladies	Méthode PM recommandée par PNUE (PNUE 2016)	2
Épuisement de la ressource en eau	Potentiel de privation d'eau de l'utilisateur (consommation d'eau pondérée en fonction de la privation)	m^3 world eq.	Available Water Remaining (AWARE) (eau disponible restante) tel que recommandé par PNUE, 2016	2
Rayonnement ionisant, santé humaine	Efficacité de l'exposition humaine par rapport à U^{235}	kBq U^{235}eq.	Modèle d'effets sur la santé humaine tel que développé par Dreicer et al., 1995 (Frischknecht et al. 2000)	2
Acidification	Accumulation d'excédents (AE)	Mol H^+eq.	Accumulation d'excédents (Seppälä et al., 2006 ; Posch et al., 2008)	2
Toxicité humaine	Unité toxique comparative pour les êtres humains (CTUh)	CTUh	Modèle USEtox 2.1 (Fankte et al. 2017)	3
Écotoxicité – eaux douces	Unité toxique comparative pour les écosystèmes (CTUe)	CTUe	Modèle USEtox 2.1 (Frakte et al., 2017)	3

20) L'indicateur « Changement climatique, total » est composé de trois sous-indicateurs : changement climatique, origine fossile ; changement climatique, origine biologique ; changement climatique, utilisation des terres et changement d'affectation des terres. Les sous-indicateurs sont décrits plus en détail dans la section 4.4.10 de la méthodologie EF. Les sous-catégories « Changement climatique, origine fossile », « Changement climatique, origine biologique » et « Changement climatique, utilisation des terres et changement d'affectation des terres » doivent être déclarées séparément si elles contribuent chacune à plus de 5 % de la note totale de changement climatique.

21) Les résultats de cette catégorie d'impact doivent être interprétés avec précaution, car les résultats ADP suite à la normalisation peuvent être surestimés. La Commission européenne entend élaborer une nouvelle méthode en passant d'un modèle fondé sur l'épuisement à un modèle fondé sur la dissipation, afin de mieux quantifier le potentiel de conservation des ressources.

2.1.7 Outils et bases de données pouvant être utilisés

À ce jour, il n'existe pas d'outil unique, complet et largement utilisé pour mesurer l'impact environnemental de l'IA. La plupart des outils sont focalisés sur la consommation d'électricité liée à la phase d'usage du matériel. On peut distinguer plusieurs catégories d'outils : les outils matériels tels que les wattmètres, les interfaces de management intégrées dans le composant comme NVIDIA smi et RAPL, les logiciels et librairies, les calculettes carbone des fournisseurs de service cloud et les pages web d'estimation d'impact.

Les critères majeurs d'analyse et de choix des outils à prendre en compte comportent :

- le mode de fonctionnement : estimation ou mesure ;
- la facilité d'utilisation : nécessité d'installation, capacité de configuration, accessibilité des données ;
- le périmètre couvert et le niveau de détail : composant, machine, système, service ;
- les limites : incertitude, explicabilité, indicateurs supportés ;
- la surcharge de consommation générée par l'exploitation de l'outil.

Quelques exemples sont repris en Annexe 1.

Les facteurs d'impact permettent d'exprimer l'importance relative des émissions (ou des extractions) liées à un équipement ou un composant pour une catégorie d'impact environnemental spécifique. Ils permettent donc des analyses de cycle de vie complètes et multi-critères. Les impacts peuvent ensuite être rapportés avec plusieurs stratégies d'allocation.

Les critères de sélection des différentes bases de facteurs d'impact comportent notamment les éléments suivants :

- Les périmètres couverts par le référentiel :
  - sectoriel ou générique
  - mono ou multi facteur
  - couverture géographique
- La transparence sur les modes de calculs, les incertitudes, et les processus de validation ;
- L'accessibilité des informations (gratuites ou payantes).

Quelques exemples sont repris en Annexe 1.

2.2 Impacts indirects liés à l'utilisation du service (deuxième ordre et ordres supérieurs)

Cette partie vise à estimer les effets potentiels liés aux modifications comportementales, économiques, ou sociétales induites par l'exploitation du service d'IA proposé. Les effets potentiels étant beaucoup plus difficiles à quantifier que ceux dus au cycle de vie des équipements, déterminer qualitativement les effets potentiels pourra constituer une première évaluation (correspondant à une évaluation de type brique 3 – ou *tier 3* – dans la recommandation ITU-T L.1480).

Évaluer les impacts environnementaux indirects du service repose sur l'idée d'une comparaison entre un **scénario de référence** qui ne fait pas intervenir l'IA et un scénario dans lequel l'IA est utilisée.

La première étape consistera donc à décrire brièvement ces scénarios. Comme indiqué dans la recommandation ITU-T L .1480 [6], le ou les scénarios de référence doivent représenter des alternatives crédibles à l'introduction du service d'IA. Par exemple, l'évaluation d'un service d'IA d'aide à la conduite ne peut pas prendre comme référence un véhicule des années 1980 mais doit considérer un véhicule moyen au moment du déploiement du service, qui comprend déjà a priori un système d'aide à la conduite.

La comparaison des indicateurs environnementaux de la mise en place du service devra être réalisée sur une période temporelle cohérente avec la durée des effets potentiels induits par l'apparition du système d'IA. Ainsi, les scénarios ne doivent pas s'arrêter à quelques mois pour des effets attendus dans plusieurs années. Le scénario de référence (c'est-à-dire sans l'apparition du système d'IA) devra s'appuyer sur des hypothèses de cadrage en phase avec une trajectoire respectant les Accords de Paris.

Il s'agira ensuite de dresser une **liste qualitative des effets potentiels** que l'on peut attendre de la proposition : effets d'obsolescence, effets rebond directs ou indirects, transformations sociétales, etc. Ici le terme « effets potentiels » de la proposition est à prendre dans un sens large : ils peuvent être induits par la proposition elle-même ou être en lien avec la ou les technologies dans lesquelles elle s'inscrit. Cette liste peut s'appuyer sur un **arbre de conséquences** (voir Figure 5 pour un modèle d'arbre de conséquences issu de la recommandation ITU-T L.1480 [6]).

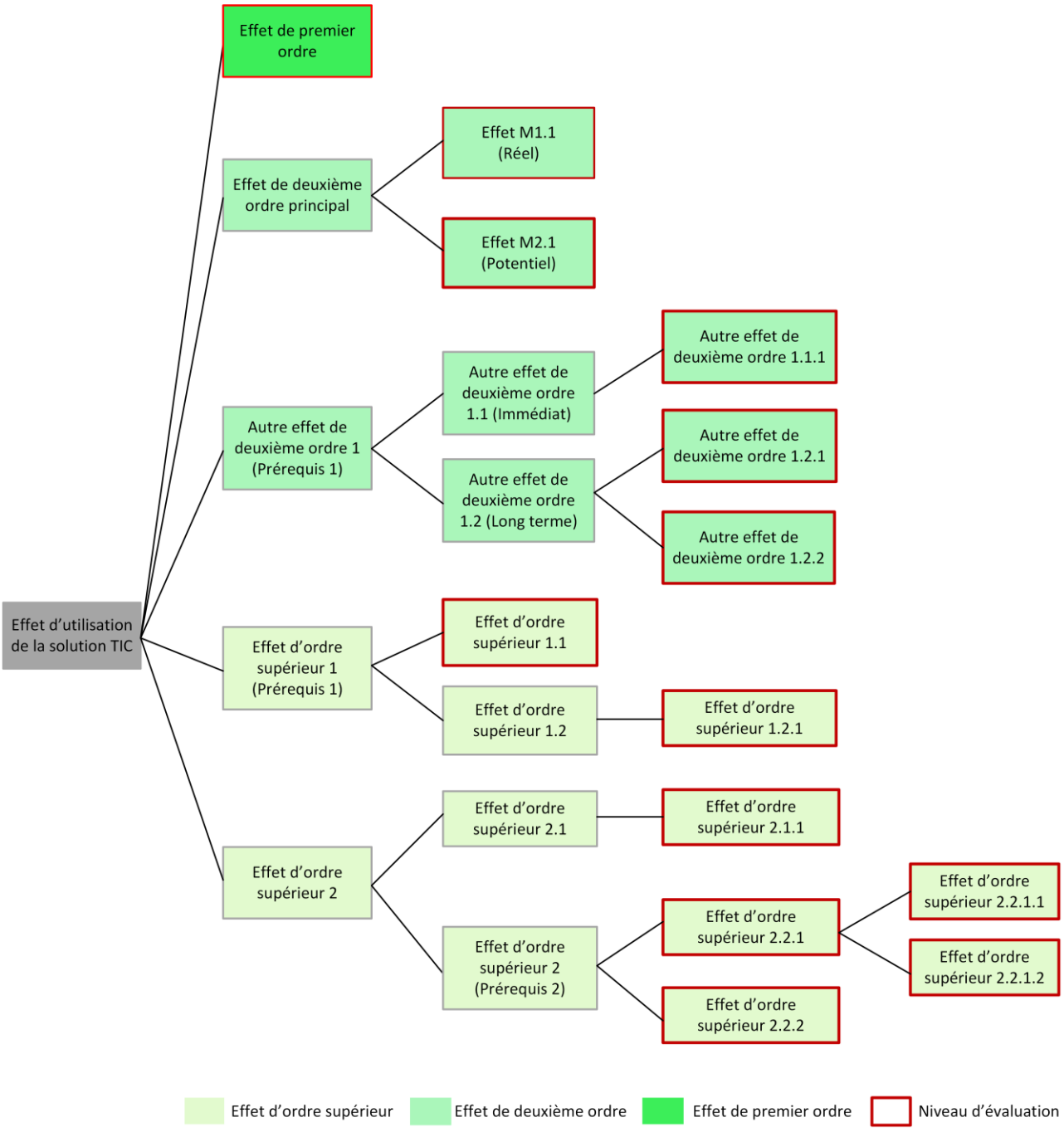

Figure 5 — Arbre de conséquences où l'utilisation du service d'IA entraine sur chaque effet une action ou événement qui a une incidence sur les impacts environnementaux

Pour chaque effet mentionné, il s'agira de préciser, le cas échéant, si une contre-mesure est prévue pour l'atténuer ou l'annuler, et indiquer l'efficacité qu'on peut attendre de cette contre-mesure. On parle de contre-mesure lorsque cette dernière agit directement sur l'effet en question. Ainsi, ne sont pas considérées comme des contre-mesures tout mécanisme de compensation environnementale.

Exemples d'effets potentiels (tableau issu du document ÉcoInfo [22]) :

Nature de l'effet	Description de l'effet	Types d'impacts environnentaux
Effet d'obsolescence	Renouvellement accéléré du parc de véhicules du fait de l'évolution très rapide de la performance des modes autonomes d'une génération à l'autre de véhicule	Liés à la fabrication des véhicules : ▪ changement climatique, ▪ épuisement des ressources…
Effets rebond direct	Augmentation des distances parcourues du fait de la baisse des coûts d'utilisation (« *platooning* »), de la valorisation des temps de trajet (grâce aux activités réalisées pendant les trajets : travail, repos, loisir…), de la plus grande accessibilité, de l'auto-parking…	Changement climatique (augmentation des émissions de gaz à effet de serre)
Effet rebond indirect	Les gains économiques dus au « *platooning* » sont dépensés par les usagers dans d'autres produits ou services à fort impact environnemental	Fonction des produits/services
Transformations sociétales	La mobilité induit des changements dans les habitudes de déplacement, de l'étalement urbain	Changement climatique (émissions de gaz à effet de serre), perte de biodiversité due à l'artificialisation des sols

2.3 Limites de la méthodologie proposée

En ce qui concerne l'évaluation des effets indirects de services numériques, il existe des méthodologies permettant d'évaluer l'impact net du déploiement de services numériques par rapport à un scénario tendanciel comme l'ITU-T L.1480 [6] ou EGDC (*European Green Digital Coalition*), mais celles-ci sont peu connues et utilisées. Elles sont aussi complexes à mettre en œuvre. C'est pourquoi nous avons proposé de s'appuyer sur le niveau le plus souple de la recommandation ITU-T L. 1480 [6]. Des initiatives françaises émergent sur le sujet, comme l'étude de l'Ademe sur des cas d'usages de solutions numériques pour éviter les impacts environnementaux d'autres secteurs, lancée début 2024.

L'obtention des données des fournisseurs d'IA (données, produits et services d'IA, infrastructure) est une condition nécessaire à la mise en œuvre de cette spécification et est le premier enjeu. Cela recouvre notamment les données d'impact des modèles propriétaires, celles des infrastructures cloud et celles de fabrication des processeurs utilisés.

22) https://hal.archives-ouvertes.fr/hal-03853135

3 Bonnes pratiques

L'objectif de ce chapitre est de proposer un état des lieux de bonnes pratiques (notées BP ci-après) **opérationnelles** en IA frugale permettant aux organisations intéressées par ce sujet de l'adopter dans les meilleures conditions.

Ce premier recueil de bonnes pratiques en IA frugale, réalisé par 83 experts réunis dans le cadre de l'AFNOR SPEC IA frugale peut donc s'apparenter à un catalogue de méthodes et pratiques à appliquer pour tendre vers l'IA frugale.

Ce premier recueil ne se veut donc pas exhaustif, la maturité de ce domaine étant encore en construction.

Les éléments détaillés dans ce chapitre sont en cohérence avec les définitions, indicateurs et méthodes de calcul abordés dans les chapitres précédents.

3.1 Approche méthodologique

3.1.1 Approche globale la plus neutre possible

Pour ce recueil de bonnes pratiques opérationnelles en IA frugale, deux défis majeurs sont apparus :

- Comment sélectionner les bonnes pratiques en IA frugale alors que les indicateurs permettant de classer ces bonnes pratiques ne sont pas encore définis ? Et encore moins recueillis ?
- Comment assurer le caractère « applicable » et « opérationnel » de ces bonnes pratiques ? Au-delà des contraintes relatives aux fournisseurs et clients d'IA, c'est en effet ce caractère « opérationnel » qui sera le facteur crucial du succès de la diffusion de ces bonnes pratiques.

Comment pallier l'absence d'indicateurs ?

Pour répondre à l'absence d'indicateurs de mesure de l'IA frugale, le groupe de travail s'est appuyé sur l'expertise et le nombre important des acteurs du groupe de travail, à la fois issus du monde de l'entreprise, du secteur public et du monde académique :

- 83 participants
- Issus de 60 organismes
  - 47 acteurs du secteur privé
    - 25 entreprises
    - 22 fournisseurs/startups IA
  - 7 acteurs du secteur public
  - 6 acteurs du monde académique et de la recherche

Comment garantir le caractère opérationnel des bonnes pratiques ?

Pour garantir le caractère opérationnel des bonnes pratiques, le groupe de travail s'est appuyé sur la volonté des participants de témoigner et de partager leurs bonnes pratiques pour diffuser l'IA frugale.

La démarche s'est déroulée de la manière suivante :

1. Recueillir les bonnes pratiques issues de projets mettant en avant l'IA frugale par les 83 participants au groupe de travail ;
2. Compléter ces bonnes pratiques expérimentées sur le terrain par les organisations par des bonnes pratiques présentes dans la littérature scientifique ;
3. Organiser ces bonnes pratiques selon :
    a. Les étapes du cycle de vie identifiées dans la section 1.5. Une catégorie, permettant de mettre en valeur les bonnes pratiques en termes de gouvernance et d'acculturation, a été ajoutée, ces deux domaines apparaissant comme fondamentaux aux participants du groupe de travail ;
    b. Les domaines d'études choisis pour un service d'IA, c'est-à-dire le service lui-même, les données et les infrastructures (*i.e.* les ressources matérielles).
4. Sélectionner les « meilleures » bonnes pratiques, en s'appuyant sur l'expertise de l'ensemble des membres du groupe de travail, tous experts de l'IA frugale en termes de gain en frugalité et d'effort de mise en œuvre :
    a. Notation des participants sur les bonnes pratiques connues par eux-mêmes (principe d'une voix par personne) sur les critères d'impact sur la frugalité et de difficulté de mise en œuvre ;
    b. Finalisation via une table ronde d'experts.

3.1.2 Limites de l'approche

L'approche adoptée pour recueillir les bonnes pratiques et les synthétiser comprend des limites que nous exposons ci-dessous :

- Absence d'indicateurs et de mesures : approche par dires d'experts des organisations du groupe de travail ;
- Représentativité des organisations constituant le groupe de travail pouvant constituer un biais ;
- Faible volume de bonnes pratiques issues de la bibliographie ;
- Domaine de l'IA frugale en forte évolution : de nouvelles bonnes pratiques liées aux évolutions technologiques apparaîtront régulièrement ;
- Maturité variée des projets à partir desquels les bonnes pratiques ont été recueillies ;
- Absence de prise en compte des effets rebond.

3.2 Description des bonnes pratiques

3.2.1 Synthèse des bonnes pratiques

Issues très majoritairement des témoignages et entretiens, 31 bonnes pratiques ont été formulées par le groupe de travail et classées selon les trois segments (service, données, infrastructures) et les étapes du cycle de vie d'une IA.

Les bonnes pratiques adressent souvent plusieurs étapes et plusieurs domaines et couvrent de manière relativement équilibrée l'ensemble des sujets à traiter. Les phases initiales (en y incluant la gouvernance, transverse) représentent un poids important dans ces bonnes pratiques, traduisant que la conception et la gouvernance ont un impact important sur l'ensemble du cycle de vie d'un projet d'IA.

Nombre de bonnes pratiques selon les domaines et les étapes du cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures	Total
0 – Transverse		9		9
1 – Initialisation	4	1	5	10
2 – Conception et Développement	10	10	6	26
3 – Vérification et validation	2	3	2	7
4 – Déploiement	2	1	3	6
5 – Exploitation et suivi	1	4	2	7
6 – Validation continue	3	4	1	8
7 – Réévaluation	4	6	1	11
8 – Mise hors service		3	4	7
Total	35	41	33	

Certaines bonnes pratiques adressant plusieurs étapes ou plusieurs domaines, le total par étape ou par domaine est donc parfois supérieur au nombre total de bonnes pratiques.

Vous pouvez accéder aux fiches explicatives des bonnes pratiques en cliquant dessus.

Répartition des bonnes pratiques selon les domaines et les étapes du cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 - Transverse	BP11 ; BP12 ; BP13 ; BP14 ; BP16 ; BP17 ; BP18 ; BP21 ; BP23		
1 - Initialisation	BP01 ; BP02 BP22 ; BP29	BP01	BP01 ; BP02 BP05 ; BP20 BP22
2 - Conception et Développement	BP01 ; BP03 BP09 ; BP22 BP23 ; BP26 BP27 ; BP28 BP29 ; BP30 BP31	BP01 ; BP06 BP07 ; BP08 BP09 ; BP10 BP15 ; BP19 BP23 ; BP31	BP01 ; BP05 BP06 ; BP20 BP22 ; BP23
3 - Vérification et validation	BP09 ; BP23	BP09 ; BP19 BP23	BP05 ; BP23
4 - Déploiement	BP09 ; BP22	BP09	BP05 ; BP20 BP22
5 - Exploitation et suivi	BP09	BP06 ; BP09 BP24 ; BP25	BP05 ; BP06
6 - Validation continue	BP09 ; BP29 BP31	BP09 ; BP24 BP25 ; BP31	BP05
7 - Réévaluation	BP04 ; BP09 BP29 ; BP31	BP04 ; BP07 BP09 ; BP24 BP25 ; BP31	BP05
8 - Mise hors service	-	BP06 ; BP10 BP15	BP05 ; BP06 BP20 ; BP22

3.2.2 Notation des bonnes pratiques

L'ensemble des bonnes pratiques a été soumis à une notation de la part des contributeurs à l'AFNOR SPEC. Ce panel de contributeurs est composé de : Entreprise (30 %) ; Fournisseur/startup IA (24 %) ; Organisme académique ou de recherche (15 %) ; Organisme public (9 %) ; Association (6 %) ; Syndicat professionnel (3 %) ; *Think Tank* (3 %) ; Autre (9 %).

Pour chaque bonne pratique, les votants ont estimé le gain de frugalité découlant de la bonne pratique (faible, modéré ou élevé) et l'effort de mise en œuvre de la bonne pratique (faible, modéré ou élevé). Chaque votant a également pu indiquer ses 5 bonnes pratiques préférées.

Ces trois classements sont consultables ci-dessous.

Bonnes pratiques avec le gain le plus fort

1. BP12 – Instruire la frugalité dans chaque projet IA
2. BP14 – Acculturer et former les parties prenantes
3. BP02 – Choisir la solution pour répondre au besoin en considérant les alternatives à l'IA
4. BP20 – Optimiser l'usage de l'équipement existant
5. BP01 – Utiliser des méthodes d'analyse de besoin pour mettre en œuvre la frugalité

Bonnes pratiques les plus faciles à mettre en œuvre

1. BP29 – Réutiliser des algorithmes entraînés et partager les algorithmes réalisés
2. BP04 – Définir des critères justifiant le ré-entraînement du modèle
3. BP15 – Faire de la compression de données
4. BP23 – Réaliser une estimation de la consommation via l'apprentissage sur une petite partie du jeu de données
5. BP19 – Utiliser des *datasets* open source pour la phase de prototypage

Bonnes pratiques les plus populaires (parmi les contributeurs à l'AFNOR SPEC)

1. BP02 – Choisir la solution pour répondre au besoin en considérant les alternatives à l'IA
2. BP14 – Acculturer et former les parties prenantes
3. BP01 – Utiliser des méthodes d'analyse de besoin pour mettre en œuvre la frugalité
4. BP29 – Réutiliser des algorithmes entraînés et partager les algorithmes réalisés
5. BP12 – Instruire la frugalité dans chaque projet IA

Le tableau ci-dessous présente une classification de l'ensemble des bonnes pratiques selon leurs notations.

Répartition des bonnes pratiques selon leur gain estimé et l'effort de mise en œuvre

Top 10 en Gain (BP avec le gain le plus fort)	BP12 BP13	BP02 BP09 BP14	BP01 BP11 BP20 BP29 BP30
BP classées de 11 à 20 en Gain	BP07 BP21 BP24	BP03 BP05 BP06 BP16 BP27 BP31	BP04
BP classées de 21 à 31 en Gain (BP avec le gain le plus faible)	BP17 BP18 BP22 BP25 BP26 BP28	BP08	BP10 BP15 BP19 BP23
Gain / Effort	BP classées de 21 à 31 en Effort (BP avec l'effort le plus fort)	BP classées de 11 à 20 en Effort	Top 10 en Effort (BP avec effort le plus faible)

Ainsi les bonnes pratiques situées dans la cellule en haut à droite représentent le meilleur ratio avec un faible effort de mise en œuvre et un gain élevé en frugalité. Il est donc recommandé de débuter par l'application de ce groupement de bonnes pratiques.

Les bonnes pratiques situées dans la cellule en bas à gauche représentent le moins bon ratio avec un effort élevé de mise en œuvre et un faible gain en frugalité. Il est donc recommandé d'appliquer ce groupement de bonnes pratiques en dernier lieu, après avoir mis en œuvre les bonnes pratiques indiquées dans les autres cellules.

3.2.3 Déroulé d'un projet d'IA frugale

Étapes clés des projets d'IA frugale

La construction, le déploiement et la production de services et systèmes d'IA frugaux nécessitent l'adaptation des processus de création et de gestion de projets en vue de :

- permettre et faciliter l'identification, l'acceptation, la construction et le déploiement de solutions IA grâce aux principes de sobriété et de frugalité d'une part ;
- intégrer l'évaluation et le suivi des impacts environnementaux d'autre part.

De fait, cela peut nécessiter le changement ou l'évolution de certaines habitudes, postures ou injonctions.

Ce besoin de transformation des processus provient d'un double constat :

- de la manière d'exprimer le besoin dépendent les solutions : pouvoir envisager l'IA frugale comme réponse doit d'abord passer par une expression de besoins ouvrant le champ des solutions et la permettant (tout comme elle permet d'autres solutions que l'utilisation de services et de systèmes d'IA) ;
- le numérique ou toute technologie comme l'IA ne sont pas durables par essence, même s'ils peuvent outiller et faciliter la réduction des impacts. En effet, ils produisent eux-mêmes des impacts environnementaux qu'il faut considérer, et leur bénéfice pour tout programme à enjeu environnemental doit être évalué en prenant en compte chaque service d'IA et l'ensemble du système (les usages) pour juger de leur pertinence et en les comparant à d'autres solutions alternatives (gestes métiers, alternatives technologiques...).

Les points de méthodologie et d'attention à adopter sont les suivants [23] :

- De manière générale et en amont, **le besoin** doit être questionné pour s'assurer que :
  - Il existe réellement, il est utile et aucune solution existante (éventuellement adaptée) ne peut y répondre ;
  - Il est ouvert et **permet, dans son expression, d'envisager tout type de solution**, i.e. des solutions autres qu'une IA à haute performance (IA frugale et solutions IT plus classiques mais plus sobres sans IA, voir bonnes pratiques N°01 et 02) ;
  - **Les scénarios d'usages sont parfaitement décrits avec leurs limites**, ce qui permettra d'évaluer les impacts des différentes solutions et d'identifier, à défaut de pouvoir les quantifier, les effets d'ordre 2 ou supérieurs à surveiller dans la phase d'exploitation.

23) Ils sont donnés dans l'ordre d'un processus projet classique mais restent autonomes et adaptables en méthodologie agile et gestion de produit.

L'injonction de développer un service d'IA doit pouvoir être remise en cause. Conjointement à un questionnement de la juste performance, **l'objectif de l'empreinte environnementale la plus faible possible devant** par ailleurs **être ajouté à la liste des exigences fortes du projet**. Cet objectif doit ouvrir le champ des possibles pour les solutions acceptables et acceptées. À ce niveau, les ateliers et critères agiles du type « definition of ready [24)] » devront être adaptés pour intégrer ces caractéristiques de l'expression des besoins et des exigences projet.

Afin d'élargir le champ des solutions possibles, il est fortement recommandé d'ajouter, en phase de cadrage du besoin, une étape d'**identification des degrés de liberté les plus larges possibles** dans les paramètres ou les caractéristiques du projet permettant de moduler les exigences et donc d'avoir une approche de sobriété et de frugalité numérique, afin d'identifier et surtout d'accepter des solutions avec ces critères (voir bonne pratique N°02). Cela concerne par exemple la précision des modèles et leur performance (voir bonnes pratiques N°03, 04, 05, 07, 10, 15, 19, 26, 27, 28, 29, 30 et 31), le fait d'autoriser des fonctionnements en mode asynchrone (pour permettre par exemple des exécutions en temps décalé, lorsque les infrastructures GPU sont moins sollicitées, ce qui permet par exemple de dimensionner ces dernières non pas par rapport au besoin maximum mais au plus juste), etc.

Dans une approche agile, cela peut se traduire par l'ajout de critères de valeurs environnementales en complément des critères de valeur métier ou par une approche adaptée de questionnement et d'affinage des besoins autour des sujets permettant la sobriété et la frugalité (performance, traitements asynchrones, etc.). Les responsables des produits doivent identifier et challenger ces degrés de liberté pour les rendre plus nombreux et adaptables, pour proposer le plus de solutions possibles et les plus ambitieuses possibles en termes de sobriété/frugalité.

- L'analyse du besoin doit passer par l'identification de toutes les solutions possibles répondant au besoin, que ce soit les solutions basées sur l'IA mais aussi celles plus classiques sans IA ou à base d'IA frugale ou encore celles s'appuyant sur l'existant (en termes de service ou d'infrastructure, voir bonnes pratiques N°20 et 22). Pour cela, les degrés de liberté précédemment identifiés doivent être exploités au maximum. **Si des solutions frugales ne peuvent pas être identifiées**, par exemple parce que la précision demandée est trop élevée, **un retour aux premières étapes du projet et un questionnement du besoin doit être possible**, par exemple pour s'interroger sur le niveau de précision attendu. Si nécessaire, des études complémentaires doivent pouvoir être menées afin de s'assurer que l'ensemble des solutions ont été identifiées. Les impacts environnementaux de ces études, si elles s'appuient sur des tests, devront être intégrés au bilan global du projet.

- Chaque solution potentielle doit être évaluée en termes d'impacts environnementaux sur l'ensemble du cycle de vie du projet et de la solution selon les scénarios d'usages documentés dans l'expression de besoins et en tenant compte des effets de bord et effets rebond possibles : l'évaluation des effets de 1er et 2nd ordre est nécessaire mais pas suffisante. Des hypothèses et des mesures de contrôle doivent être identifiées et parfaitement documentées pour évaluer et garder les effets d'ordres supérieurs, notamment les effets rebond, sous contrôle (voir le chapitre précédent et les bonnes pratiques N°21, 23, 24 et 25). L'analyse doit également permettre de **dégager des conditions de pertinence des différentes solutions**. Ces conditions devront être surveillées en phase de production de la solution : si l'une de ces conditions vient à ne plus être vérifiée, une évaluation des conséquences de cette nouvelle situation devra être faite et une décision devra être prise.

24) LA « definition of ready » consiste, dans le cadre d'une approche agile de type Scrum, en la construction d'une liste de critères à valider pour que l'équipe de développement accepte de prendre en charge une user story (un élément fonctionnel à développer). Voir https://blog.octo.com/la-definition-of-ready-dor

Dans le cas des solutions basées sur ou faisant intervenir une IA, la question de la frugalité de celle-ci doit être posée sur l'ensemble du cycle de vie, sur chaque étape et de manière globale [25], y compris lors des activités de gouvernance ou des phases de R&D préalables (l'apprentissage mais pas seulement : l'ensemble des tests et tentatives pour identifier l'algorithme, etc.). En clair, cela doit se faire dans un cadre de gouvernance révisé intégrant la question de la frugalité (voir bonnes pratiques N°06, 08, 09, 11, 12, 17 et 18). Par ailleurs, le choix de la solution, qu'elle soit basée sur l'IA ou non, doit prendre en compte une évaluation de l'ensemble du système de matériel nécessaire pour fournir le service numérique : du choix d'un hébergeur écoresponsable à une optimisation du matériel existant via la mutualisation par exemple (voir bonnes pratiques N°20, 22, 24 et 25).

Si une solution intègre l'achat de matériel spécifique, comme des cartes GPU, l'évaluation des impacts doit intégrer la seconde vie ou la fin de vie de ces équipements (voir bonne pratique N°11) : le devenir de ces équipements après le projet doit être décrété dès le départ.

- Les processus de décision et de pilotage projet doivent intégrer une évaluation des impacts environnementaux (voir chapitre précédent). À défaut de mesure spécifique au système d'IA souhaitée, les principes standards de mesure et de vigilance applicables à tout système numérique s'appliquent (RCP service numériques [2], référentiels du type Référentiel Général de l'Écoconception des Services Numériques [10], etc.). L'instruction et les choix définitifs doivent être parfaitement documentés et partagés. Cela facilitera le suivi du projet et la capitalisation pour d'autres projets (voir bonnes pratiques N°11 et 12). On peut intégrer la notion de budget carbone aux processus décisionnels pour cadrer l'empreinte carbone maximale allouée aux projets, ce qui peut forcer à un certain niveau de sobriété, voire permettre d'orienter l'empreinte globale à la baisse et à la rationalisation en demandant par exemple que, pour tout nouveau projet, une application existante d'impact comparable soit décommissionnée en parallèle. Le budget carbone devra au maximum être cohérent avec une trajectoire respectant les Accords de Paris.

- Le déploiement de la solution doit être outillé pour pouvoir surveiller et mesurer précisément son utilisation réelle (voir chapitre précédent) et la comparer avec les conditions de pertinence, les scénarios d'usage de l'expression de besoins et les prévisions réalisées lors des étapes précédentes du projet. Si une condition de pertinence vient à ne plus être vérifiée ou en cas de dérive par rapport à l'attendu (par exemple le scénario d'usage ne s'applique pas en raison d'un usage imprévu), la situation et la projection dans le temps des impacts devront être réévaluées et des décisions, potentiellement de révision profonde voire de décommissionnement, devront être prises.

- En parallèle, les rôles et responsabilités des différentes parties prenantes des projets doivent être réadaptés pour tenir compte de ce qui précède et intégrés dans les pratiques courantes de l'entreprise. Par ailleurs, dans une optique d'industrialisation de la nouvelle approche projet, il est possible de construire et de déployer un modèle de maturité permettant d'évaluer le niveau de prise en compte des problématiques liées à l'IA frugale. Au final, c'est toute la gouvernance projet qui doit être adaptée pour intégrer la frugalité (par exemple en questionnant l'élargissement du besoin et la profondeur des cas d'usages pour détecter les effets rebond et autre effets indirects), la mesure de l'empreinte environnementale du projet et les processus de décision tant en cours de projet (argumentation et documentation des décisions, nécessité d'une analyse mesurée au préalable, etc.) qu'en phase d'opération pour s'assurer qu'ils se feront en toute connaissance de cause.

25) En effet, une somme d'optimum locaux ne faisant pas nécessairement un optimum global, il est nécessaire de se poser la question de la frugalité dans l'ensemble, d'où le questionnement initial du besoin.

Principaux facteurs clés de succès

Chaque bonne pratique possède ses propres facteurs clés de succès. L'adaptation et le succès de la méthodologie projet adaptée telle que décrite précédemment supposent toutefois plusieurs conditions de succès propres :

- Mettre en place une gouvernance permettant de piloter et maîtriser les impacts environnementaux de l'IA, en les intégrant dans les processus d'organisation générale, notamment les prises de décision, le modèle de pilotage des projets ;
- Anticiper suffisamment pour pouvoir valider le besoin et réaliser les différentes évaluations ;
- Garantir une posture d'ouverture des parties prenantes : acceptation des solutions alternatives et renoncement si nécessaire ;
- Accepter d'ajuster les ambitions de la finalité visée notamment en termes de performance en vue d'optimiser l'équilibre économico-socio-environnemental ;
- Animer et faciliter la création/participation à une communauté de pratiques ou de partage existante sur le thème de l'IA frugale basée sur les compétences ou non. Cela permettra de partager et de capitaliser sur les retours d'expériences et d'autres bonnes pratiques ;
- Organiser la formation et l'acculturation aux enjeux environnementaux de l'ensemble des acteurs impliqués dans le projet (voir bonnes pratiques N°14 et 16), en commençant par les acteurs clés (chefs de projet IA, architectes d'entreprises, data scientists, etc.) y compris au niveau des instances décisionnaires. Il est important que les acteurs connaissent et puissent proposer et qualifier les services et systèmes IA au regard de leur frugalité. Aussi, cela peut donner lieu à la formalisation et à la valorisation de compétences spécifiques via un programme RH de montée en compétence par exemple.

Le passage à l'échelle de l'IA grâce à un déploiement autour de 7 thématiques

Les différentes pratiques analysées ont permis de construire 7 thématiques structurelles pour faciliter le déploiement à l'échelle des systèmes et services d'IA frugaux :

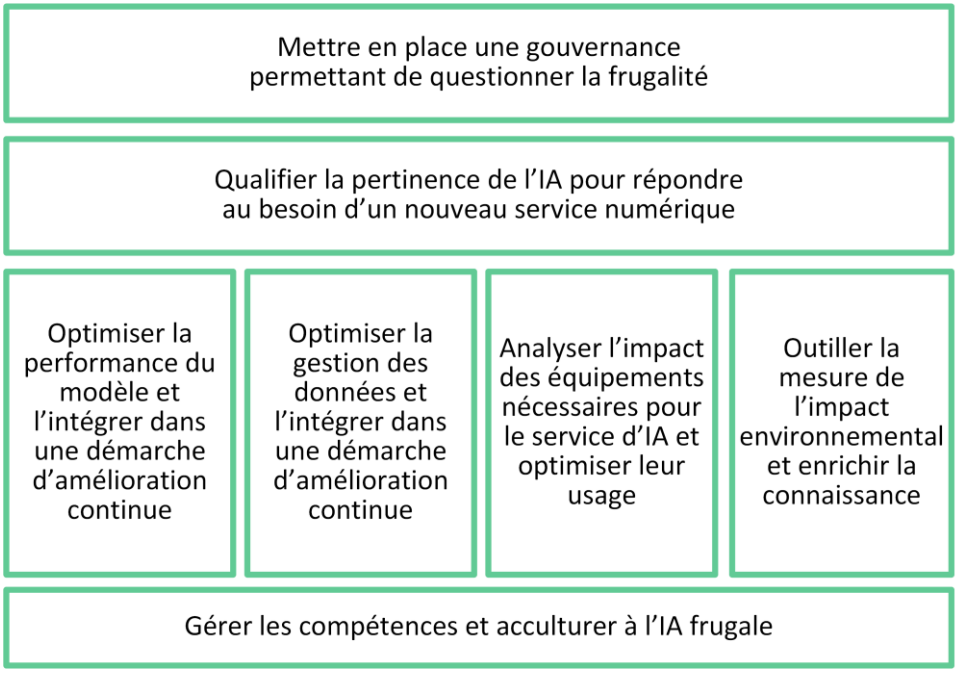

Figure 6 — Les 7 thématiques englobant les bonnes pratiques

3.3 Pour aller plus loin

Conscient des limites de ce recueil de bonnes pratiques, plusieurs actions d'accompagnement peuvent être entreprises pour stimuler l'adoption de l'IA frugale par la filière :

- Faciliter la prise de conscience en IA frugale :
  - Créer une grille d'auto-évaluation à disposition des organisations, accompagnée des avantages liés à l'IA frugale
- Créer une grille d'auto-évaluation de l'éligibilité des outils existants à la présente AFNOR SPEC
- Démocratiser les bonnes pratiques :
  - Organiser un cycle de témoignages sur ces bonnes pratiques
- Améliorer les bonnes pratiques :
  - Publier annuellement une mise à jour du recueil de bonnes pratiques. Cette publication annuelle sera ainsi l'occasion d'offrir :
    - Une meilleure identification des bonnes pratiques sur l'ensemble du cycle de vie, la récurrence réduisant ainsi le biais lié au déficit d'identification de ces bonnes pratiques potentielles ;
    - Une meilleure pertinence de ces bonnes pratiques, du fait du dynamisme du domaine de l'IA frugale.
  - Quantifier l'impact des bonnes pratiques en utilisant les indicateurs définis au Chapitre 2.

4 Communication

Dans le cas où un fournisseur d'IA ou un producteur d'IA cherche à publier des informations environnementales sur des systèmes ou des services d'IA, ce document propose quelques lignes directrices simples pour outiller les entreprises. Un rapport de durabilité établi dans le cadre de la directive CSRD pourra faire référence aux évaluations réalisées en se basant sur ce référentiel en précisant la date de publication de l'évaluation. Ces recommandations peuvent être utilisées en second lieu par des acteurs qui souhaitent évaluer la qualité des allégations environnementales d'autres acteurs.

4.1 Dans le cas d'une évaluation quantitative d'indicateurs environnementaux sur le cycle de vie

Le fournisseur ou le producteur d'IA devra préciser les informations suivantes pour faire des déclarations sur des indicateurs environnementaux sur le service :

- Précision du périmètre de l'analyse en cycle de vie : unité fonctionnelle, frontières du système, méthode d'allocation (par volume, par temps d'exécution, par nombre de requêtes...), catégories d'impact, précision des hypothèses [26)] adoptées dans le cas d'une incertitude (par exemple sur la localisation opérationnelle des ressources de calcul) ;
- Source des méthodologies utilisées (par exemple un référencement aux méthodes de calcul de cette AFNOR SPEC ou une méthodologie future qui reprend les mêmes objectifs) ;
- Base de données d'analyse en cycle de vie utilisée pour le calcul d'impact ;
- Guide d'utilisation du service d'IA pour obtenir la performance environnementale attendue sur les inférences (dans le cas où la phase d'utilisation fait partie des étapes du cycle de vie les plus importantes du système d'IA) ;
- Localisation des ressources de calcul et de stockage sur le cycle de vie de l'IA ;
- Précision sur l'évaluation : revue critique par un tiers extérieur ou non, données instantanées ou date de réalisation de l'analyse en cycle de vie, éléments sur la qualité des données à la base de l'ACV.

NOTE Au vu des avancées technologiques régulières dans le domaine de l'IA, il est recommandé de renouveler cette évaluation quantitative au moins une fois par an.

Les informations suivantes peuvent venir compléter les allégations environnementales ou être disponibles à la demande aux clients :

- Volume de données utilisées pour l'entraînement et le ré-entraînement ;
- Volume de transferts réseaux, y compris ceux liés aux sauvegardes et réplications ;
- Inventaire des flux/processus et description des calculs réalisés.

---

26) Les calculs se basent sur diverses hypothèses de volumétrie de données et/ou de nombre de réentraînement des modèles qui peuvent ne pas être vérifiées dans la pratique. Une partie de ces hypothèses peuvent faire partie des conditions d'usage ou du guide d'utilisateur du service d'IA, comme le nombre de réutilisations par exemple. Dans le cas idéal, il faudrait également indiquer la gouvernance ou a minima un canal d'actualisation des données et des mesures environnementales.

4.2 Pour communiquer sur le caractère frugal d'un service d'IA

Le fournisseur ou le producteur d'IA devra donner les informations suivantes :

- Une évaluation quantitative des indicateurs environnementaux sur le cycle de vie du service [27] (avec les précisions issues de la partie 4.1.) : les indicateurs prioritaires devront être reportés obligatoirement (Chapitre 2).
- Le détail de mise en œuvre des bonnes pratiques adoptées pour le service, avec un référencement de leurs sources (par exemple aux bonnes pratiques du Chapitre 3 de l'AFNOR SPEC) et une évaluation qualitative de l'impact de ces bonnes pratiques ;
- Une liste qualitative des effets négatifs potentiels de second ordre et d'ordre supérieur qu'on peut attendre du service d'IA (par exemple : éléments sur les effets anticipés d'obsolescence accélérée et de massification des usages) et des contre-mesures mises en place, le cas échéant.

4.3 Pour communiquer sur le bilan positif pour une catégorie d'impact d'un service frugal d'IA

Le fournisseur ou le producteur d'IA devra donner l'ensemble des informations suivantes :

- Toutes les informations relatives à la communication sur le caractère frugal du service d'IA (voir 4.2.) ;
- Une évaluation quantitative des impacts environnementaux des usages d'un service d'IA (à partir d'une méthodologie qui doit être précisée avec la source) et une comparaison avec les impacts environnementaux du cycle de vie du service. On parlera ici du premier ordre et des ordres supérieurs. Cette évaluation fera apparaître clairement les impacts négatifs en face des impacts positifs pour un même indicateur ;
- Les potentiels transferts d'impact du service (par exemple, un effort sur la consommation d'énergie liée au service peut entraîner une augmentation de la consommation en eau pour le refroidissement des serveurs...).

NOTE Les méthodes de compensation ne devront pas être prises en compte, de même que l'achat d'énergie bas carbone ou renouvelable. Une évaluation basée sur la localisation des centres de calcul est obligatoire (cf. GHG Protocol[28] ou Bilan GES[29]). La dissipation « intelligente » de chaleur fatale (recyclage pour le chauffage, par exemple) ne peut pas non plus être prise en compte.

---

27) Le service d'IA frugale comprend les composants IA et non IA du service numérique, il doit être absolument évalué sous les angles des algorithmes, des données et du matériel.

28) https://ghgprotocol.org/

29) https://bilans-ges.ademe.fr/ressources/points-cles-methodologiques

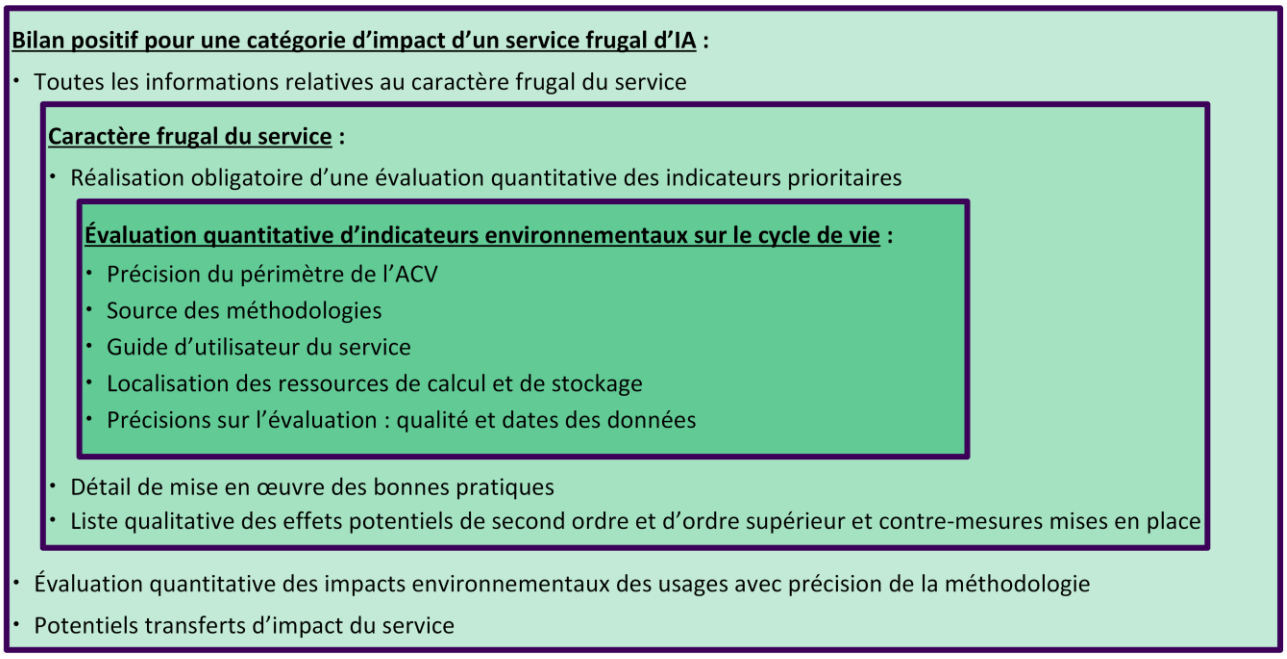

Figure 7 — Schéma des prérequis en terme de communication autour de l'IA frugale

Annexe 1 — Outils et bases de données pouvant être utilisés

Sans en faire faire ici une liste exhaustive, nous vous proposons un tableau de comparaison des outils les plus utiles et pertinents.

Méthode	Mesure de la consommation	Estimation intégrée de la consommation	Estimation de la consommation
But	Mesurer la consommation d'un programme et son empreinte carbone associée pendant son exécution	Estimer et suivre la consommation d'énergie et les impacts environnementaux de l'utilisation de l'IA générative via API	Estimer la consommation d'un programme et son empreinte carbone associée pendant son exécution
Facilité d'utilisation	Moyen : requiert l'installation d'un package et l'implémentation dans le code	Facile/Moyen – requiert une installation d'un package python, mais l'implémentation ne nécessite peu/pas de changement du code existant	Facile : formulaire en ligne
Exemples	CodeCarbon, CarbonTracker	EcoLogits	GreenAlgorithms
Avantages	▪ Mesure directe : pas besoin de mesurer le temps de calcul ▪ Possibilité d'isoler des composants du système dans certains cas ▪ Suit les variations de charge de calcul	▪ Facilité d'implémentation dans un projet existant ou en cours de développement ▪ Modélise la consommation d'énergie et les impacts multicritères (GWP, ADPe, Énergie Primaire) ▪ Intègre la phase de fabrication du matériel, l'intensité du mix électrique dans la région et le PUE du data centre	▪ Facilité d'utilisation ▪ Utilisation possible a priori pour estimer l'impact avant de lancer l'expérience ▪ La surestimation due à l'utilisation du *Thermal Design Power* au lieu de la consommation réelle des processeurs peut compenser la sous-estimation du manque de mesures d'autres composants (alimentation, carte mère...)
Limites	▪ Mesure directe pas toujours possible ▪ Les outils font des choix pas toujours explicités, par exemple pour le facteur d'émission utilisé ▪ Ajout d'un léger surcoût énergétique	▪ Repose sur de nombreuses hypothèses de modélisation par manque de transparence des fournisseurs de servide d'IA générative ▪ Ne supporte que la génération de texte avec LLM (à ce stade)	▪ Il faut connaître le temps de calcul ainsi que la machine qui sera utilisée ▪ Moins précis si la charge de calcul n'est pas constante
Cas typiques d'utilisation	Benchmark de composants Mesure de l'inférence/utilisation ou de l'entraînement d'un modèle	Suivi des impacts de l'utilisation de l'IA générative (monitoring) Estimation des impacts d'une inférence LLM via un fournisseur de service en API	Estimation avant le début du projet Estimation a posteriori si impossibilité de mesure à chaud

Bases de facteurs d'impact existantes

Il existe aujourd'hui différentes bases adoptant différentes méthodologies et approches, en voici quelques exemples :

Base	Accès	Type	Commentaire
Base Empreinte	Libre	Générique et Sectorielle	Base de données publique officielle de facteurs d'émission et de jeux de données d'inventaire nécessaires à la réalisation d'exercices de comptabilité carbone des organisations. La base de données (format CSV) n'est pas spécifique à l'impact du numérique mais contient une cinquantaine de données d'impact provenant de la base NegaOctet. Elle a vocation à intégrer au fur et à mesure de nouveaux facteurs d'impacts sur le numérique.
Boavizta API	Libre	Sectorielle	Données au format de portrait-robot, tiers 1 et 3 représentés. Les impacts intégrés sont évalués au niveau des composants et agrégés au niveau de l'appareil selon une approche bottom-up.
NegaOctet	Payant via EIME	Sectorielle	Les données couvrent des services numériques et une partie des 3 tiers (500+ données). Base de données d'ACV d'équipements physiques selon 16 catégories d'impact et sur les 4 étapes de cycle de vie (fabrication, distribution, utilisation, fin de vie). Cinq niveaux de données sont proposés : wafer (base pour les semi-conducteurs), composant, équipement, système, service numérique.
Resilio DB	Payant	Sectorielle	Modélisation d'équipements des briques 1 et 3 principalement.
Sphera Managed LCA	Payant	Générique	Données produits et composants, adaptée pour la brique 1.
GaBi	Payant	Générique	Base de données d'ACV permettant de calculer les facteurs d'impact de la production de différentes puces de calcul et de mémoires, avec plusieurs méthodes de calcul d'impact.
EcoInvent	Payant	Générique	Données génériques et quelques composants (brique 1). Base de données d'ACV permettant de calculer les facteurs d'impact de la production de wafer avec plusieurs méthodes de calcul d'impact.

Outils pour réduire l'impact environnemental de projets d'IA

La plateforme de deep learning Aidge (https://www.deepgreen.ai/) dédiée à l'embarqué a pour but de faciliter le design et l'optimisation de réseaux de neurones pour un déploiement des applications d'IA embarquée facilité sur différentes cibles matérielles à faible consommation d'énergie et faible empreinte mémoire. Pour répondre aux exigences industrielles et environnementales, Aidge dispose de fonctions d'optimisation pour réduire la complexité des modèles comme la quantification et offre des implémentations compatibles avec une large gamme d'architectures matérielles (MCU, CPU, GPU et accélérateurs neuronaux). Disponible en open-source au travers de la Fondation Eclipse, la plateforme se développe et monte en maturité via les projets DeepGreen et NeuroKit2E, regroupant un large panels d'industriels et universitaires pour devenir la plateforme de référence souveraine pour l'IA embarquée.

Annexe 2 — Schéma fonctionnel du service d'IA Générative Stable Diffusion

Cette annexe présente un exemple de schéma fonctionnel du service d'IA générative « Stable Diffusion » adapté de l'article de Berthelot *et al*. (2024).

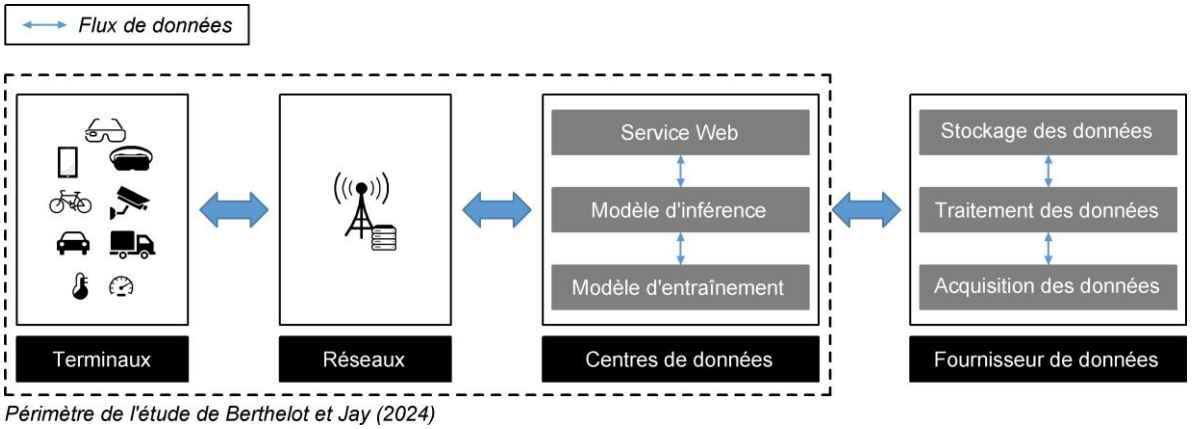

Figure 8 — Schéma fonctionnel du service d'IA générative Stable Diffusion adapté de Berthelot et al. (2024)

Les icônes du schéma sont issues de « *Study on the Economic Potential of Far Edge Computing in the Future Smart Internet of Things* ».

Annexe 3 — Calcul des coûts environnementaux pour le service d'IA générative Stable Diffusion

Cette annexe présente un exemple de méthodes de calcul des coûts environnementaux pour les modèles d'inférence et d'entraînement du service d'IA générative Stable Diffusion. Ces méthodes de calcul sont adaptées de l'article de Berthelot et al. (2024).

Notations communes aux méthodes de calcul

Nom de la variable	Définition
I	Impact environnemental
e	Équipement e
F_e	Empreinte environnementale de l'équipement e incluant les étapes de fabrication, transport et fin de vie
EGM_g	Impact environnemental du mix électrique dans la zone géographique g
PUE	L'indicateur d'efficacité énergétique (Power Usage Effectiveness) du Centre de données
$a_e(t)$	Allocation du temps d'usage t de l'équipement e pendant toute sa durée de vie (c'est à dire sa durée de vie fois le pourcentage d'utilisation en mode actif).

Calcul des coûts environnementaux pour le modèle d'inférence

L'inférence est réalisée par un équipement « à la demande », dans ce cas un GPU e. Le calcul d'une inférence i sur un équipement e consomme de l'électricité et prend un temps t. Le coût de l'électricité est ensuite multiplié par le PUE et l'EGM_g.

Le reste de l'impact est calculé avec une allocation de temps $a_e(t)$ sur l'équipement e, en supposant que les GPU ne puissent pas effectuer d'autres tâches en parallèle. En ce qui concerne les terminaux d'utilisateurs finaux, l'allocation est proportionnelle à la durée de vie totale multipliée par son taux d'utilisation active. Il représente le fait que les appareils à la demande fournis par les centres de données ne sont pas toujours utilisés en permanence.

$$I_{Inférence} = \sum_i C_{i,e} \times EGM_g \times PUE + a_e(t) \times F_e$$

Avec la définition des variables suivantes :

Nom de la variable	Définition
i	Inférence réalisée sur un GPU
$C_{i,e}$	Consommation d'électricité de l'inférence i réalisée avec l'équipement e

Calcul des coûts environnementaux pour le modèle d'entraînement

Cette section se concentre sur le serveur de centre de données nécessaire pour effectuer un ou plusieurs entraînements tr pour le modèle d'IA. Comme précédemment, l'allocation de temps est basée sur la durée de vie de l'équipement, la durée de l'entraînement et son taux d'utilisation active.

Cependant, une attention particulière est accordée à l'estimation de la consommation d'électricité. Comme la reproduction de l'entraînement est trop coûteuse et la modélisation pure insatisfaisante, une solution intermédiaire est nécessaire. Comme l'entraînement est divisé en étapes à coût constant, il suffit de mesurer la consommation d'électricité de quelques étapes pour estimer le coût total de l'entraînement, en supposant la connaissance de l'entraînement initial.

Enfin, l'impact de l'entraînement est un coût fixe nécessaire au lancement du service. Lors de l'évaluation du coût d'un service à l'aide d'un modèle au cours de sa durée de vie, le coût total de l'entraînement est pris en compte. Toutefois, lorsqu'on évalue le coût d'une utilisation unique d'un service en effectuant une inférence à l'aide du modèle, le coût de la formation doit être attribué à chaque inférence.

$$I_{Entraînement} = \sum_{tr} C_{tr,e} \times EGM_g \times PUE + a_e(t) \times F_e$$

Avec la définition des variables suivantes :

Nom de la variable	Définition
tr	Entraînement du modèle d'IA
$C_{tr,e}$	Consommation d'électricité de l'entraînement tr réalisée avec l'équipement e

Les méthodes de calcul des coûts environnementaux des tiers « terminaux » et « réseau », ainsi que la brique fonctionnelle « service web » sont disponibles dans l'article de Berthelot *et al.* (2024).

Recueil des fiches de bonnes pratiques

Table des bonnes pratiques, par thématique :

Analyser l'impact des équipements nécessaires pour le service d'IA et optimiser l'usage (économie de la fonctionnalité)

- AFNOR SPEC IA Frugale-BP N°05,
  Mettre en œuvre des mesures d'éco-conception en phase de développement
- AFNOR SPEC IA Frugale-BP N°20,
  Optimiser l'usage de l'équipement existant
- AFNOR SPEC IA Frugale-BP N°22,
  Favoriser les terminaux utilisateurs/salariés existants pour l'entraînement ou l'inférence du service d'IA
- AFNOR SPEC IA Frugale-BP N°24,
  Assurer la frugalité des infrastructures tout au long de l'exploitation

Gérer les compétences et acculturer à l'IA frugale

- AFNOR SPEC IA Frugale-BP N°14,
  Acculturer et former les parties prenantes
- AFNOR SPEC IA Frugale-BP N°16,
  Identifier & mobiliser le vivier de compétences IA frugale

Mettre en place une gouvernance permettant de questionner la frugalité

- AFNOR SPEC IA Frugale-BP N°09,
  Intégrer la frugalité dans les critères de pertinence de l'IA
- AFNOR SPEC IA Frugale-BP N°11,
  Prévoir la fin de vie dans la gestion d'un projet IA
- AFNOR SPEC IA Frugale-BP N°12,
  Instruire la frugalité dans chaque projet IA grâce au cycle de vie
- AFNOR SPEC IA Frugale-BP N°17,
  Mettre en place et animer le référentiel unique de « services IA » frugaux
- AFNOR SPEC IA Frugale-BP N°18,
  Avoir une offre de produits numériques IA sur étagère favorisant la frugalité

Optimiser la gestion des données et l'intégrer dans une démarche d'amélioration continue (ou management de la qualité)

- AFNOR SPEC IA Frugale-BP N°06,
  Maîtriser le volume des données
- AFNOR SPEC IA Frugale-BP N°07,
  Travailler sur la qualité des données
- AFNOR SPEC IA Frugale-BP N°08,
  Utiliser un jeu de données pertinent pour concevoir le service d'IA
- AFNOR SPEC IA Frugale-BP N°10,
  Définir des règles de stockage des données en fonction des usages
- AFNOR SPEC IA Frugale-BP N°15,
  Faire de la compression de données
- AFNOR SPEC IA Frugale-BP N°19,
  Utiliser des jeux de données open source pour le prototypage

Optimiser la performance du modèle et l'intégrer dans une démarche d'amélioration continue (ou management de la qualité)

- AFNOR SPEC IA Frugale-BP N°03,
  Utiliser des méthodes de compression pour réduire l'empreinte des algorithmes d'IA
- AFNOR SPEC IA Frugale-BP N°04,
  Définir des critères justifiant le ré-entraînement du modèle
- AFNOR SPEC IA Frugale-BP N°26,
  Écrire du code pouvant être amélioré par plusieurs personnes et ré-implémenté sur plusieurs environnements
- AFNOR SPEC IA Frugale-BP N°27,
  Rationaliser les modèles
- AFNOR SPEC IA Frugale-BP N°28,
  Décomposer un gros modèle d'IA en plusieurs petits modèles
- AFNOR SPEC IA Frugale-BP N°29,
  Réutiliser des algorithmes entraînés et partager les algorithmes réalisés (OpenSource)
- AFNOR SPEC IA Frugale-BP N°30,
  Privilégier des modèles plus frugaux
- AFNOR SPEC IA Frugale-BP N°31,
  A/B Testing de modèles pour identifier le modèle avec le meilleur ratio performance/ressources

Qualifier la pertinence de l'IA pour répondre au besoin d'un nouveau service numérique

- AFNOR SPEC IA Frugale-BP N°01,
  Utiliser des méthodes d'analyse de besoin pour mettre en œuvre la frugalité
- AFNOR SPEC IA Frugale-BP N°02,
  Choisir et développer la solution pour répondre spécifiquement au besoin, en considérant les alternatives à l'IA

Outiller la mesure de l'impact environnemental et enrichir la connaissance

- AFNOR SPEC IA Frugale-BP N°21,
  Créer un référentiel des impacts environnementaux des projets
- AFNOR SPEC IA Frugale-BP N°23,
  Réaliser une estimation de la consommation du modèle a priori
- AFNOR SPEC IA Frugale-BP N°25,
  Faire évoluer les stratégies de mesure en fonction des enjeux et des contraintes pour maintenir la frugalité du service d'IA

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

BP 01

N° 01	Utiliser des méthodes d'analyse de besoin pour mettre en œuvre la frugalité	
Gain de frugalité : Élevé	**Effort mise en œuvre :** Modéré	**Qualifier la pertinence de l'IA pour répondre au besoin d'un nouveau service numérique**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation	X	X	X
2 – Conception et développement	X	X	X
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

L'usage de l'IA doit être motivé par la confirmation du besoin via une méthode centrée sur l'utilisateur : « l'emploi d'une solution IA pour répondre à quel usage/finalité ? »

Mise en œuvre

Voici la démarche pour confirmer le besoin via une méthode centrée sur l'utilisateur :

- prise en compte des critères 1.1 et 1.2 du RGESN (Référentiel Général de l'Écoconception des Services Numériques) ; [10]
- prise en compte dans l'AFNOR SPEC 2201 [11], des critères 5.1.1 « collecter et questionner les besoins et usages » et 5.1.2 « Analyser et dimensionner au plus juste les besoins et les usages » ;
- pdopter une approche itérative telle que l'UX design permettant de travailler le besoin de manière à le circonscrire et le ramener à l'essentiel en le centrant sur l'utilisateur. Il est en effet essentiel de se concentrer sur les besoins fondamentaux et pérennes et d'exclure les besoins accessoires, occasionnels, jetables ou de précaution.

Si la solution IA est confirmée pour répondre au besoin, alors l'usage du service IA doit être optimisé dans le parcours de son utilisateur. Il s'agit ainsi d'optimiser son accès par rapport à l'usage attendu (au bon moment et dans le bon contexte), afin de garantir un usage raisonné de la solution pour répondre au juste besoin.

Facteurs clés de succès :

- circonscription du besoin à l'essentiel et au plus juste, sans besoins accessoires, occasionnels, jetables ou de précaution.

Les démarches itératives peuvent amener à de nombreux cycles d'essais et de tests. En matière d'IA, si ceux-ci impliquent de nombreuses phases d'apprentissage, les impacts, notamment énergétiques, pourraient ne pas être négligeables sur l'ensemble du cycle du service IA.

En tout état de cause, il convient de tenir compte systématiquement des impacts des expérimentations et tests dans le bilan global du projet et du cycle de vie du service. Cela signifie que, pour que ce dernier soit le plus réduit possible, une étude réfléchie et frugale des tests doit également être menée et documentée.

Exemple : à chaque test, on ne cherchera pas à obtenir la précision la plus élevée possible de manière, d'une part, à réduire les temps d'apprentissage et, d'autre part, à réduire l'utilisation de ressources énergétiques et matérielles et donc l'impact global des tests.

Sources : Témoignage et bibliographie (cf AFNOR SPEC 2201 « Écoconception des services numériques » [11])

Secteur : Multisectoriel

BP 02

N° 02	**Choisir et développer la solution pour répondre spécifiquement au besoin, en considérant les alternatives à l'IA**	
Gain de frugalité : Élevé	**Effort mise en œuvre :** Modéré	**Qualifier la pertinence de l'IA pour répondre au besoin d'un nouveau service numérique**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation	X		X
2 – Conception et développement			
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

Du fait de leurs spécificités matérielles, les services d'IA présentent une empreinte environnementale potentiellement supérieure aux systèmes numériques plus traditionnels.

Les impacts liés à la fabrication des microprocesseurs spécialisés, ainsi que les prévisions de consommation d'électricité et d'eau supplémentaires générées par l'IA, doivent amener à s'interroger en premier lieu sur la nécessité de recourir à l'IA pour répondre aux besoins des utilisateurs. Afin de limiter l'impact environnemental des projets, il convient donc de s'interroger sur la pertinence de l'IA dans la réponse à apporter au besoin.

Ainsi, l'IA ne doit pas devenir une réponse systématique à tous les besoins : c'est un type de solution dont l'intérêt et l'impact doivent être évalués et comparés à d'autres solutions plus classiques pour identifier la solution la plus pertinente, notamment sur le plan environnemental.

Mise en œuvre

Le questionnement de la pertinence de l'IA et des Systèmes d'IA pour répondre au besoin doit se faire tout au long du cycle de vie du projet mais, en premier lieu, lors des phases amont du projet, notamment pour concevoir la solution et sa frugalité (basée sur l'IA ou non) dans sa globalité, la somme des optimisations locales ne générant pas nécessairement l'optimisation globale :

- avant de développer le modèle d'IA :
  - challenger les besoins exprimés par les utilisateurs et notamment la performance et la qualité des résultats pour arriver aux besoins fondamentaux et permettre de proposer diverses approches et non uniquement celles basées sur l'IA (voir également la bonne pratique N° 01). Les besoins doivent être formulés de telle sorte que toutes les solutions, IA et non-IA, soient possibles. Cela peut signifier un travail spécifique sur les critères de performance attendus de manière à réduire au maximum leur niveau d'exigence. Pour autoriser cette recherche la plus large possible de solutions, l'expression de besoins peut être qualifiée au regard d'un certain nombre de critères (liste non exhaustive) :
    - l'expression de besoins décrit les cas d'usage standards auxquels elle répond,
    - l'expression de besoins ne mentionne pas de solution et permet de proposer plusieurs solutions de différentes natures,
    - l'expression de besoins ne donne pas de critères de dimensionnement de la solution ou si elle en donne, elle doit les justifier,

N° 02	Choisir et développer la solution pour répondre spécifiquement au besoin, en considérant les alternatives à l'IA	
Gain de frugalité : Élevé	**Effort mise en œuvre :** Modéré	**Qualifier la pertinence de l'IA pour répondre au besoin d'un nouveau service numérique**

- l'expression de besoins ne donne pas de critère de la performance (« précision de X % », « temps de réponse de X secondes max. ») ou si tel est le cas, cette performance doit être clairement justifiée au regard des cas d'usage précédemment décrits (pourquoi on ne peut pas accepter une performance inférieure),
- l'expression de besoins doit détailler comment est acceptée ou gérée la non-précision totale si une précision reste exigée (par exemple des vérifications manuelles complémentaires seront nécessaires) et expliquer pourquoi cela ne peut pas constituer le mode de fonctionnement standard de la solution :
  - si la non-précision totale est acceptée, pourquoi ne peut-on pas abaisser davantage le seuil, voire retirer cette exigence de l'expression de besoins ?
  - si une (ou plusieurs) opération supplémentaire est nécessaire pour gérer une précision < 100 %, pourquoi ces opérations ne pourraient-elles pas constituer finalement le besoin ?
- L'expression de besoins indique, et justifie si ce n'est pas acceptable, si un mode de fonctionnement asynchrone est envisageable. Le mode asynchrone doit être décrit de telle sorte qu'il accepte des temps de réponse les plus longs possible,
- l'expression de besoins décrit le fonctionnement accepté en mode dégradé. Ce mode dégradé pourra être considéré comme le critère minimal d'acceptation d'une solution (sorte de mode dégradé permanent) ou comme base de discussion pour les solutions alternatives ;

- ne pas réinventer ce qui existe et répond déjà au besoin : la meilleure solution à un besoin est celle qui existe déjà. L'adaptation d'une solution existante ou d'une règle métier est une solution alternative à envisager. L'expression de besoins doit justifier en quoi les solutions IT existantes ne permettent pas de répondre au besoin, même si elles sont adaptées ;
- faire attention au poids de l'apprentissage pour l'IA. Cela doit passer par une évaluation (pour prime évaluation des impacts et éclairer les choix) et une mesure (pour confirmer les hypothèses d'évaluation, éclairer les décisions et caractériser concrètement les impacts de la ou des solutions) des besoins et des impacts (liés à la fabrication et à l'usage, selon les différents scénarios de déploiement et d'usage) de cette phase d'apprentissage ;
- ne pas oublier de considérer les alternatives à l'IA dans le choix d'algorithme. En effet, l'expérience a montré, par exemple, qu'un recours à une solution plus classique de traitement du signal est plus rapide et moins coûteuse qu'une solution basée sur l'IA. Le recours à l'IA doit être pleinement justifié ;
- une évaluation globale quantifiée, couvrant toutes les phases du cycle de vie, des différentes solutions identifiées doit être menée de la manière la plus exhaustive possible sur les effets de 1er, 2nd et 3ème ordre (pour ces derniers, l'évaluation étant délicate, une caractérisation plus qualitative peut être nécessaire ; dans tous les cas, les limites et les risques d'effets rebond doivent être documentés). Cette évaluation permettra d'évaluer le caractère frugal des solutions et servira de point d'entrée pour les processus décisionnels (voir les Bonnes Pratiques relatives à la gouvernance), étant entendu que la solution à l'impact environnemental le plus faible devrait figurer parmi les solutions à privilégier, ainsi que pour le suivi en production pour valider les hypothèses et identifier les éventuelles dérives qui devraient mener à une nouvelle décision pour réduire l'impact de la solution. L'évaluation devra se faire dans le respect des meilleures pratiques en la matière (voir le chapitre consacré de l'AFNOR SPEC) et en se basant sur un retour d'expérience avec mesure réelle (outillée) si possible.

N° 02	Choisir et développer la solution pour répondre spécifiquement au besoin, en considérant les alternatives à l'IA	
Gain de frugalité : Élevé	**Effort mise en œuvre :** Modéré	**Qualifier la pertinence de l'IA pour répondre au besoin d'un nouveau service numérique**

À ce stade, il est difficile d'avoir une liste exhaustive de critères permettant d'évaluer la frugalité d'une solution. On peut toutefois en proposer quelques-uns à partir de la définition d'un service d'IA frugal :

- la contrainte sur les ressources disponibles (énergie, données, infrastructure, etc.) est-elle intégrée dans l'expression de besoins et lui donne-t-on la priorité ? À défaut, est-elle compatible avec de telles contraintes ?
- la solution respecte-t-elle les contraintes acceptées par la définition du besoin sur les ressources ?
- La solution s'appuie-t-elle sur une autre solution déjà existante ? A-t-on envisagé l'utilisation d'alternatives notamment basées sur des solutions existantes ? Moins la solution requiert de nouveaux éléments, plus elle paraît frugale ; les nouveaux éléments doivent être envisagés uniquement pour ce qui ne dispose pas d'alternative ;
- les nouveaux éléments requis par la solution sont-ils pérennes ou *a contrario* temporaires ou transitoires ? La pérennité signifie que la durée de vie est importante et qu'une suite des investissements a été prise en compte, signe d'une certaine durabilité ;
- entre deux systèmes d'IA équivalents en termes d'infrastructure et de performance, celui requérant le moins de réévaluation ou de réentraînement sera *a priori* [30] plus frugal que l'autre ;
- une somme d'optimisations locales ne fait pas une optimisation globale. Cependant, une solution construite de telle sorte que, sur chacun des axes service (algorithme), données et infrastructure, le minimum de ressources possible est utilisé (le moins d'énergie possible, le moins de stockage, le moins de données d'apprentissage, le moins de ressources computationnelles, etc.), a de fortes chances de constituer une solution frugale si ce n'est pas la solution la plus frugale possible (superlatif qui doit être démontré). Ainsi, à défaut, toute solution dite frugale ne démontrant pas explicitement qu'il n'existe pas d'autres solutions plus frugales mais simplement qu'elle est la plus frugale de celles envisagées ne peut pas être déclarée comme étant la plus frugale possible : c'est une solution *a priori* frugale.

Dans tous les cas, il convient de requestionner le besoin si la solution envisagée n'apparaît ou ne semble pas frugale. Il s'agit de se concentrer sur la réponse aux besoins fondamentaux et pérennes, et d'exclure ceux qui peuvent être accessoires, jetables, ou incertains par exemple. Une démarche centrée utilisateur est un moyen d'y répondre (voir la Bonne Pratique N° 01).

La gouvernance projet doit également être adaptée pour permettre les différentes solutions et les comparer notamment au travers d'indicateurs environnementaux et énergétiques (voir les Bonnes Pratiques N° 6, 8, 9, 11 et 12).

Facteurs clés de succès :

- délais suffisamment confortables pour identifier et qualifier les différentes solutions (IA et non-IA), et pour travailler le besoin de sorte qu'il autorise des solutions alternatives à l'IA ;
- sponsoring fort car les activités à mener sont nombreuses et requièrent du temps, notamment pour sensibiliser et embarquer tous les acteurs dans la démarche de frugalité ;
- par extension, la gouvernance des projets doit être adaptée pour intégrer les délais, les différentes solutions alternatives et les critères environnementaux et énergétiques dans les décisions et les indicateurs de suivi ;
- évaluation précise de l'infrastructure à mettre en œuvre pour les différentes solutions afin d'identifier leur empreinte énergétique respective liée à la fabrication du matériel ;

30) *A priori* car la frugalité se juge sur l'ensemble du cycle de vie de la solution (l'équivalence de la performance supposée ici fait que les effets de 3e ordre n'entrent pas en jeu dans la comparaison mais de manière générale, ces effets de 3e ordre doivent être intégrés au calcul).

N° 02	**Choisir et développer la solution pour répondre spécifiquement au besoin, en considérant les alternatives à l'IA**	
Gain de frugalité : Élevé	**Effort mise en œuvre :** Modéré	**Qualifier la pertinence de l'IA pour répondre au besoin d'un nouveau service numérique**

- mise en place de sondes ou d'outils permettant une mesure de l'empreinte énergétique lors des différentes phases du cycle de vie des solutions (évaluation de l'empreinte en exploitation lors des phases de développement, voire de qualification des solutions, mesure et suivi régulier de la consommation énergétique totale de la solution en phase de production, mesure à rapprocher du prévisionnel pour éventuelle correction du déploiement et/ou de l'architecture d'exécution, etc.) ;
- mise en place d'indicateurs de suivi de la performance et des usages ;
- compétences techniques pour définir l'architecture et développer ;
- sensibilisation des différents acteurs, y compris les partenaires, aux impacts environnementaux du numérique et des calculs (GPU, etc.) ;
- raisonnements/raisons qui mènent à interroger les solutions basées sur l'IA et raisonnements/raisons qui ont amené, le cas échéant, à proposer et choisir des solutions alternatives ;
- explicabilité des choix/décisions (volet gouvernance) et des résultats des solutions pour les utilisateurs : les parties prenantes doivent comprendre (et accepter) les choix faits pour répondre aux besoins et comprendre en quoi la solution apporte une solution satisfaisante.

Sources : Témoignages et Bibliographie (cf. https://hal.science/hal-03009741/)

Secteur : Multisectoriel

BP 03

N° 03	Utiliser des méthodes de compression pour réduire l'empreinte des algorithmes d'IA	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Optimiser la performance du modèle et l'intégrer dans une démarche d'amélioration continue**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et développement	X		
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

Après le développement du modèle, il existe des méthodes de compression automatique des modèles. Cela permet de réduire le volume des solutions stockées.

Mise en œuvre

- Il existe plusieurs méthodes pour réduire le poids d'un modèle tout en essayant de préserver ses performances avec des méthodes comme la Quantization, le Pruning ou la Distillation. Comme ces méthodes peuvent modifier les performances, il peut être utile de mesurer l'impact de ces optimisations avec des vérités terrain ou toute autre approche qualitative.

Sources : Témoignage

Secteur : Multisectoriel

BP 04

N° 04	Définir des critères justifiant le ré-entraînement du modèle	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Faible	**Optimiser la performance du modèle et l'intégrer dans une démarche d'amélioration continue**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et développement			
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation	X	X	
8 – Mise hors service			

Description

Le réentraînement d'un modèle est une des étapes les plus énergivores de son cycle de vie. Il faut donc éviter de le réentraîner trop régulièrement et justifier ce besoin en s'appuyant sur des critères objectifs.

Mise en œuvre

- Les critères peuvent être la modification significative des outputs attendus, l'augmentation ou la modification du jeu de données d'entrée.
- Cf. Critère 9.1 du RGESN (Référentiel Général de l'Écoconception des Services Numériques). [10]

Sources : Témoignage

Secteur : Multisectoriel

BP 05

N° 05	Mettre en œuvre des mesures d'éco-conception en phase de développement	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Analyser l'impact des équipements nécessaires pour le service d'IA et optimiser l'usage (économie de la fonctionnalité)**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			X
1 – Initialisation			X
2 – Conception et développement			X
3 – Vérification et validation			X
4 – Déploiement			X
5 – Exploitation et suivi			X
6 – Validation continue			X
7 – Réévaluation			X
8 – Mise hors service			X

Description

Pour développer un service d'IA, il est nécessaire de recourir à des équipements informatiques (serveurs, ordinateurs ...). Ces appareils ont un impact environnemental tant au niveau de la fabrication qu'au niveau de l'utilisation. Pendant la phase de développement, qui peut prendre plusieurs mois, différents environnements peuvent être créés (production, développement, tests ...) nécessitant eux aussi des ressources.

Pour réduire l'impact environnemental, différentes solutions existent : éteindre automatiquement certains environnements/machines, contraindre les ressources allouées, être vigilant à la localisation des machines et à la charge carbone instantanée de l'électricité.

Mise en œuvre

- De manière générale, appliquer toutes les pratiques d'écoconception des services numériques.
- Mettre en place une métrologie de la consommation de ressources des systèmes de développement en transparence avec les équipes et la gouvernance projet (voir les Bonnes Pratiques N° 21 et 23).
- Maximiser l'utilisation d'équipements existants, de postes de développement standards existants, et décourager l'acquisition de postes de travail spécifiques « gonflés » pour ces activités de développement/test.
- Confiner l'espace mémoire/CPU/stockage de l'environnement de développement pour inciter le développeur à une modération de ses itérations par rapport au volume et à la qualité de données.
- Lorsque le besoin le justifie, encourager l'utilisation de plateformes mutualisées et réutilisables pour la montée en volume du projet de développement en gardant les mêmes préoccupations de suivi des métriques de consommation de ressources.
- Encourager le partage de bonnes pratiques de frugalité de développement et de prise en compte des facteurs de charge sur les ressources partagées (autres projets, contraintes d'effacement électrique...), en relation avec les Bonnes Pratiques N° 20 et 23.

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et
réduire l'impact environnemental de l'IA

N° 05	Mettre en œuvre des mesures d'éco-conception en phase de développement	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Analyser l'impact des équipements nécessaires pour le service d'IA et optimiser l'usage (économie de la fonctionnalité)**

- Prévoir des dispositifs techniques de départ différé, de mise en pause et d'extinction des environnements de développement prenant en compte des phases d'inactivité ou de forte tension sur la demande électrique carbonée. Un entraînement doit pouvoir être initié à 22 h et non à 18 h, puis générer une extinction des environnements à la fin de son exécution nocturne.
- Inciter à la réalisation des phases d'entraînement en période de basse activité des infrastructures et de disponibilité d'électricité décarbonée (nuit et/ou après-midi suivant localisation des ressources de calcul).
- Adopter sur le projet des pratiques minimisant les volumes de déplacements carbonés des collaborateurs.
- Ne pas redévelopper ce qui existe déjà et utiliser des bibliothèques éprouvés de haut niveau.
- Être vigilant à l'impact des entrées/sorties qui peut fortement dépendre de l'infrastructure cible (pose de travail vs cluster spécialisé).

En fonction de l'intensité carbone de l'électricité, évaluer l'intérêt respectif de traitements locaux, régionaux ou nationaux (ex. : Inde vs Pologne vs Allemagne vs Suède ou France...).

Sources : Témoignage et bibliographie

- https://calculator.green-algorithms.org/
- https://graal.ens-lyon.fr/~llefevre/greendays2024/GreenDays2024_Mokhtari.pdf

Secteur : Multisectoriel

BP 06

N° 06	Maîtriser le volume des données	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Optimiser la gestion des données et l'intégrer dans une démarche d'amélioration continue**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et développement		X	X
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi		X	X
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service		X	X

Description

Le volume de données stockées a un impact sur les infrastructures. Il peut aussi être un révélateur d'une faible qualité des données, ce qui a un impact sur la conception du service d'IA et le choix des algorithmes. En effet, des algorithmes plus énergivores peuvent être choisis pour compenser la qualité des données.

Toute démarche qui permet de limiter les duplications de données et les flux et ainsi de maîtriser le volume de données stockées permet de réduire l'empreinte environnementale du projet d'IA.

Mise en œuvre

- Représenter les données dans le cadre de l'architecture d'entreprise et dans les référentiels de modèles de données d'entreprise.
- Mettre en place une gouvernance de la donnée pour piloter les volumes de données stockées.
- Mettre en place le principe d'unicité de la donnée et de fait, limiter au strict minimum les duplications de données dans l'architecture SI pour répondre aux besoins.
- Définir une politique d'archivage et de suppression des données en décrivant notamment correctement la fréquence.
- Utiliser des outils de compression des données.

Sources : Témoignages et Bibliographie (cf méthodes DMBOK, TOGAF® Series Guide: Environmentally Sustainable Information Systems (opengroup.org))

Secteur : Multisectoriel

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

BP 07

N° 07	Travailler sur la qualité des données	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Optimiser la gestion des données et l'intégrer dans une démarche d'amélioration continue**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et développement		X	
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

En termes de consommation de ressources, la donnée est un enjeu crucial, que cela soit pour son stockage mais aussi parce que la consommation de ressources pour entraîner une IA est fonction de la quantité de données qu'elle reçoit pour son entraînement.

Une bonne qualité de données permet de réduire la quantité nécessaire à une bonne optimisation du modèle et donc la quantité de ressources consommées. Il faut donc privilégier la qualité à la quantité, une mauvaise qualité augmentant aussi le temps et les biais d'apprentissage.

Mise en œuvre

- Définir des critères de qualité d'une donnée.
- Nettoyage des données : supprimer les doublons, les valeurs manquantes, les erreurs de saisies et les données aberrantes qui peuvent fausser les résultats.
- Normalisation des données : convertir les données dans un format cohérent et standardisé. Par exemple, convertir toutes les dates dans le même format ou en normalisant les unités de mesure.
- Vérification de la cohérence : vérifier que les données soient cohérentes entre elles et avec les règles métier. Par exemple, vérifier que les dates de naissance sont antérieures aux dates de décès.
- Documentation des données : documenter les données de manière précise et exhaustive en décrivant leur origine, leur format, leur qualité et leur utilisation prévue.

Sources : Témoignage

Secteur : Multisectoriel

BP 08

N° 08	Utiliser un jeu de données pertinent pour concevoir le service d'IA	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Optimiser la gestion des données et l'intégrer dans une démarche d'amélioration continue**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et développement		X	
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

Moins il y a de données et plus l'impact en ressource est maîtrisé, mais pour cela il faut que les données utilisées soient pertinentes.

Un jeu de données pertinent peut être un jeu de données déjà annoté, ayant déjà fait l'objet d'un tri et d'une sélection qualitative et quantitative.

Ce type de jeu est couramment appelé « produit data » dans le cadre d'une gouvernance de la donnée moderne.

Lors de la conception de l'architecture et du modèle, il est possible d'élaborer un jeu de données qui soit le plus frugal possible, en analysant pour chaque donnée considérée ces critères : leur type (image, texte, vidéo, données structurées ...), leur disponibilité, leur accessibilité, leur qualité (déjà utilisée, auditée par un organisme tiers) ou leur spécificité métier.

Mise en œuvre

- Utiliser des jeux de données disponibles en open data (consulter les annuaires de données open data existants).
- Participer à des communautés entre pairs afin de partager des jeux de données de qualité.
- Définir des métriques pour évaluer et partager la qualité des données (par exemple : exactitude, complétude, pertinence métier, accessibilité, fraîcheur ...).
- Mettre en place une gouvernance de la donnée pour faciliter la création de jeux de données pertinents avec des fournisseurs de données et des data engineers.
- Impliquer les producteurs de données et les utilisateurs finaux dans la réalisation du jeu de données.
- Afin de réduire le volume de données, créer des données synthétiques plus pertinentes.

Sources : Témoignages et bibliographie (cf méthodologie DMBOK, TOGAF)

Secteur : Multisectoriel

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

BP 09

N° 09	Intégrer la frugalité dans les critères de pertinence de l'IA	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Mettre en place une gouvernance permettant de questionner la frugalité**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et développement	X	X	
3 – Vérification et validation	X	X	
4 – Déploiement			
5 – Exploitation et suivi	X	X	
6 – Validation continue	X	X	
7 – Réévaluation	X	X	
8 – Mise hors service			

Description

L'identification d'une IA pertinente dépend énormément du ou des projets auxquels elle s'applique. Il existe plusieurs types de systèmes et de services d'IA, d'où l'existence de plusieurs critères à prendre en compte pour faire le bon choix :

- besoins opérationnels et spécifications ;
- compatibilité de ces besoins avec la frugalité ;
- critères intrinsèques du modèle (licence, capacités…) ;
- analyse du cycle de vie de l'IA au regard de la frugalité ;
- privilégier les fournisseurs de technologies claires sur leur politique environnementale ;
- favoriser les technologies connues ou accepter de monter en compétence pour effectuer le bon choix :
- prendre en compte l'aspect de maintenabilité durant toute la phase d'exploitation du système ou service d'IA.

Mise en œuvre

Critères d'IA pertinente : il existe plusieurs types d'IA (connexionniste, symbolique, hybride, RL, LLM…).

En plus des critères de pertinence de l'IA sur les plans algorithmiques et de réponse au besoin, il convient de prendre en compte des critères de frugalité.

L'IA pertinente est celle qui saura répondre aux besoins opérationnels tout en garantissant la prise en compte des conditions de frugalité. Cela signifie qu'il est souvent utile d'itérer sur l'établissement des spécifications du besoin. Certains types d'IA pourront être écartés en amont car incompatibles avec les besoins opérationnels.

Par exemple, le besoin d'explicabilité peut imposer des solutions issues de l'IA symbolique.

De même, l'IA générative peut s'imposer du fait de son efficacité pour répondre aux spécifications. Le sujet est de travailler le besoin pour confirmer que seule une IA peut y répondre (voir Bonne Pratique N° 02).

Avant toute sélection ou engagement, il est important de se poser ces questions simples : Les enjeux environnementaux sont-ils considérés dans la solution IA potentielle ? Sont-ils déjà compatibles ou bien y a-t-il un antagonisme entre les autres besoins et la frugalité (par exemple le temps réel ou la maximisation de la détection vs la quantization ou la puissance de calcul …).

N° 09	Intégrer la frugalité dans les critères de pertinence de l'IA	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Mettre en place une gouvernance permettant de questionner la frugalité**

Le choix du modèle de licence (MIT…) peut avoir un impact sur le choix des modèles. La capacité à réentraîner le modèle peut également être un critère de choix. Une bonne pratique peut être de tracer l'analyse réalisée pour sélectionner le meilleur modèle pour les projets futurs.

Il faut mener la réflexion de la frugalité conjointement sur les deux critères suivants : durée d'exploitation et conception (pouvant intégrer l'apprentissage). On pourra accepter une consommation plus importante lors de la conception car la durée de vie du système sera longue et le résultat sera plus efficace énergétiquement. Mais il faudra prendre en compte la frugalité à chaque étape du cycle de vie du système.

Concernant les modèles préentraînés et/ou les raisonneurs/solveurs pour la symbolique, il est préférable de sélectionner ceux frugaux et mettant à disposition des critères quantitatifs pour rendre compte de leur impact environnemental. De manière générale, il vaut mieux valoriser la chaîne de valeur de la frugalité entre clients et fournisseurs à tous les niveaux. La même démarche peut s'appliquer pour les fournisseurs cloud.

Dans les critères de sélection, on peut également évoquer les compétences disponibles au sein de l'organisation. Une technologie prometteuse mais non maîtrisée va non seulement coûter plus cher à mettre en œuvre, coût qui ne sera pas forcément rentabilisé en interne et qui peut passer au second plan, sauf si la gouvernance l'impose en interne, mais présente également le risque d'un impact environnemental plus important. Pour être frugal, il est nécessaire d'organiser le double regard sur le besoin considéré, technique (choix du modèle, hyper paramètres …) et métier (par exemple, détection dans l'air différente du milieu aquatique, besoins différents). Cela se définit et répartit directement au sein de l'équipe projet.

Le niveau de compétences permet aussi de proposer des modèles qui seront plus efficaces pour répondre à une problématique donnée, depuis la phase de conception jusqu'à l'exploitation et la réévaluation. Il est donc important que les aspects de frugalité soient traités auprès de référents d'IA frugale, afin d'opter pour la meilleure stratégie à adopter pour la frugalité.

Un système IA peut devenir obsolète avec le temps ou du fait d'une évolution technologique et réglementaire, et il est nécessaire de le détecter car un tel système ne répondra plus aux besoins et il peut même devenir inutile. Il faut penser à un moyen de détecter cette obsolescence et la corriger, cela pouvant être un réapprentissage incrémental ou complet voire même la nécessité de revoir complètement le modèle (retour à la phase conception et développement depuis l'exploitation). Il faut réfléchir lors de la conception s'il est possible d'ajouter une fonctionnalité de détection d'obsolescence au système IA, ce qui a pour conséquence aussi de rajouter un traitement et donc de consommer des ressources. Ce dernier point peut être traité techniquement et/ou organisationnellement. Pour les gros systèmes, il est suggéré d'opter pour les deux. Pour les petits systèmes avec des équipes plus réduites, une solution de monitoring technique simple peut être considérée.

Sources : Témoignages

Secteur : Multisectoriel

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

BP 10

N° 10	Définir des règles de stockage des données en fonction des usages	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Optimiser la gestion des données et l'intégrer dans une démarche d'amélioration continue (ou management de la qualité)**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et développement		X	
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service		X	

Description

Le stockage des données entraîne une certaine consommation de ressources, ici l'électricité des centres de données et le matériel utilisé (disques), en fonction de la performance de ces derniers. En effet, si ceux-ci garantissent une grande disponibilité (grande vitesse de lecture/écriture des données) ou un haut niveau de redondance (pour garantir la non-corruption des données), leur consommation augmente sensiblement.

Le stockage de données par le service doit ainsi opérer une distinction entre les données dont le stockage est nécessaire et celles dont l'usage ne requiert pas une grande disponibilité. Par ailleurs, chaque donnée ne gardant pas un niveau de redondance constant, une réévaluation de celle-ci doit être effectuée afin d'optimiser son niveau de disponibilité, garantissant une utilisation optimale des ressources.

Il est donc intéressant de trouver un compromis entre performance du serveur et sa frugalité au regard de l'importance des données concernées et de la fréquence à laquelle elles sont utilisées. Il peut être utile de se référer à la Bonne Pratique N° 07 afin de privilégier la qualité des données plutôt que la quantité.

Mise en œuvre

- Définir clairement une politique de stockage et de conservation des données.
- Mettre en place des routines automatiques de réévaluation de la redondance, de suppression et de déplacement des données vers des stockages « plus froids » (moins consommateurs de ressources).
- Supprimer les données à la fin du projet.

Sources : Témoignage

Secteur : Multisectoriel

BP 11

N° 11	Prévoir la fin de vie dans la gestion d'un projet IA	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Mettre en place une gouvernance permettant de questionner la frugalité**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse	X	X	X
1 – Initialisation			
2 – Conception et développement			
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

Chaque solution potentielle doit être évaluée en termes d'impacts environnementaux sur l'ensemble du cycle de vie du projet d'IA et de la solution selon les scénarios d'usages documentés dans l'expression de besoins et en tenant compte des effets de bord et rebond possibles.

La question de la frugalité des solutions d'IA doit également être posée sur l'ensemble du cycle de vie, allant de la phase de R&D préalable (l'apprentissage mais pas seulement : l'ensemble des tests et tentatives pour identifier l'algorithme, etc.) jusqu'à la fin de vie du modèle et des données utilisées et/ou collectées.

Si une solution intègre l'achat de matériel spécifique on-premise (logiciel sur site) comme des cartes GPU, l'évaluation des impacts doit intégrer la seconde vie ou le décommissionnement de ces équipements : le devenir de ces équipements post projet doit être identifié dès le départ.

Mise en œuvre

Dès la création du projet, il est nécessaire d'avoir dans l'équipe projet tous les acteurs représentatifs du cycle de vie de la solution IA et de l'ensemble des équipements associés (terminaux utilisateurs/capteurs, logiciels, réseaux, serveurs, hébergement).

La fin de vie d'un logiciel étant immatérielle, les considérations suivantes doivent être prises en compte :

- définir les règles et la mécanique de purge des données qui pourront devenir inutiles ;
- mettre en place une stratégie de réutilisation et de seconde vie des modèles d'IA en interne et potentiellement avec la communauté. Ces modèles d'IA peuvent en effet enrichir une bibliothèque interne et partagée entre les collaborateurs. Par ailleurs, de tels modèles peuvent également avoir une seconde vie en les mettant en open-source. Ils pourraient ainsi à terme devenir un commun numérique pour la communauté. De tels partages permettent d'éviter des réentraînements coûteux et énergivores.

La fin de vie du hardware on-premise nécessite d'être prise en compte indépendamment et tout au long de la gestion du projet. Il est primordial de prendre en compte la durée de vie du matériel utilisé et de définir des scénarios de décommissionnement à la fin du projet et de revalorisation des équipements utilisés.

Sources : Témoignages et Bibliographie

Secteur : Multisectoriel

BP 12

N° 12	Instruire la frugalité dans chaque projet IA grâce au cycle de vie	
Gain de frugalité : Élevé	**Effort mise en œuvre :** Élevé	**Mettre en place une gouvernance permettant de questionner la frugalité**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse	X	X	X
1 – Initialisation			
2 – Conception et développement			
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

Le cycle de vie de l'IA est aujourd'hui défini dans une norme ISO internationale. Il peut être facilement utilisé dans la gestion de tout projet IA et ainsi, faciliter la matérialité de la frugalité pour tout système et/ou service d'IA. La question de la frugalité de l'IA se pose ainsi dans le déploiement successif des étapes de la vie d'un projet IA et, de façon plus transverse, dans les axes traditionnels du pilotage du projet.

Étant donné sa dimension novatrice, cette approche, lorsqu'elle est porteuse d'impact, peut dépasser le cadre des projets d'IA pour s'appliquer à tout type de gestion de projet.

À l'inverse, certains freins ou obstacles à cette approche émergent des témoignages. Il reste primordial d'organiser la prise de hauteur sur chaque évolution et/ou nouveau projet. Une simple modification d'un système même frugale peut occasionner des effets de bord indirects sur le système, d'autres systèmes voir d'autres services.

Mise en œuvre

- Analyser tout projet IA au regard de son cycle de vie de l'IA.
- Intégrer l'instruction de la frugalité dans la gestion de projet, dans le déploiement successif des étapes de la vie d'un projet. Ici dans 9 étapes identifiées (les illustrations données sont des exemples, et ne revêtent pas de caractère exhaustif) :
  - cadrage du projet *Ex. Dès le démarrage des projets d'IA, intégrer une réflexion sur les enjeux environnementaux ;*
  - expression des besoins ;
  - identification au démarrage des KPIs du projet (build) ;
  - identification au démarrage des KPIs du projet une fois lancé (run) *Ex. Adopter une logique de financier en évaluant le ROI carbone de l'investissement, incluant une estimation de la durée de vie des projets* ;
  - documentation technique *Ex. Intégrer dans la documentation technique du projet tous les éléments d'instruction de la frugalité IA* ;
  - développement technique ;
  - phase de tests (technique et métier) *Ex. optimisation de la consommation d'énergie des environnements* ;
  - mise hors service à la fin du projet *Ex. désactivation des environnements à la fin du projet* ;

N° 12	Instruire la frugalité dans chaque projet IA grâce au cycle de vie	
Gain de frugalité : Élevé	**Effort mise en œuvre :** Élevé	**Mettre en place une gouvernance permettant de questionner la frugalité**

- suivi de la performance de la solution d'IA en usage *Ex. intégration d'une vision systémique de l'IA frugale dans le suivi des indicateurs (du moins une partie).*

- Intégrer l'instruction de la frugalité dans la gestion de projet, dans les axes traditionnels du pilotage de projet :
  - gouvernance *Ex. une gouvernance qui fait le pont entre le projet d'IA frugale et les autres instances de gouvernance projet au sein de l'organisation, afin d'infuser les bonnes pratiques* ;
  - pilotage *Ex. Compléter les axes traditionnels de pilotage du projet IA (coûts, qualité, délai) par celui de l'impact environnemental (valoriser les plus-values extra financières ou les impacts environnementaux évités dans le business model, définir un « budget CO_2 » à ne pas dépasser)* ;
  - comitologie *Ex. intégrer le suivi de la frugalité du projet d'IA dans toutes les instances pertinentes (daily, weekly, stand up, validation, …)* ;
  - gestion du changement *Ex. 1 : Afin de s'assurer du soutien des équipes métier dans le projet, utiliser des leviers comme la CSRD pour engager et mobiliser en interne. / ex 2 : Entretenir une culture de projet centrée sur l'impact à long terme et l'efficacité collective* ;
  - RACI *Ex. Intégrer dans l'équipe projet tous les acteurs représentatifs du cycle de vie de la solution et de l'ensemble des équipements associés (terminaux utilisateurs/capteurs, logiciels, réseaux, serveurs, hébergement).*

Les obstacles :

- la difficulté à mobiliser les dirigeants, notamment au niveau de la gouvernance ;
- la préférence pour atteindre des objectifs de court terme (financiers) versus long terme (environnementaux) ;
- maintenir une efficacité constante dans le projet.

Sources : Témoignage

Secteur : Multisectoriel

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

BP 13

N° 13	Piloter la performance environnementale des systèmes d'IA	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	Mettre en place une gouvernance permettant de questionner la frugalité

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse	X	X	X
1 – Initialisation			
2 – Conception et développement			
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

Dans un contexte de très forte demande en puissance de calcul, le pilotage de la performance environnementale des systèmes d'IA a pour objectif de réduire l'impact environnemental pendant toutes les phases du cycle de vie.

Il s'agit notamment d'intégrer ce suivi dans la gouvernance des projets et de piloter des plans d'action de réduction dans une approche d'amélioration continue.

Par exemple, une mesure systématique et précise des émissions de gaz à effet de serre s'avère nécessaire pour mieux évaluer l'empreinte carbone de l'IA et du *Machine Learning*, à la fois en phase de R&D et de production.

NOTE Cette Bonne Pratique N° 13 est complémentaire à la Bonne Pratique N° 05.

Mise en œuvre

- Analyser les indicateurs environnementaux et les métriques associées dans un tableau de bord de suivi.
- Intégrer le suivi du respect des indicateurs dans la gouvernance des projets et dans le système de management des risques.
- Mettre en place les processus et les outils pour collecter les données nécessaires aux indicateurs de suivi de l'impact environnemental.
- Mesurer les performances tout au long du cycle de vie des systèmes d'IA (entraînement, inférence, réentraînement, nombre d'équipements utilisés …).
- Décrire le processus de remontée des alertes.
- Définir des plans d'action de réduction des impacts environnementaux dans une démarche d'amélioration continue.
- Assurer une veille pour la mise à jour des indicateurs et des systèmes de mesures pour tenir compte des évolutions technologiques et des pratiques d'évaluation.
- Intégrer dans les contrats avec les fournisseurs, l'accès aux données pour alimenter les indicateurs.

Sources : Témoignages et Bibliographie (cf. Impact AI Green IT boite à outils, Je code les bonnes pratiques)

Secteur : Multisectoriel

BP 14

N° 14	Acculturer et former les parties prenantes	
Gain de frugalité : Élevé	**Effort mise en œuvre :** Modéré	**Gérer les compétences et acculturer à l'IA frugale**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse	X	X	X
1 – Initialisation			
2 – Conception et développement			
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

La compréhension des facteurs qui ont un impact sur la performance environnementale des systèmes et services d'IA permet de réaliser des arbitrages en faveur de la frugalité tout au long du cycle de vie.

Selon les métiers et les personas, les compétences et les modalités d'actions sont différentes.

Les 4 grands types de personas sont les suivants :

- les spécialistes (architectes, data scientist , DevOPS, data engineer),
- les généralistes (chef de produit, designers),
- les décideurs,
- les utilisateurs.

Un socle commun de connaissances est nécessaire pour tous ces personas afin que soient maîtrisés :

- les facteurs qui influent sur la performance environnementale du modèle d'IA,
- la stratégie de réduction de l'empreinte environnementale du modèle d'IA et les indicateurs de pilotage mis en place,
- la compréhension du service rendu par un algorithme pour répondre à un besoin et ses impacts environnementaux afin de faire un arbitrage service rendu/frugalité.

Les grandes modalités d'actions sont les suivantes : former, sensibiliser, acculturer, animer une communauté, recruter.

Pour les spécialistes, il s'agit d'intégrer la frugalité au cœur de leurs pratiques.

L'optimisation des modèles est au cœur du métier des data scientists car un modèle qui est performant a les caractéristiques suivantes :

- précision élevée : le modèle doit être capable de produire des résultats précis et cohérents, avec un taux d'erreur minimal ;
- capacité à généraliser : le modèle doit être capable de s'adapter à de nouveaux exemples de données et de fournir des résultats précis même lorsque les données d'entrée varient ;
- efficacité : le modèle doit être capable de traiter les données rapidement et efficacement, sans nécessiter de ressources informatiques excessives ;
- interprétabilité : le modèle doit être conçu de telle manière que les utilisateurs puissent comprendre comment il prend ses décisions, ce qui est important pour établir la confiance et l'adoption de l'IA ;
- évolutivité : le modèle doit être capable de gérer des ensembles de données volumineux et complexes, ainsi que d'évoluer avec les besoins de l'entreprise.

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

N° 14	Acculturer et former les parties prenantes	
Gain de frugalité : Élevé	**Effort mise en œuvre :** Modéré	**Gérer les compétences et acculturer à l'IA frugale**

L'autoformation et le travail entre pairs est un moyen pour les data scientists d'augmenter la performance de leurs modèles. La maîtrise d'outils de mesure permet de vérifier la performance environnementale.

Une architecture d'IA est performante si elle est conçue en fonction des besoins et qu'elle est suffisamment flexible pour évoluer. Elle doit respecter les critères suivants :

- modularité : l'architecture doit être conçue de telle manière que les différents composants puissent être facilement remplacés ou mis à niveau sans affecter l'ensemble du système ;
- intégration : l'architecture doit prendre en charge l'intégration de différents outils et technologies d'IA, tels que les frameworks d'apprentissage automatique, les bibliothèques de traitement du langage naturel et les outils de visualisation de données ;
- sécurité : l'architecture doit prendre en compte les considérations de sécurité, telles que la protection des données et la confidentialité, pour garantir que les modèles d'IA soient déployés de manière sûre et fiable ;
- efficacité : l'architecture doit être conçue pour optimiser les performances des modèles d'IA, en utilisant des techniques telles que la parallélisation et la distribution des charges de travail ;
- interopérabilité : l'architecture doit prendre en charge l'interopérabilité entre différents systèmes et plateformes, afin de faciliter l'intégration de l'IA dans les processus métier existants.

Les choix d'architecture peuvent aller à l'encontre des besoins de la frugalité :

- un composant ou une librairie peuvent être choisis en fonction de la complexité de l'architecture existante ;
- les contraintes de sécurité peuvent nécessiter des composants matériels et logiciels supplémentaires, tels que des pare-feu, des systèmes de détection d'intrusion, des outils de chiffrement et des mécanismes d'authentification. Ces composants peuvent augmenter la complexité de l'architecture et entraîner des coûts supplémentaires en termes de matériel, de logiciels et de main-d'œuvre.

La frugalité est une compétence spécifique à intégrer pour faire de la frugalité by design dans l'architecture.

Le métier du DevOPS peut intégrer des compétences clés pour suivre les consommations et piloter les indicateurs environnementaux. Le DevOPS peut produire des dossiers d'arbitrage sur la performance environnementale avec des analyses fiables et précises. Il faut bien veiller à aller au-delà du FinOPS car le financier peut être décorrélé de l'impact environnemental, et les acteurs du DevOPS doivent en avoir conscience.

Le métier du data engineer est facilité s'il existe une bonne gouvernance qui permet de maîtriser le cycle de vie et la qualité de la donnée. Si les fournisseurs de données mettent à disposition des produits data respectant des principes d'interopérabilité et d'autoservice permettant leur usage alors le data engineer peut proposer des architectures de données les plus frugales possible.

Mise en œuvre

- Cartographier les compétences de son organisation sur les impacts environnementaux et proposer des plans de formation pour faire monter en compétence les parties prenantes.
- Recruter des experts environnementaux du numérique (réalisation d'ACV et de mesure de l'empreinte carbone) et/ou sur la conformité réglementaire (CSRD, AI Act, …).
- Créer et intégrer des communautés de pairs pour échanger sur les bonnes pratiques environnementales au service d'une IA Frugale.
- Proposer des sessions de formation pour différents publics avec des partages de bonnes pratiques.
- Intégrer dans les fiches de poste les compétences environnementales nécessaires (la performance des modèles, la frugalité des architectures, la mesure et pilotage) et assurer un suivi.
- Proposer des preuves de concepts ou des sessions type hackathon pour tester des approches de frugalité dans les systèmes d'IA.
- Proposer des formations de prise en main aux utilisateurs en mentionnant les impacts environnementaux du service.
- Embarquer les sachants sur la frugalité & maintenir les compétences des sachants (formations avec les éditeurs de solution).

Sources : Témoignages

Secteur : Multisectoriel

BP 15

N° 15	Faire de la compression de données	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Faible	**Optimiser la gestion des données et l'intégrer dans une démarche d'amélioration continue (ou management de la qualité)**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et développement		X	
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service		X	

Description

Le stockage de la donnée étant un facteur consommateur de ressources, il est nécessaire d'identifier des solutions pour réduire leur consommation énergétique. Diminuer le volume des données peut-être intéressant. En effet, les algorithmes d'IA n'utilisent parfois qu'une partie des données, ou une version dégradée de celles-ci (par exemple une image en plus basse résolution que celle stockée).

De même, après la mise hors service d'un service d'IA, si la suppression des données n'est pas envisageable, compresser la donnée peut être un compromis intéressant entre le fait de capitaliser sur les données d'exploitation du service et le fait de garder une approche frugale.

Mise en œuvre

- Stocker les données sous une forme compressée/réduite/dégradée suffisante à leur bonne exploitation. Par exemple :
  - pour des images : stocker en faible résolution plutôt que des images 2K ou 4K, d'autant plus si le prétraitement du service induit un redimensionnement (comme c'est souvent le cas) ;
  - pour des données textuelles : stocker un embedding (vecteur représentant le texte) ou une version nettoyée du texte ne contenant que les données pertinentes ;
  - pour des données numériques : stocker les données transformées par des analyses statistiques (comme la PCA).
- Ne pas hésiter en fin de projet à appliquer cela aux données d'exploitation et à appliquer une compression encore plus forte pour l'archivage de ces données.

Sources : Témoignage

Secteur : Multisectoriel

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

BP 16

N° 16	Identifier & mobiliser le vivier de compétences IA frugale	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Gérer les compétences et acculturer en IA frugale**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse	X	X	X
1 – Initialisation			
2 – Conception et développement			
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

La mise en place et le maintien à jour des connaissances et compétences liées à l'IA frugale sont essentiels pour assurer leur pérennité. Elle se décline via 4 objectifs, auxquels répond le vivier de compétences :

- acculturer & maintenir les compétences des utilisateurs de services/solutions d'IA (exemple : ateliers de sensibilisation, hackathon ...) ;
- embarquer les sachants dans la frugalité de l'IA et maintenir les compétences des sachants (formations avec les éditeurs) ;
- participer à la qualification des cas utilisateurs ;
- participer à la réalisation de projets de mise en œuvre de services/solutions d'IA.

Il y a 4 typologies de connaissances/compétences à considérer :

- les connaissances/compétences transversales & universelles (projet, gouvernance, outils) ;
- les connaissances/compétences spécifiques à :
  - l'analyse et l'optimisation des modèles,
  - l'analyse et l'optimisation de la data utile,
  - l'analyse et l'optimisation des équipements (infrastructure/device).

Mise en œuvre

- **CARTOGRAPHIER** les 4 typologies de sachants qui disposent de connaissances théoriques ou compétences opérationnelles :
  - les sachants généralistes « Frugalité » (aussi appelés ambassadeurs ou défricheurs) ;
  - 3 sortes de spécialistes pour l'optimisation de la frugalité :
    - pour les modèles,
    - pour les données,
    - pour l'infrastructure.

N° 16	Identifier & mobiliser le vivier de compétences IA frugale	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Gérer les compétences et acculturer en IA frugale**

Si un pool de compétences semblable préexiste pour l'une des typologies, il est préférable de perfectionner la connaissance et compétence des personnes de ce pool et de ne pas créer de pool supplémentaire. Sinon, il sera nécessaire de le mettre en place à une échelle transversale. Pour faciliter la cartographie des profils clés de ce pool, voici quelques exemples de rôles concernés :

- parmi les généralistes :
  - leader numérique responsable/greenIT/finOps,
  - architectes d'entreprise,
  - architectes data,
  - UX,
  - chef de projets IA ;
- parmi les spécialistes :
  - data scientists,
  - DevOps,
  - architecte solution.

- **MOBILISER** : le pool de compétences doit définir les modalités de mobilisation des sachants selon 3 formats :
  - en avant-projet, afin d'accompagner le demandeur dans la définition du modèle de mise en œuvre optimal du service/solution d'IA (principalement les sachants généralistes) :
    - il réalisera une partie du développement et déléguera une autre,
    - il déléguera au pool de compétences ;
  - au lancement du projet de mise en œuvre d'un service/solution d'IA, réservation des « sachants généralistes et spécialistes » pour rejoindre le projet selon 2 moments clés :
    - dès le début « frugalité by design »,
    - lors de la mise en production, réaliser un passage de responsabilité interéquipes ;
  - hors projet de mise en œuvre de service/solution d'IA, en fil rouge : interventions ponctuelles lors de projets d'acculturation à l'échelle entreprise.

Sources : Témoignages

Secteur : Multisectoriel

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

BP 17

N° 17	Mettre en place et animer le référentiel unique de « services IA » frugaux	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Mettre en place une gouvernance permettant de questionner la frugalité**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse	X	X	X
1 – Initialisation			
2 – Conception et développement			
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

La mise en place d'un unique référentiel de l'intégralité des « services d'IA » et « systèmes IA», permettant la capitalisation de l'analyse de la frugalité de l'IA. Les « services d'IA » et « systèmes d'IA » seront évalués avec, *a minima,* ces 2 types de critères :

- le premier critère questionne la frugalité du service/système ;
- le second critère portant sur le fait d'avoir une approche frugale d'un point de vue organisationnel de ce service/solution.

Mise en œuvre

- Mettre en place un référentiel d'offres de « services d'IA » et « systèmes d'IA » avec *a minima* :
  - un identifiant unique par service d'IA ;
  - son statut de déploiement ;
  - l'identification de l'équipe en charge de sa gestion ;
  - la description des modèles du service d'IA ;
  - concernant les informations liées à la donnée et l'infrastructure, 2 options :
    - dans le cas de la préexistence d'une base de données décrivant la donnée ou l'infrastructure, il est nécessaire de faire le lien avec ce référentiel d'offres ;
    - dans le cas contraire, recenser les informations relatives à la donnée et l'infrastructure dans ce référentiel.
- Créer un espace spécifique d'analyse de la frugalité dans ce référentiel des services d'IA et ajouter au niveau de chacun d'entre eux :
  - un critère déclaratif d'évaluation de la frugalité du service d'IA (exemple : non étudié, en cours d'analyse, étudié, audité ...) ;
  - un critère déclaratif de l'existence d'une procédure de réutilisation (re-use) et d'évolution du service/système d'IA produite par l'équipe qui assure la gestion du service d'IA ;
  - dans l'idéal, l'analyse de la frugalité peut être complétée par des informations plus détaillées décrivant chaque indicateur de frugalité par service d'IA :
    - chaque indicateur est évalué au regard de la donnée disponible pour établir cet indicateur ;
    - chaque indicateur dispose d'une ressource externe décrivant le procédé d'analyse réalisé.

Sources : Témoignages

Secteur : Multisectoriel

BP 18

N° 18	Avoir une offre de produits numériques IA sur étagère favorisant la frugalité	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Mettre en place une gouvernance permettant de questionner la frugalité**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse	X	X	X
1 – Initialisation			
2 – Conception et développement			
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

Proposer un catalogue de produits numériques d'IA, aussi bien de type services ou systèmes, disponibles sur étagère permet de :

- éviter la création de services d'IA déjà existants ;
- faciliter l'amélioration des produits existants afin d'éliminer plus rapidement les produits devenus obsolètes.

Pour cela, l'offre de services d'IA sera organisée dans un ordre facilitant la frugalité organisationnelle :

- faciliter la réutilisation de n'importe quel service ou système IA, avec un processus de mise en œuvre simplifiée, afin d'intégrer un produit proposant un service ou système préexistant en l'état (sans développement de fonctionnalités supplémentaires) ;
- participer à l'évolution fonctionnelle d'un produit proposant un service ou système IA sans changer la manière de l'utiliser. Pour cela il est nécessaire d'offrir certaines possibilités au demandeur :
  - accéder au backlog facilement,
  - accéder à la procédure de gestion d'une nouvelle demande (mise à jour ou évolution),
  - afficher les règles de priorisation pour toute demande;

La responsabilité de la gestion de l'offre de produits numériques IA doit être attribuée à une équipe. Lors de la mise en place d'une IA Factory, il apparaît que cette responsabilité est à lui confier.

Mise en œuvre

- Mettre en place un catalogue de produits IA accessible par tous.
- Pour mettre en place un nouveau produit numérique d'IA, proposant un service d'IA, il est nécessaire de :
  - créer un nouveau produit en suivant les bonnes pratiques de product management ;
  - offrir les différents types de développement et/ou d'intégration suivants :
    - le demandeur a la capacité de développer et/ou d'intégrer le service d'IA par lui-même (3 formats : interne, externe, hybride),
    - le demandeur souhaite faire développer ou intégrer le service ou système IA par une équipe tierce (2 sous variantes possibles : interne/externe) ;

N° 18	Avoir une offre de produits numériques IA sur étagère favorisant la frugalité	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Mettre en place une gouvernance permettant de questionner la frugalité**
○ prendre en compte les deux contextes du service ou système d'IA : ▪ celui de l'implémentation du service ou système d'IA, ▪ celui de l'utilisation du service ou système d'IA ; Ex : Pour un mailbot, de nombreuses variantes sont possibles. ○ à la fin de chaque création de service d'IA, mettre à disposition une procédure de réutilisation et de mise à jour de ce service ou système d'IA.		
Facteur clé de succès : • gestion de produit efficace qui saura identifier les limites de la mutualisation et des évolutions fonctionnelles au regard de la frugalité. Il pourrait en effet être plus intéressant sur le plan environnemental de créer deux produits plutôt que d'intégrer des fonctionnalités supplémentaires dans un produit existant.		
Sources : Témoignage		
Secteur : Multisectoriel		

BP 19

N° 19	Utiliser des jeux de données open source pour le prototypage	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Optimiser la gestion des données et l'intégrer dans une démarche d'amélioration continue (ou management de la qualité)**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et développement		X	
3 – Vérification et validation		X	
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

Rassembler et annoter des jeux de données est une tâche consommatrice de temps et de ressources. Parfois, des données semblables ou proches du cas d'usage sont à disposition de la communauté et peuvent être réemployées pour qualifier la faisabilité ou dimensionner un service.

Mise en œuvre

- Utiliser un jeu de données open source dans la mesure du possible selon les étapes du projet :
  - pour la faisabilité : les jeux de données open source étant souvent assez propres et bien annotés, ils peuvent donner une bonne estimation de la qualité maximale qu'un algorithme sera capable d'atteindre. Cela permet de choisir l'algorithme avant de lancer une campagne de collecte de données ;
  - pour le dimensionnement du service : utiliser ces données pour entraîner une première version du réseau permet aussi de régler en amont au plus juste la quantité de ressources (quantité de données, puissance de calcul …) à mettre à disposition du service avant même de réunir des données spécifiques au problème à résoudre.
- Suivre les grands portails publics de données (gouvernement ...) et les catalogues de données mis à disposition par la communauté (ex. HuggingFace).

Facteur clé de succès :

- qualité du jeu de données open source.

Sources : Témoignage

Secteur : Multisectoriel

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

BP 20

N° 20	Optimiser l'usage de l'équipement existant	
Gain de frugalité : Élevé	**Effort mise en œuvre :** Modéré	**Analyser l'impact des équipements nécessaires pour le service d'IA et optimiser l'usage (économie de la fonctionnalité)**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			X
2 – Conception et développement			X
3 – Vérification et validation			
4 – Déploiement			X
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			X

Description

L'impact environnemental de la fabrication d'un équipement correspond à plusieurs années de son utilisation, c'est pourquoi une vigilance toute particulière doit être portée à la réutilisation en priorité de l'existant avant tout achat additionnel.

Mise en œuvre

- Cf. Critère 9.5 du RGESN (Référentiel Général de l'Écoconception des Services Numériques).
- Mutualiser les équipements existants pour atteindre une utilisation maximale ou une plus haute puissance de calcul sans rachat de matériel.
- Plus une machine est puissante, plus elle pourra exécuter des calculs rapidement. Inversement, étalonner ses calculs sur des périodes plus longues (pendant le week-end ou la nuit par exemple) afin de devoir solliciter moins de puissance permet de retarder l'achat d'une machine plus puissante.

Sources : Témoignage et Bibliographie

Secteur : Multisectoriel

BP 21

N° 21	Créer un référentiel des impacts environnementaux des projets	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Outiller la mesure de l'impact environnemental et enrichir la connaissance**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse	X	X	X
1 – Initialisation			
2 – Conception et développement			
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

Établir pour chaque projet un référentiel des impacts environnementaux (prédictifs puis réellement constatés) sur toute la vie du projet afin de se familiariser sur les ordres de grandeur et établir des comparaisons d'un projet à un autre.

Mise en œuvre

- Construire un référentiel des impacts environnementaux des projets par étape de vie.
- Alimenter ce référentiel de façon anticipée puis réelle via différents outils. Alimenter ce référentiel avec l'estimation que fournissent les outils (Code Carbon, Green Algorithms, Carbon Tracker, MLCO2Impact ..), puis la réelle consommation de l'entraînement. Cela peut permettre de comparer l'estimation et la réelle consommation, ainsi que d'affiner les prédictions des futurs entraînements.

Sources : Témoignage

Secteur : Multisectoriel

BP 22

N° 22	**Favoriser les terminaux utilisateurs/salariés existants pour l'entraînement ou l'inférence du service d'IA**	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Analyser l'impact des équipements nécessaires pour le service d'IA et optimiser l'usage (économie de la fonctionnalité)**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation	X		X
2 – Conception et développement	X		X
3 – Vérification et validation			
4 – Déploiement	X		X
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			X

Description

De plus en plus de modèles peuvent être exécutés voire entraînés en local, c'est-à-dire sur des terminaux déjà à disposition dans l'entreprise ou auprès des utilisateurs finaux. Ceci n'implique pas de nouvelle fabrication, d'accès au réseau ou encore de consommation énergétique importante comme dans un centre de données. Ce modèle de déploiement doit être envisagé comme une solution alternative et doit être évalué pour challenger la frugalité du service final.

Mise en œuvre

- Identifier des petits modèles d'IA exécutables, voire entraînables, sur les terminaux salariés/utilisateurs existants.
- Mettre en place des techniques d'optimisation et de compression de modèles (voir Bonne Pratique N° 03).
- Évaluer les impacts de cette approche (voir chapitre 2 de l'AFNOR SPEC) : cette bonne pratique ne peut pas être appliquée de manière systématique mais de manière raisonnée et prouvée au regard des cas d'usage prévus, et doit donc être évaluée au même titre que toutes les autres solutions envisageables.
- Intégrer cette approche dans les réponses possibles au besoin : elle doit être challengée comme toutes les autres solutions possibles (voir Bonne Pratique N° 02 et celles relatives à la gouvernance).

Facteurs clés de succès :

- évaluation des impacts de différentes solutions dont la solution initialement envisagée ;
- identification des conditions limites de pertinence de cette solution (nombre de déploiements, consommation réseau ou électrique, etc.) : ce seront les conditions à surveiller une fois la solution déployée en production ;
- comparaison des solutions et choix raisonné ;
- monitoring des impacts des solutions déployées pour vérifier que les conditions de pertinence sont toujours rencontrées : si ce n'est plus le cas, une décision devra être prise car les impacts risquent de ne plus être sous contrôle.

Sources : Témoignages

Secteur : Multisectoriel

BP 23

N° 23	Réaliser une estimation de la consommation du modèle a priori	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Faible	**Outiller la mesure de l'impact environnemental et enrichir la connaissance**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et développement	X	X	X
3 – Vérification et validation	X	X	X
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

Des outils de simulation permettent d'anticiper l'impact environnemental d'une phase d'apprentissage d'un modèle d'IA.

Une autre approche peut consister à exécuter une phase d'apprentissage sur une petite partie du jeu de données pour obtenir des données réelles sur l'impact environnemental généré qui serviront à extrapoler l'impact environnemental sur l'ensemble de la phase d'apprentissage et sur la totalité du jeu de données.

Cette approche présente certaines limites, notamment en ce qui concerne l'extrapolation des résultats à des jeux de données hétérogènes. En effet, lorsque les données présentent une grande variabilité ou une complexité importante, il peut être difficile d'obtenir des résultats précis en extrapolant à partir d'un sous-ensemble limité de données. Il est donc important de prendre en compte ces limites lors de l'interprétation des estimations obtenues.

Il est essentiel de se concentrer sur l'influence des hyperparamètres du modèle sur la consommation d'énergie afin d'optimiser l'efficacité énergétique de l'IA frugale. En effet, le choix des hyperparamètres peut avoir un impact significatif sur la consommation d'énergie du modèle et il est donc important de les ajuster de manière à minimiser cette consommation tout en maintenant des performances acceptables. Cette optimisation peut se faire en utilisant des techniques d'apprentissage automatique et en prenant en compte les contraintes spécifiques du matériel sur lequel le modèle sera exécuté.

Mise en œuvre

- Réaliser un apprentissage sur une petite partie du jeu de données et calculer l'impact environnemental à la fin de l'apprentissage pour extrapoler l'impact environnemental que donnerait un apprentissage sur tout le jeu de données.
- Il est crucial de sélectionner soigneusement l'échantillon de données pour garantir la représentativité des résultats par rapport à l'ensemble des données. Le choix de l'échantillon dépend de plusieurs facteurs, tels que la taille de l'ensemble de données, sa complexité, sa qualité et sa variabilité. Il est recommandé d'utiliser un échantillon aléatoire pour assurer la représentativité statistique ou un échantillon stratifié si les données présentent des sous-groupes distincts. Dans certains cas, il peut être judicieux de choisir un échantillon de données plus complexe pour évaluer la performance du modèle dans des conditions difficiles. Cependant, il est important de s'assurer que l'échantillon sélectionné est représentatif de l'ensemble des données pour éviter tout biais dans les résultats.
- Pour estimer et mesurer la consommation d'énergie lors de l'apprentissage, il est possible d'utiliser divers outils tels que *CodeCarbon*, *TensorFlox Profiler* ou encore *Nvidia-smi*. *CodeCarbon* permet de mesurer l'empreinte carbone et la consommation énergétique des modèles de Machine Learning, tandis que *TensorFlow Profiler* fournit des informations détaillées sur les performances et l'utilisation des ressources du modèle. De son côté, *nvidia-smi* permet de surveiller l'utilisation de la mémoire, de la bande passante et de l'énergie des GPU Nvidia. Ces outils peuvent être utilisés conjointement pour obtenir une mesure précise et complète de la consommation d'énergie lors de l'apprentissage.

Sources : témoignage et bibliographie

Secteur : Multisectoriel

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

BP 24

N°24	Assurer la frugalité des infrastructures tout au long de l'exploitation	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Analyser l'impact des équipements nécessaires pour le service d'IA et optimiser l'usage (économie de la fonctionnalité)**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et développement			
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			X
6 – Validation continue			X
7 – Réévaluation			X
8 – Mise hors service			

Description

La consommation en ressources d'un service d'IA n'est pas forcément homogène pendant tout son cycle de vie. Si la frugalité est un objectif qui doit être pensé en amont de la mise en ligne du service, il est tout aussi important de se donner les moyens de la maintenir en exploitant ses infrastructures de manière frugale, d'autant que beaucoup de services d'IA connaissent un fort engouement à leur lancement et une utilisation au long cours beaucoup plus modeste.

Mise en œuvre

- Mutualiser et adapter les ressources au plus juste pendant la phase de lancement du service (phase souvent critique haute en termes d'utilisation de ressources) et bien poser les scénarios si les métriques opérationnelles ne sont pas disponibles.
- Monitorer l'activité de l'IA frugale, basculer dès que possible les mesures sur des données au plus près de la couche physique et définir une politique de désallocation de ressources matérielles selon l'utilisation réelle du service.
- Étudier la consommation des ressources techniques des équipements/infrastructures utilisés, sur la base de métriques réelles et opérationnelles (il arrive souvent que les indicateurs proposés soient uniquement établis d'un point de vue économique et donc non basés sur la consommation réelle). Pour cela, il est notamment opportun de faire une analyse des ressources dédiées et une analyse de chaque type de ressources mutualisées (réseau, cluster, conteneurisation ...).

Sources : Témoignage

Secteur : Multisectoriel

BP 25

N° 25	Faire évoluer les stratégies de mesure en fonction des enjeux et des contraintes pour maintenir la frugalité du service d'IA	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Outiller la mesure de l'impact environnemental et enrichir la connaissance**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et développement			
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi		X	
6 – Validation continue		X	
7 – Réévaluation		X	
8 – Mise hors service			

Description

Si les stratégies de mesure d'impact du service d'IA doivent être pensées en amont du déploiement, elles doivent faire l'objet de réévaluations tout au long du cycle de vie du service d'IA.

En effet, les risques liés aux ressources environnementales évoluent dans le temps, et des externalités positives (utilisation de la chaleur fatale des centres de données pour le chauffage collectif) peuvent par exemple devenir des contraintes en cas d'épisode caniculaire ou de stress hydrique régional. De même, les constructeurs, les hébergeurs, ou même des communautés travaillant sur le service d'IA concerné, peuvent mettre à disposition de nouveaux services gérés et des outils implémentables pour envisager davantage de frugalité dans le projet.

Mise en œuvre

- Planifier de manière régulière (pluriannuelle) :
  - la réévaluation des stratégies de mesure et des critères de frugalité ;
  - l'évolution des fonctionnalités optimisant la mesure de la frugalité via des outils de mesure ;
  - l'intégration de nouvelles fonctionnalités favorisant la frugalité dans les outils de mesure.
- Mener une veille constante sur les sujets de frugalité dans les canaux d'information de l'entreprise portant sur :
  - la rationalisation des outils ;
  - l'optimisation de l'utilisation de ses outils ;
  - l'arrivée de nouvelles fonctionnalités afin de contribuer à la frugalité.

Sources : Témoignage

Secteur : Multisectoriel

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

BP 26

N° 26	**Écrire du code pouvant être amélioré par plusieurs personnes et ré-implémenté sur plusieurs environnements**	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Optimiser la performance du modèle et l'intégrer dans une démarche d'amélioration continue**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et développement	X		
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

Le code est une partie importante dans la conception d'une IA. La façon de coder peut donc avoir un fort impact sur l'environnement. En effet, coder de façon responsable permet d'optimiser les ressources, d'améliorer la maintenabilité, l'évolutivité et l'accessibilité du modèle d'IA, de réduire les coûts associés à son développement, sa maintenance et son support technique.

Mise en œuvre

- Appliquer les règles d'écoconception. Les 3 piliers de l'écoconception sont simplicité, frugalité, pertinence. La simplicité est une démarche qualitative, alors que la frugalité est plutôt quantitative. Il s'agit de limiter les fonctionnalités et leur qualité au minimum (sobriété fonctionnelle). La pertinence, quant à elle, est une équation entre l'utilité, la rapidité et l'accessibilité.
- Optimiser le code pour qu'il soit un maximum réutilisable pour réentraîner facilement le modèle ou pouvoir utiliser des modules dans d'autres projets. Cela permet d'optimiser les ressources matérielles et de permettre la compilation multiplateforme/architecture.
- Utiliser un langage de programmation performant du point de vue environnemental : soit un langage compilé (C, C++, Cuda) - soit un interpréteur optimisé (Pythran ou Numba pour Python) pour les langages plus haut niveau (plus accessibles par exemple aux data scientists). Si cela est possible dans la limite des compétences en développement des data scientists et des choix stratégiques de la société à laquelle ils appartiennent, favoriser le langage en C pour développer l'IA avec la norme ANSI C99 et en utilisant des librairies standards.
- Utiliser des librairies appropriées pour chaque étape du cycle de vie. Ne pas développer de nouveau des outils et méthodes déjà existantes.
- Pérenniser le code :
  - utiliser un maximum les librairies courantes ;
  - vérifier les licences d'utilisation des langages (risque de portabilité et de surcoût dans le temps) ;
  - privilégier des langages multienvironnements ;
  - utiliser des langages évolutifs avec des supports opérationnels garantis (communautés de développeurs et/ou sociétés spécialisées) ;
  - vérifier la compatibilité des langages avec l'environnement cible (BDD, etc.).

Sources : Témoignage et bibliographie

Secteur : Multisectoriel

BP 27

N° 27	Rationaliser les modèles	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Optimiser la performance du modèle et l'intégrer dans une démarche d'amélioration continue**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et développement	X		
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

Pendant les phases de test et d'entraînement, il est nécessaire de trouver un équilibre entre le coût d'entraînement pour améliorer l'efficacité du modèle d'IA et les gains d'efficacité obtenus grâce à cet entraînement. Autrement dit, il s'agit de déterminer la quantité optimale d'entraînement nécessaire pour obtenir le meilleur rapport coût-efficacité pour le modèle.

Mise en œuvre

- Prendre en compte le niveau de performance d'acceptabilité du donneur d'ordre.
- Réaliser une étude approfondie afin de choisir le modèle, en adéquation avec le besoin du donneur d'ordre, ainsi que les paramètres de celui-ci, afin de limiter l'impact environnemental de la phase d'expérimentation. Par exemple, dans le cas d'utilisation de modèle de *Deep Learning*, on peut utiliser le paramètre *early stopping* afin d'arrêter l'apprentissage du modèle lorsque le gain de performance devient trop faible.
- Le nombre de fonctionnalités fournies en entrée du modèle joue également un rôle important dans la complexité des calculs. Il est donc indispensable d'utiliser des techniques de sélection de fonctionnalités n'impliquant pas le réentraînement des modèles (type MRMR, régression de Ridge, Lasso, etc.), lorsque cela est possible.
- Cf. Critères 9.2 et 9.3 du RGESN (Référentiel Général de l'Écoconception des Services Numériques).

Sources : Témoignage

Secteur : Multisectoriel

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

BP 28

N° 28	Décomposer un gros modèle d'IA en plusieurs petits modèles	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Optimiser la performance du modèle et l'intégrer dans une démarche d'amélioration continue**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et Développement	X		
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

La décomposition d'un modèle d'IA généraliste en plusieurs modèles spécialisés peut permettre de réduire l'empreinte globale du projet. En effet, l'empreinte d'un modèle d'IA généraliste est supérieure à la somme de l'empreinte des modèles spécialisés. Autrement dit, réduire la taille et la complexité de chacun des petits modèles permet de réduire les ressources nécessaires pour l'entraînement et l'inférence comme le temps de calcul, l'espace mémoire ou encore l'énergie par rapport au développement d'un modèle généraliste.

De plus, cette décomposition permet de mutualiser l'utilisation d'un modèle spécialisé dans d'autres projets.

Cette décomposition permet également de limiter l'empreinte environnementale du réentraînement et de l'inférence. En effet, il sera possible de réentraîner une partie du métamodèle au lieu de réentraîner l'ensemble de la solution. Le métamodèle est ainsi plus adaptable, plus maintenable et évolutif.

Mise en œuvre

- Traiter les cas évidents avec un modèle léger et réserver les cas moins évidents à traiter pour l'emploi d'un modèle plus lourd en entraînant le modèle lourd sur un jeu de données plus spécifique.
- Développement d'un orchestrateur de modèles permettant de diriger la demande vers le modèle spécialisé et le mieux adapté.

Facteurs clés de succès :

- besoin adapté à ce genre d'approche ;
- le délai accordé à la conception et au développement doit être suffisant pour permettre ce genre d'approche ;
- niveau de compétence suffisant pour la mise en œuvre de modèles multiples.

Sources : Témoignage

Secteur : Multisectoriel

BP 29

N° 29	**Réutiliser des algorithmes entraînés et partager les algorithmes réalisés (OpenSource)**	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Faible	**Optimiser la performance du modèle et l'intégrer dans une démarche d'amélioration continue**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation	X		
2 – Conception et développement	X		
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue	X		
7 – Réévaluation	X		
8 – Mise hors service			

Description

De nombreuses plateformes de modèles en Open Source existent et permettent de facilement trouver des modèles déjà entraînés sur des tâches spécifiques. Il est également possible de partager ses modèles au sein de la communauté.

Mise en œuvre

- De nombreux modèles Open Source sont stockés et accessibles sur la plateforme HuggingFace.
- Privilégier le *transfer learning* à la conception *from scratch.*

Sources : Témoignage

Secteur : Multisectoriel

AFNOR SPEC 2314
Référentiel général pour l'IA frugale
Une AFNOR SPEC pour mesurer et réduire l'impact environnemental de l'IA

BP 30

N° 30	Privilégier des modèles plus frugaux	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Optimiser la performance du modèle et l'intégrer dans une démarche d'amélioration continue**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et développement	X		
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue			
7 – Réévaluation			
8 – Mise hors service			

Description

Il existe de nombreuses manières d'appréhender le besoin d'un client avec de l'IA. Certains modèles d'IA sont moins énergivores et ont besoin de moins de données en entrée que d'autres modèles.

Mise en œuvre

- Mener un état de l'art des solutions existantes, de la moins complexe vers la plus énergivore.
- Évaluer si l'IA symbolique peut être un bon candidat pour répondre aux besoins. Si cela ne suffit pas, débuter l'analyse comparative par des modèles de *Machine Learning* puis de D*eep Learning :*
  - utiliser des outils sur mesure adaptés aux problèmes au lieu d'approches généralistes ;
  - cf. Chapitre 9 du RGESN (Référentiel Général de l'Écoconception des Services Numériques).

Facteur clé de succès :

- le besoin doit être construit de telle sorte que les modèles les moins énergivores soient utilisables.

Sources : Témoignage

Secteur : Multisectoriel

BP 31

N° 31	A/B Testing de modèles pour identifier le modèle avec le meilleur ratio performance/ressources	
Gain de frugalité : Modéré	**Effort mise en œuvre :** Modéré	**Optimiser la performance du modèle et l'intégrer dans une démarche d'amélioration continue**

Positionnement de la bonne pratique sur le cycle de vie

Étape du cycle de vie	Service	Données	Infrastructures
0 – Transverse			
1 – Initialisation			
2 – Conception et développement	X	X	
3 – Vérification et validation			
4 – Déploiement			
5 – Exploitation et suivi			
6 – Validation continue	X	X	
7 – Réévaluation	X	X	
8 – Mise hors service			

Description

Afin d'identifier le modèle avec le meilleur ratio performance/ressources, tester différents modèles avec un échantillon du jeu de données le plus petit possible mais représentatif de la globalité du jeu d'entraînement. Cette étape permet d'éviter de tester les différents modèles possibles sur l'ensemble du jeu de données et ainsi de limiter l'impact environnemental de la phase expérimentale.

Cette approche présente certaines limites, notamment en ce qui concerne le choix de l'échantillon représentatif du jeu de données. En effet, lorsque les données présentent une grande variabilité ou une complexité importante, il peut être difficile d'obtenir des résultats précis en extrapolant à partir d'un sous-ensemble limité de données. Il est donc important de prendre en compte ces limites lors de l'interprétation des résultats des performances. Il est bien entendu évidemment que chaque modèle doit être testé sur le même échantillon.

Mise en œuvre

- Sélection d'un échantillon représentatif du jeu d'entraînement :
  - une étude statistique des données peut permettre de faciliter la sélection de l'échantillon de données ;
  - il est crucial de sélectionner soigneusement l'échantillon de données pour garantir la représentativité des résultats par rapport à l'ensemble des données ;
  - le choix de l'échantillon dépend de plusieurs facteurs, tels que la taille de l'ensemble de données, sa complexité et sa variabilité. Il est recommandé d'utiliser un échantillon aléatoire pour assurer la représentativité statistique ou un échantillon stratifié si les données présentent des sous-groupes distincts.

Dans certains cas, il peut être judicieux de choisir un échantillon de données plus complexe pour évaluer la performance du modèle dans des conditions difficiles. Cependant, il est important de s'assurer que l'échantillon sélectionné est représentatif de l'ensemble des données pour éviter tout biais dans les résultats.

Sources : Témoignage

Secteur : Multisectoriel

Acronymes

ISO	International Organization for Standardization
AFNOR	Association Française de Normalisation
ITU	International Communication Union
ACV	Analyse de cycle de vie
RCP	Référentiel par Catégorie de Produit
UF	Unité Fonctionnelle
Ademe	Agence pour l'environnement et la maîtrise de l'énergie
Arcep	Autorité de régulation des communications électroniques, des Postes et de la distribution de la Presse
Arcom	Autorité de régulation de la communication audiovisuelle et numérique
GT	Groupe de Travail