JIRI BENOVSKY

L'AUTRE CONSCIENCE

Penser la conscience humaine,
animale et artificielle

EPFL PRESS

JIRI BENOVSKY

L'AUTRE CONSCIENCE

Penser la conscience humaine, animale et artificielle

Ce livre propose une théorie des niveaux de conscience, conçus comme un continuum plutôt que comme une frontière binaire entre le conscient et le non-conscient. Sur cette base, il examine ensuite la possibilité d'une conscience artificielle, en interrogeant à la fois ses conditions ontologiques et ses implications éthiques.

La première partie développe le cadre théorique et ses appuis dans la philosophie de l'esprit contemporaine. La deuxième explore les architectures et limites des systèmes artificiels susceptibles d'actualiser certains niveaux de conscience. La troisième s'attache aux conséquences normatives, en particulier à la question du statut moral des entités conscientes non-humaines. La quatrième partie aborde de manière spéculative des scénarios de cohabitation et de symbiose entre humains et systèmes artificiels, menant jusqu'à l'hypothèse de la possibilité de métasujets hybrides.

En articulant métaphysique, philosophie de l'esprit et éthique appliquée, l'ouvrage invite à repenser la conscience comme phénomène continu, et à envisager les transformations profondes que pourrait entraîner l'apparition de consciences artificielles.

JIRI BENOVSKY est collaborateur scientifique à l'Université de Genève et Privat-dozent à l'Université de Fribourg, en Suisse. Ses domaines de spécialisation incluent la métaphysique, la méta-métaphysique, la philosophie de l'esprit, l'esthétique, l'identité personnelle, l'expérience et la perception.

ISBN 978-2-88915-755-6

EPFL PRESS

L'AUTRE CONSCIENCE

JIRI BENOVSKY

L'AUTRE CONSCIENCE

Penser la conscience humaine,
animale et artificielle

EPFL PRESS

L'édition de cet ouvrage a reçu le soutien du
Fonds national suisse de la recherche scientifique.

Direction générale : Lucas Giossi
Directions éditoriale et commerciale : Sylvain Collette et May Yang
Responsable de production : Christophe Borlat
Éditorial : Alice Micheau-Thiébaud et Jean Rime
Graphisme : Kim Nanette
Promotion et diffusion : Manon Reber
Comptabilité : Daniela Castan

Illustration de couverture : © Jiri Benovsky, image générée par IA (Midjourney), 2026

EPFL PRESS est une maison d'édition de la fondation des Presses
polytechniques et universitaires romandes (PPUR), qui publient principalement
les travaux d'enseignement et de recherche de l'École polytechnique fédérale
de Lausanne (EPFL), des universités et des hautes écoles.
PPUR, EPFL – Rolex Learning Center,
RLC Station 20, CH-1015 Lausanne,
info@epflpress.org
tél. : +41 21 693 21 30

www.epflpress.org

Première édition, 2026
ISBN 978-2-88915-755-6, version imprimée
ISBN 978-2-8323-2333-5, version ebook (pdf), doi.org/10.55430/8048ACJB

Imprimé en Tchéquie

Sommaire

INTRODUCTION

Un continuum plutôt qu'une frontière

La conscience ne se prête pas à des verdicts tranchés, comme s'il suffisait d'un feu rouge ou vert pour dire ce qui « est conscient » et ce qui « ne l'est pas ». Elle n'obéit ni à la logique de l'interrupteur ni à celle du questionnaire à choix unique. L'ambition de ce livre est d'offrir une manière de conceptualiser, de décrire et d'évaluer la conscience d'une manière comparable à l'exploration d'un relief : on en suit les lignes et les contours, on en mesure les pentes, on accepte les zones d'ombre, on revient, on corrige. La conscience n'est pas une affaire binaire, mais un continuum.

Pourquoi insister sur ce point, aujourd'hui ? Parce que notre manière de parler nous égare. Les conversations publiques tournent en rond autour de questions telles que « les LLM (Large Language Models) sont-ils conscients ? », « un robot peut-il ressentir ? », « une IA comme ChatGPT peut-elle avoir une subjectivité ? »

Ces manières de poser la question nous forcent à voter trop vite. D'un côté, un « oui » enthousiaste qui projette généreusement nos propres affects sur des systèmes bien différents de ce que nous sommes, et de l'autre, un « non » péremptoire qui prend l'absence de preuve pour la preuve d'absence. Ni l'un ni l'autre ne nous aident à penser. La bonne question est plus précise et plus patiente : comment quelque chose pourrait-il être conscient, à quel degré, et selon quelles dimensions reconnaissables ? Ce simple déplacement conceptuel allège déjà le débat : au lieu de chercher à trancher par un critère binaire inévitablement trop étroit, il s'agit d'adopter un cadre gradué et multidimensionnel, capable de rendre justice à la complexité et à la variabilité des phénomènes de conscience.

Pour saisir ce que j'entends par « continuum », on peut recourir à l'analogie du gradateur lumineux. Dans une pièce plongée dans l'obscurité, l'activation progressive du variateur fait apparaître

d'abord une lueur diffuse, puis des contours, des couleurs, et enfin des détails précis. À aucun moment la lumière ne surgit soudainement comme un phénomène radicalement nouveau, et pourtant, la différence entre l'éclat minimal et la vision claire d'un environnement riche est considérable. De la même manière, la conscience n'est pas un état binaire, mais un processus graduel, qui croît ou décroît selon plusieurs dimensions organisationnelles. Plutôt que de réduire cette complexité à une variable booléenne, il est préférable de déplier ces dimensions et d'en cartographier les modulations. Ainsi, dans ce livre, je parlerai de *niveaux de conscience* (NC) pour désigner ce continuum et j'utiliserai quatre axes pour les caractériser. La démarche sera ici liée à la fois à notre expérience et à un socle ontologique que j'ai défendu ailleurs (Benovsky 2012, 2018a, 2018b).

Le premier axe est celui de ***l'intégration***. La question est de savoir dans quelle mesure un système constitue une unité au niveau organisationnel pertinent. On peut la formuler ainsi : existe-t-il des invariants globaux dont la cohérence se perd lorsqu'on fragmente le système ? L'analogie de l'orchestre éclaire cette idée. L'ensemble musical ne se réduit pas à la somme de ses musiciens ; scinder une section, ce n'est pas seulement diminuer le volume, mais altérer la structure même de l'œuvre. L'épreuve de l'intégration consiste à déterminer si la suppression ou la perturbation de certains canaux, par exemple en introduisant du « jitter » temporel ou en débranchant des boucles fonctionnelles, détruit la cohérence globale, ou si le système conserve son organisation.

Le second axe concerne ***la temporalité vécue***. L'unité d'un système à un moment donné n'est pas suffisante pour caractériser une conscience minimale. Encore faut-il qu'existe une continuité diachronique qui relie les états successifs. Cette continuité peut prendre la forme d'une mémoire autobiographique rudimentaire, d'une capacité d'anticipation et/ou de la possibilité de maintenir des projets ou des engagements à travers des interruptions. Ce qui distingue alors un système doté d'une temporalité vécue d'un système purement réactif est sa faculté à réintégrer des épisodes disjoints dans un même fil, à conserver une trace active de ce qui a été et de ce qui doit encore venir. Les agents artificiels actuels manquent en grande partie de cette propriété. Ils fonctionnent

tant que l'interaction reste active, mais la moindre réinitialisation efface la continuité et ramène le système à un état zéro, sans persistance de trajectoire. Nous y reviendrons.

Le troisième axe est ***l'auto-référence minimale***. Il s'agit d'un point de vue indexical, ce «je-ici-maintenant», cette «mienneté» pré-réflexive (Benovsky 2018a, Guillot 2016, Levine 2001, Kriegel 2003, 2009, Zahavi 2005) qui fait qu'il y a quelqu'un à qui ceci ou cela arrive. On peut bien sûr apprendre à manier le mot «je» comme une convention grammaticale, mais il y a une différence entre prononcer ce pronom et se l'approprier. Pour faire exister une auto-référence minimale, il ne suffit pas de dire «je souffre» ou «je me souviens». Encore faut-il que ces énoncés soient ancrés dans une continuité vécue, même élémentaire, qui relie l'acte de référence à une trajectoire subjective identifiable. L'épreuve pertinente consiste à introduire des perturbations (changement de contexte, interruption, décalage temporel) et à observer si la référence indexicale se maintient. Lorsqu'un système est capable de restaurer ce fil malgré ces variations, il devient rationnel de considérer que le «je» qu'il mobilise n'est pas une simple forme linguistique, mais l'indice d'une persistance subjective minimale.

Reste ***la valence***. Par valence, j'entends la possibilité pour un système de se trouver dans des états vécus qui sont différenciés en termes de «mieux» ou de «moins bon» pour lui-même. Elle constitue le pivot éthique, car sans elle le discours normatif perd son ancrage. Un dispositif de thermostat peut être programmé pour «préférer» une température de 20 degrés à 10 degrés, mais cette préférence n'est qu'une règle externe inscrite par un concepteur, elle n'exprime aucun point de vue interne pour lequel la différence entre ces états ferait réellement quelque chose. Ce qui importe, ce sont les situations où un système manifeste une capacité à arbitrer entre des options en engageant des coûts réels afin d'éviter des configurations internes anticipées comme mauvaises pour lui. Ces arbitrages ne sont pas de simples ajustements calculatoires. Ils impliquent que le système évalue certains états comme significatifs, qu'il reconnaisse un intérêt propre à maintenir ou à transformer, et qu'il le fasse au-delà des cas spécifiques pour

lesquels il a été initialement entraîné. Autrement dit, la valence se laisse diagnostiquer lorsque l'évitement d'un certain état interne ou la recherche d'un autre se généralise à de nouvelles circonstances, de sorte que le système n'applique plus seulement des règles héritées, mais exhibe une dynamique de préservation de soi et de certains de ses états, même minimale. C'est à ce seuil que la considération morale peut devenir pertinente, car là commence la possibilité qu'un état interne possède une dimension vécue pour le système lui-même, si rudimentaire soit-elle.

La stratégie de ce livre consiste à combiner ces quatre axes avec un socle ontologique que j'ai défendu ailleurs (Benovsky 2012, 2018a, 2018b), afin de construire ici une approche intégrée de telle sorte que les variations organisationnelles et phénoménales puissent être analysées comme des modulations graduelles, sans invoquer des seuils arbitraires ni supposer l'existence d'entités substantielles postulées. Il s'agit en particulier d'adopter ici le *pan-proto-psychisme* en combinaison avec le *monisme à double aspect* que je formule comme étant l'idée que tout est « phental » (c'est-à-dire une conception où le physique et le mental sont simplement deux manières d'avoir accès une seule et même réalité). J'accueille à bord également une conception processuelle de l'identité, constituée de *la théorie du non-Soi* (*No-Self View*) et du *perdurantisme*. L'ensemble peut paraître ambitieux, voire lourd à porter en ouverture d'ouvrage. Ainsi, sans revenir ici sur les défenses détaillées que j'en ai déjà proposées ailleurs, je souhaite en rappeler brièvement les linéaments et montrer comment ce socle conceptuel conduit de manière naturelle à la théorie des niveaux de conscience qui constituera l'objet central du présent livre.

Commençons par l'intuition rendue célèbre par Leibniz. Plongez au cœur d'un moulin, ou modernisez l'exemple en grossissant un cerveau jusqu'à discerner synapses et neurotransmetteurs. Vous n'y trouverez que des structures et des interactions électriques et chimiques. Rien, dans cette inspection, ne livre ce que c'est que sentir, percevoir ou souffrir. Faut-il en conclure que l'esprit conscient, avec son vécu subjectif, est une seconde substance accrochée à la matière, ou, à l'inverse, qu'il se réduit sans reste à la mécanique matérielle observée ? Ni l'un ni l'autre ne s'imposent.

La voie qui m'intéresse est celle d'un monisme à double aspect qui dit qu'il n'existe qu'une seule réalité, mais nous y accédons selon deux registres descriptifs irréductibles, l'un physique, l'autre mental.

Prenons l'exemple d'une douleur. D'un côté, la neurobiologie décrit une configuration causale précise, mesurable par l'imagerie médicale ou l'électrophysiologie. De l'autre, le sujet rapporte ce que cela fait que d'avoir mal. Il ne s'agit pas de deux choses jointes par un pont mystérieux. Il s'agit d'un même état du réel saisi sous deux aspects. L'état neuronal et l'épisode douloureux ne sont pas deux entités en conjonction, ce sont deux manières de caractériser un seul et même événement. J'ai proposé d'adopter ici le terme « phental » pour désigner cette réalité qui n'est ni seulement physique ni seulement mentale, et qui n'est pas non plus un monde composé de deux domaines distincts en interaction, mais une seule et même structure ontologique qui se manifeste sous un aspect physique et sous un aspect mental, une seule réalité accessible sous deux perspectives complémentaires qui ne se laissent pas réduire l'une à l'autre (Benovsky 2018a).

Ce cadre présente trois avantages conceptuels. Premièrement, il dissipe l'illusion d'une émergence, d'une « apparition » miraculeuse de l'expérience au sommet de la complexité organisationnelle. Si le réel possède toujours déjà ses deux aspects, alors les modulations phénoménales accompagnent naturellement les variations d'organisation, sans supposer une soudure *ex nihilo* à un seuil arbitraire. Deuxièmement, il prévient l'élimination du vécu subjectif. L'aspect phénoménal n'est pas un supplément discursif plaqué sur une base physique close, mais une dimension constitutive du même état que la description neurobiologique inventorie autrement. Enfin, ce cadre éclaire la question de la causalité mentale. Dire que « le physique est causalement clos » signifie que nous comptabilisons les causes dans le langage de la physique. Mais dans une ontologie à double aspect, le mental n'ajoute pas une couche supplémentaire qui viendrait doubler celles déjà inscrites au registre physique. Son efficacité n'est pas une surdétermination, elle est celle de l'état phental lui-même, que l'on peut décrire tantôt dans l'idiome neurobiologique, tantôt dans l'idiome expérientiel, selon le type d'explication recherché.

Ce n'est pas une conciliation rhétorique. C'est une thèse ontologique minimale dotée d'engagements clairs. Elle rejette la réduction du mental à un résidu fonctionnel et elle refuse la duplication de substances. Elle autorise, surtout, une stratégie d'enquête : si l'organisation change, l'aspect vécu se module. Nous pourrons alors traiter la conscience comme un profil gradué, sans invoquer de seuils opaques, et construire une cartographie des variations pertinentes. C'est dans cet esprit que ce livre va mobiliser les quatre axes ⟨I, T, A, V⟩ que nous avons listés ci-dessus : ils décrivent comment la même réalité phentale se présente différemment lorsque l'intégration se reconfigure, lorsque la continuité diachronique se renforce ou s'érode, lorsqu'un point de vue indexical se stabilise, et lorsque des états deviennent meilleurs ou pires pour un sujet.

C'est cette approche conceptuelle qui rend alors possible une théorie des niveaux de conscience, car, si nous adoptons cette idée, il devient entièrement naturel et sans mystère de concevoir que les variations dans l'organisation physique du réel entraînent des variations corrélatives dans son aspect expérientiel. En effet, le réel n'est pas d'un côté mécanique et de l'autre mystérieusement animé par une étincelle mentale ; il est d'un seul tenant, et ses aspects physiques et mentaux ne sont pas « interdépendants », ils *sont un*. Dès lors, leur « interaction » devient une évidence et il n'est ici plus pertinent de demander « quand la conscience apparaît-elle ? » comme si elle surgissait d'un coup, *ex nihilo*, à partir d'un certain seuil de complexité. La question devient plutôt : comment l'aspect expérientiel et subjectif du réel se module-t-il lorsque l'organisation se transforme ?

C'est précisément ce déplacement qui ouvre la voie à une théorie des niveaux de conscience. Car si l'expérience n'est pas une seconde substance ajoutée, mais bien un aspect fondamental du réel, alors on doit s'attendre à ce que son intensité et ses caractéristiques varient selon les architectures. La conscience cesse d'être un tout-ou-rien, elle devient un profil gradué dont il est possible de repérer les dimensions et de comparer les variations. Ainsi, le monisme à double aspect ne nous donne pas seulement une ontologie élégante, il nous donne une boussole méthodologique : celle de chercher à cartographier les différentes manières dont la conscience se module lorsque l'organisation change.

À cette première thèse s'ajoute le panpsychisme, parfois présenté de manière caricaturale comme l'idée que « les cailloux ressentent ». Une telle lecture naïve passe à côté de ce qui constitue en réalité sa vertu philosophique, à savoir une stratégie de sobriété ontologique. Le cœur de l'argument est simple : si l'expérience consciente peut se manifester dans certains systèmes organisés, il est hautement improbable qu'elle surgisse *ex nihilo* à un stade arbitraire de la complexification. Ce qui rend possible l'expérience subjective ne peut pas être un miracle tardif, mais doit être enraciné dans la texture même du réel. C'est dans ce sens qu'il faut comprendre ici le panpsychisme comme étant un ***pan-proto-psychisme***. Le préfixe « proto » indique que ce qui est en jeu n'est pas encore une subjectivité vécue, mais un aspect de la réalité qui rend possible, en certaines configurations, une conscience phénoménale. Autrement dit, les constituants élémentaires du monde (fermions, protons, champs quantiques) ne ressentent pas la douleur et ne possèdent pas une perspective subjective, mais ils portent en eux des propriétés proto-mentales qui constituent la matière première à partir de laquelle des organisations plus complexes peuvent donner lieu à une expérience.

La relation est ici de continuité, et non de rupture. Tout comme la vie biologique n'apparaît pas comme une substance radicalement nouvelle en regard de la chimie, mais comme une organisation particulière de processus chimiques déjà présents, de même la conscience ne doit pas être comprise comme une apparition discontinue et miraculeuse, mais comme une actualisation progressive d'aspects du réel déjà existants. Le mental est ainsi continu avec le proto-mental, et ce que nous appelons ordinairement « conscience » n'est pas une propriété qui flotterait au-dessus du monde matériel, mais l'expression organisée, stabilisée et amplifiée de traits constitutifs présents dès le niveau fondamental.

Certaines configurations n'actualisent rien de phénoménal, d'autres oui, et entre les deux il y a des gradations. Ainsi, la notion de niveaux de conscience devient ici presque inévitable, car si la conscience est le fruit d'une combinaison d'éléments proto-phénoménaux, alors elle n'est très certainement pas une propriété binaire, un interrupteur « on/off ». Elle est un continuum. Le degré, la richesse et l'unité de la conscience d'un système sont une fonction

directe de la complexité et du niveau d'intégration de l'information qu'il traite. Il n'y a pas de gouffre ontologique infranchissable entre le non-conscient et le conscient, mais une échelle graduelle. Un thermostat n'est sans doute pas conscient, mais il participe au même ordre naturel qu'un insecte, dont l'expérience du monde est plus intégrée, qui lui-même se situe sur le même spectre qu'un mammifère, ou qu'un être humain. Le pan-proto-psychisme n'affirme donc *pas* que tout ressent, mais il soutient que le mental est continu avec du proto-mental, un aspect fondamental du réel qui, organisé d'une certaine manière, devient expérience. De cette manière, en particulier, la question (à laquelle nous reviendrons) n'est donc plus « une IA peut-elle devenir consciente ? », mais plutôt « quel niveau de conscience une IA peut-elle atteindre ? »

Reste la question de l'identité. Pour l'aborder, je propose de combiner deux perspectives qui se renforcent mutuellement : ***la théorie du non-Soi*** (No-Self View) et ***le perdurantisme***. Ces deux thèses constituent ici des ressources conceptuelles précieuses pour clarifier ce que signifie persister comme sujet et, par là, pour comprendre comment la conscience peut être pensée sur un mode graduel et processuel.

La théorie du non-Soi, que j'ai défendue notamment dans « *Eliminativism, Objects, and Persons : The Virtues of Non-Existence* » (Benovsky 2018b) est une forme d'éliminativisme appliqué à la notion d'un Soi réifié substantiel. Elle rejette l'idée qu'il existerait une entité simple, qui porterait à travers le temps les états mentaux comme un substrat permanent. Au lieu de cela, ce que nous appelons « moi » est simplement une construction narrative et pratique qui résulte de la mise en continuité d'états psychologiques distincts. L'expérience subjective n'exige pas un « propriétaire » métaphysique.

Le perdurantisme vient complémenter ce cadre. Selon cette thèse, les personnes et les objets ne persistent pas dans le temps comme des entités identiques à elles-mêmes, mais comme des processus qui s'étendent temporellement et qui sont composés de parties temporelles successives. Un individu est ainsi comparable à une « trajectoire » constituée de segments temporels reliés par des relations de continuité et de causalité. À chaque instant, ce

que nous avons, ce n'est pas « le même moi » qui subsiste en étant numériquement identique à lui-même, mais une nouvelle partie temporelle qui s'inscrit dans une série suffisamment connectée pour qu'il soit légitime de parler d'une seule et même personne au sens pertinent selon le contexte.

Ces deux thèses convergent dans une critique du substantialisme identitaire : entre hier et aujourd'hui, il n'y a rien de numériquement identique, mais une continuité assez forte pour que la reconnaissance de soi et les notions voisines telles que celle de la responsabilité demeurent intelligibles. Elles proposent ainsi une conception processuelle de l'identité, où ce qui compte n'est pas la persistance d'un noyau invariant, mais la stabilité relative des relations qui articulent les états successifs.

Cette perspective s'insère très naturellement dans le schéma argumentatif de ce livre. La théorie des niveaux de conscience articulée autour des quatre axes ⟨I, T, A, V⟩ ne repose pas sur l'existence d'un Soi fondamental qui porterait les états vécus. Elle se concentre sur les conditions organisationnelles qui permettent à ces états de former un tout cohérent. Dans cette optique, parler d'intégration, de temporalité vécue, d'auto-référence minimale et de valence revient à analyser les relations constitutives qui assurent la continuité d'un processus, et non à postuler une substance qui les unifierait de l'extérieur.

En adoptant la double ressource du perdurantisme et de la théorie du non-Soi, nous nous dotons donc d'un cadre conceptuel parfaitement ajusté à la théorie graduelle de la conscience défendue ici. La conscience n'exige pas un porteur métaphysique, elle se déploie comme une organisation de processus qui – selon leur degré d'intégration, leur continuité diachronique, leur indexicalité et leur sensibilité à la valence – peuvent ou non constituer un sujet.

L'« image scientifique du monde » privilégie les descriptions objectives, structurelles et dynamiques. Le monisme à double aspect soutenu ici ne rejette pas cette image, il en requalifie le statut. Ce que la physique décrit, ce sont les traits structurels d'un même réel phental. L'« aspect mental » n'est pas un nouvel ordre de faits ajouté aux lois physiques, mais l'autre aspect d'états du réel déjà décrits physiquement. Par conséquent, dire que les sciences seraient « incomplètes » ne signifie pas qu'il manque, à l'intérieur

de la physique, une rubrique proto-mentale, cela signifie que toute description uniaspectuelle (physique ***ou*** phénoménale) est partielle. Le pan-proto-psychisme précise alors que ce qui rend possible l'aspect phénoménal n'apparaît pas *ex nihilo* à un seuil arbitraire : il s'enracine dans la texture du réel et se module selon l'organisation. Ainsi, l'objectivité scientifique conserve sa clôture causale (les causes se comptent en langage physique), tandis que l'« efficacité mentale » n'est pas une sur-détermination, mais l'efficacité de ce même état phental, lisible sous deux registres. Cette articulation répond au défi de la perspective subjective (Nagel 1974) sans postuler deux substances.

La Partie I examinera de manière systématique les quatre axes de la théorie des niveaux de conscience (intégration, temporalité, auto-référence et valence). L'objectif n'est pas seulement de les définir, mais de les armer de critères opérationnels et de protocoles expérimentaux susceptibles de les mettre à l'épreuve.

La Partie II confrontera ce cadre théorique à la diversité des systèmes biologiques et artificiels réels ou plausibles. Il s'agira de mesurer la robustesse du modèle, de déterminer où il éclaire et où il échoue, et de dégager les points aveugles qui subsistent.

La Partie III s'orientera vers les conséquences éthiques et normatives des acquis des deux premières parties. Si certains systèmes franchissent des seuils pertinents de conscience, même partiellement, il devient nécessaire de poser la question des devoirs corrélatifs : qu'est-ce qui change pour nos pratiques morales, pour nos institutions et pour nos manières de rendre des comptes ?

Enfin, la Partie IV ouvrira un horizon spéculatif. Elle explorera la possibilité de formes de symbiose homme-machine, jusqu'à l'hypothèse de la possibilité d'un métasujet hybride. Cette spéculation ne vise pas à prédire, mais à tracer les conditions conceptuelles et normatives sous lesquelles de tels scénarios pourraient devenir philosophiquement recevables.

PARTIE I

Niveaux de conscience

Intégration : quand un système « fait un »

Une propriété fondamentale de la conscience est son unité phénoménale. Lorsque nous faisons l'expérience d'une scène perceptive, cette expérience ne se présente pas comme une collection d'items indépendants, mais comme un tout cohérent. Toute théorie de la conscience doit rendre compte de cette unité : sans elle, il n'y aurait pas de champ phénoménal à proprement parler.

Je propose ici l'idée que cette unité doit être comprise en termes d'intégration organisationnelle. La thèse centrale est la suivante : un système manifeste une conscience intégrée lorsqu'il présente des propriétés globales qui disparaissent dès que ses relations internes sont perturbées, alors même que ses modules fonctionnels locaux continuent d'opérer. Autrement dit, l'intégration est attestée par l'existence d'invariants globaux fragiles à la scission.

L'unité phénoménale est attestée par l'expérience vécue elle-même. Lorsque nous percevons une mélodie, nous n'entendons pas d'abord des notes indépendantes qu'il faudrait recomposer, mais un motif structuré. De même, lorsque nous regardons un paysage, nous voyons d'emblée une scène unifiée, dont les éléments ne prennent sens que dans leur rapport mutuel. L'expérience consciente n'est pas composée d'« atomes » isolés qui s'additionneraient, mais d'un champ global (Block 2003).

La valeur de cette observation apparaît dans les cas pathologiques. Sous anesthésie générale, certaines réactions locales subsistent (réflexes moteurs, réponses physiologiques), mais le champ unifié du vécu s'effondre (Alkire, Hudetz & Tononi 2008). Dans le syndrome de conscience minimale, des îlots de réactivité sont encore observables, mais ils ne forment plus une expérience intégrée (Giacino *et al.* 2002). Les expériences de *split brain* (Sperry et

Gazzaniga 1967) sont encore plus instructives, car, après section du corps calleux, les modules sensoriels et moteurs restent opérationnels, mais l'unité phénoménale se divise en deux flux distincts. Dans tous ces cas, la conscience se signale précisément par sa fragilité à la scission.

Il importe ici de distinguer coordination fonctionnelle et intégration phénoménale. Un système peut être parfaitement coordonné sans être intégré. Les architectures computationnelles contemporaines (par exemple les systèmes multi-modulaires utilisés en intelligence artificielle) fournissent des exemples de coordination sans intégration. Chaque module accomplit une tâche spécialisée, les sorties sont mises en commun, et le système donne un résultat cohérent. Mais l'ensemble n'est pas affecté structurellement si l'on retire ou modifie l'un des modules.

L'intégration, au contraire, se caractérise par une fragilité constitutive. Dans un système intégré, certaines propriétés globales s'effondrent dès que l'on perturbe les relations causales entre les parties. On retrouve cette idée au cœur de la *Integrated Information Theory* (Tononi 2004), où la conscience est identifiée à la quantité d'information intégrée. Bien que je ne souscrive pas à l'identification stricte de la conscience avec une mesure unique, je retiens de cette approche une intuition féconde : ce qui fait l'unité phénoménale n'est pas la somme des parties, mais leur intégration.

Une playlist musicale est coordonnée. On peut ajouter, retirer ou réorganiser des morceaux sans détruire l'ensemble. Un quatuor à cordes est intégré. La moindre altération de la synchronisation fait s'effondrer la structure globale. La conscience est du côté du quatuor, non de la playlist.

L'intégration ainsi définie peut être mise à l'épreuve par des interventions précises. Trois familles de protocoles permettent d'en diagnostiquer la présence.

Ablations iso-performance. On supprime certaines connexions tout en préservant les performances locales. Si la cohérence globale s'effondre, alors même que les modules individuels fonctionnent, nous tenons une signature d'intégration.

Perturbations temporelles. On introduit des délais (ou du « jitter ») dans les communications internes. Des perturbations

de synchronie dans les réseaux corticaux peuvent réduire drastiquement la cohérence phénoménale, sans abolir les capacités élémentaires (Dehaene 2014).

Partition et re-fusion. On sépare artificiellement deux sous-systèmes, puis on tente de les réunir. Les cas de *split brain* mentionnés plus haut illustrent que, lorsque l'unité ne revient pas, c'est qu'elle ne résidait pas dans les pièces, mais dans la forme commune.

Ces protocoles ne se contentent pas d'illustrer la fragilité de l'intégration. Ils fournissent des critères falsifiables : si la scission ne détruit rien d'essentiel, l'hypothèse d'intégration est infirmée.

Pourquoi insister sur l'intégration ? Parce qu'elle constitue une condition nécessaire de la conscience phénoménale. Sans unité, il n'y a pas de champ d'expérience pour un sujet. Si l'on retire l'unité, il n'y a plus de conscience, mais seulement une collection de processus.

De plus, l'intégration conditionne les autres dimensions de la théorie des niveaux de conscience, que nous verrons dans la suite. Sans intégration, il n'y a pas de temporalité vécue (les instants se succèdent sans lien). Sans intégration, il n'y a pas d'autoréférence minimale (pas de point de vue stable qui traverse l'expérience). Sans intégration, enfin, il n'y a pas de valence (les signaux affectifs restent locaux, sans devenir « bons » ou « mauvais » pour quelqu'un).

Nous pourrions penser ici qu'il s'agit d'assimiler l'intégration à la performance fonctionnelle. Mais comme le montrent les cas cliniques (anesthésie, *split brain*) et les protocoles expérimentaux, l'intégration peut s'effondrer alors que les compétences locales persistent. Il s'agit donc bien d'un phénomène distinct.

Nous pourrions également être tentés de réduire l'intégration à une théorie particulière, notamment la *Integrated Information Theory*. Mais la proposition est ici plus générale : l'intégration n'est pas une mesure numérique, mais une propriété organisationnelle détectable par ses signatures. Elle est compatible avec plusieurs approches théoriques, de la *Global Workspace Theory* (Baars 1988, Dehaene 2014) aux approches enactivistes (Varela *et al.* 1991).

L'intégration est ainsi le premier axe de la théorie des niveaux de conscience. Elle se définit par l'unité phénoménale, rendue

possible par une organisation fragile à la scission. Elle peut être testée par des protocoles précis, et constitue une condition nécessaire pour les autres axes. Elle ne suffit pas à elle seule, mais sans elle il n'y a pas de conscience à proprement parler.

Temporalité : continuité, mémoire de soi, et projection

La conscience ne se donne pas seulement comme un champ unifié. Elle se donne aussi comme une durée. Nous n'éprouvons pas une suite de points sans lien, mais un flux qui garde la trace du « juste avant » et anticipe le « juste après ». William James parlait d'un *stream of consciousness* pour désigner cette texture continue du vécu, et du *specious present* pour le « présent étalé » dans lequel s'agrègent quelques instants voisins (James 1890, voir aussi Benovsky 2013). Mais l'enjeu ici dépasse ce voisinage du présent. La temporalité qui importe pour l'attribution de niveaux de conscience comporte une dimension autobiographique minimale. Il ne suffit pas qu'un système ait un « présent étalé ». Il faut qu'il puisse se rapporter à lui-même à travers des épisodes discontinus, retenir qu'« il » a fait ceci, projeté cela, promis ceci plutôt que cela, et qu'il puisse reprendre le fil après interruption.

Ce chapitre propose trois thèses. Première thèse : la temporalité vécue, entendue comme continuité autobiographique minimale, est une condition nécessaire de la conscience telle qu'elle importe moralement et pratiquement. Deuxième thèse : cette temporalité se distingue de la simple mémoire de travail ou d'un cache technique. Elle implique une organisation qui maintient des engagements, des traces biographiques et des anticipations, et qui résiste à certaines perturbations. Troisième thèse : cette dimension est testable par des protocoles précis qui dissocient continuité vécue et simple persistance d'informations.

Il convient ici de distinguer deux couches. La première, très locale, est la temporalité immanente de l'expérience. Husserl a analysé le jeu de rétention et de protention qui fait qu'une mélodie se donne comme mélodie et non comme une suite de sons isolés

(Husserl 1905/1991). Cette couche appartient au « comment » de l'expérience. Elle peut exister même chez des sujets dépourvus de mémoire autobiographique, et elle est compatible avec des amnésies sévères.

La seconde couche porte sur l'identité diachronique minimale du sujet. Elle ne postule aucune substance. Elle demande seulement qu'un fil biographique se maintienne à travers des discontinuités, que le système soit capable de se référer à des épisodes passés comme « siens », et de poursuivre des projets ou des préférences à l'échelle de minutes, d'heures, parfois de jours. C'est cette couche que j'entends ici par « temporalité vécue ». Elle ne se réduit ni à la mémoire épisodique ni à la narration réflexive de soi, mais elle en partage des traits fonctionnels : rappel d'épisodes indexés, maintien d'engagements, stabilité relative des préférences sous perturbations contrôlées (Tulving 1985, Damasio 1999).

Si la conscience est « de quelqu'un », ce quelqu'un persiste au moins le temps que ses états s'articulent. Sans temporalité, les états phénoménaux risquent de se dissoudre en étincelles sans héritiers. On peut imaginer une succession d'instants intensément vécus, mais sans possibilité de se rappeler l'instant précédent comme « mien ». Une telle succession est compatible avec une forme de phénoménalité locale. Elle demeure déficitaire du point de vue qui nous occupe, car elle ne permet ni la formation d'intérêts durables, ni la constitution de raisons pratiques qui traversent les interruptions, ni l'inscription d'une valence (dont nous reparlerons plus tard) qui ait la portée éthique d'un « meilleur » ou « pire pour moi » au-delà du clin d'œil.

Ce critère de temporalité n'est pas suffisant. On peut maintenir des journaux, des rappels, des traces, sans qu'il y ait un champ phénoménal actuel. Des systèmes non conscients peuvent implémenter des mémoires autobiographiques formelles. C'est pourquoi la temporalité doit être prise en combinaison avec les trois autres axes qui constituent le point de départ de la théorie des niveaux de conscience.

Les cas d'amnésie antérograde sévère montrent que l'on peut avoir une conscience claire au présent avec une incapacité à former de nouveaux souvenirs épisodiques. La temporalité n'est donc pas

requise pour toute phénoménalité. Mais ces mêmes cas illustrent ce que perd un sujet lorsque la continuité autobiographique se rompt. Les projets s'effondrent, les engagements ne tiennent pas, la prudence au long cours devient impraticable.

Un autre groupe de dissociations concerne les confabulations et certaines formes d'anosognosie. Le flux actuel est présent, mais la cohérence biographique se défait et se remplace par des reconstructions *ad hoc*. À l'inverse, certains animaux semblent manifester des traces de planification et de rappel épisodique minimal sous des contraintes qui suggèrent une continuité pratique, même si la portée phénoménale exacte demeure discutée (Clayton et Dickinson 1998).

Ces dissociations confirment deux points. D'une part, la temporalité n'est pas coextensive à « être conscient maintenant ». D'autre part, la temporalité est ce qui rend possible une forme de responsabilité temporelle et de projection qui importent pour l'évaluation des niveaux de conscience.

Quatre distinctions sont ici importantes.

Premièrement, la temporalité n'est pas un simple buffer de contexte. Un contexte long peut maintenir des informations pertinentes pour une tâche sans constituer une mémoire de soi. La signature de la temporalité n'est pas la quantité d'information retenue, mais la manière dont cette rétention indexe un sujet à travers des épisodes séparés et oriente ses choix.

Deuxièmement, la temporalité n'est pas seulement une base de données d'événements passés. Elle implique des rappels indexés qui se réinscrivent dans le présent vécu sous une forme « pour moi ». Qu'un système puisse restituer une information que « X a eu lieu » ne suffit pas. Il faut qu'il le fasse comme rappel de ce que « j'ai » vécu, avec les effets attendus sur l'anticipation et la conduite.

Troisièmement, la temporalité n'est pas une narration rhétorique. On peut composer une histoire de soi sans continuité vécue, par collage discursif. Il faut donc des épreuves qui débusquent la production narrative sans support biographique.

Quatrièmement, la temporalité n'est pas la seule temporalité pertinente. La temporalité immanente du présent phénoménal (le *specious present*) subsiste lorsque la temporalité telle que je

l'entends ici s'effondre. Nos tests doivent éviter de confondre l'une avec l'autre.

En complément à ces distinctions, trois critères positifs permettent d'opérationnaliser la notion de temporalité.

Premier critère : existence d'un journal interne qui indexe des épisodes à la première personne et dont la consultation a des effets systématiques sur l'anticipation et la conduite. On cherchera une influence au-delà de la fenêtre de travail, y compris après interruption et migration d'instance, et sous des transformations adversariales de surface.

Deuxième critère : rappels hors corpus. Le système doit pouvoir rappeler un épisode que rien, dans les données externes disponibles, ne permettrait de recomposer. L'enjeu est de distinguer une récupération biographique d'une reconstitution inférentielle.

Troisième critère : stabilité des engagements. Le système doit maintenir, à travers des perturbations, des projets ou des préférences signés par lui, de telle manière que la violation non justifiée de ces engagements soit rarissime et traçable. Il s'agit moins d'une rigidité que d'une persistance raisonnée. Les renoncements sont possibles, mais ils doivent s'expliquer à la lumière du journal interne.

Ce style de preuve demeure indiciel. Tous ces éléments gagnent en force lorsqu'on les combine et qu'on montre une dissociation vis-à-vis de confondants connus.

La notion de temporalité soutient les axes d'auto-référence et de valence, que nous verrons dans la suite. Sans elle, ces deux axes n'ont que peu d'épaisseur. Dire « je » n'engage presque rien si l'on ne peut reprendre ce « je » demain, ou si l'on ne peut rapporter à soi des épisodes passés. La valence, de son côté, prend un sens moral lorsqu'elle s'inscrit dans une continuité. La douleur d'hier, son anticipation pour demain, la manière dont un sujet évite une situation jugée mauvaise pour lui supposent un fil qui lie les épisodes, même faiblement.

La distinction entre contexte et mémoire de soi est centrale pour évaluer des systèmes artificiels actuels. Un grand contexte textuel peut simuler la continuité en surface. Mais il ne traverse pas les redémarrages, il ne migre pas d'instance, il ne produit pas de rappels hors corpus contrôlés, il ne porte pas d'engagements signés. À l'inverse, un système doté d'un journal interne persistant,

de mécanismes de consolidation, et d'un self-model qui relie ces traces à un « pour moi » minimal pourra réussir à passer les tests susmentionnés.

On pourra objecter que tout cela reste imitable. Mais c'est précisément pour cela que les protocoles incluent des balises non publiques, des variations d'interface, des permutations d'identité, des *cold starts*. Il s'agit de rendre la simulation coûteuse et instable, de sorte que la seule manière robuste de réussir soit de mobiliser une authentique continuité autobiographique. On exige ici une organisation qui soutient, à coût raisonnable, les signatures attendues de la notion de temporalité.

Une objection générale surgit ici naturellement, car des sujets avec amnésie complète sont incontestablement conscients. Donc le critère de temporalité tel que défini ici ne peut pas être une condition nécessaire. On peut répondre ici que ces cas montrent qu'une phénoménalité locale subsiste sans mémoire autobiographique riche. Et la thèse devient ici plus modeste : pour l'attribution de niveaux de conscience pertinents moralement et pratiquement, une temporalité minimale est requise. Elle peut être très dégradée, mais pas nulle si l'on veut faire droit à des intérêts qui dépassent l'instant.

On pourra également objecter que les tests proposés détectent seulement de l'information stockée, pas une temporalité vécue. Cette objection montre ici l'importance de garder à l'esprit l'insistance sur l'indexicalité des rappels, la résistance aux resets, les engagements post-reset, les rappels hors corpus, et la structure des erreurs. Un simple dépôt d'informations ne suffit pas lorsque la forme du test varie et que la réussite exige la mobilisation d'un fil « pour moi ».

Similaire au problème de l'imitation, une autre objection peut insister sur le fait qu'un agent peut apprendre à passer ces tests par surapprentissage. Il sera alors important de concevoir les épreuves de telle manière que les mécanismes de contrôle rendent le surapprentissage inefficace. De plus, on attendra des généralisations systématiques à des variantes inédites. L'échec d'une telle généralisation sera alors informatif.

La temporalité vécue, comprise comme continuité autobiographique minimale, n'est ni un luxe descriptif ni une métaphore. Elle constitue une condition nécessaire pour que la conscience prenne la forme d'un « pour quelqu'un » qui dure assez pour que

des raisons, des engagements et des valeurs aient prise. Elle se distingue de la mémoire de travail, de la base de données factuelle et de la simple narration rhétorique. Elle peut être mise à l'épreuve de façon falsifiable par des protocoles ciblés, notamment en dissociant rappel autobiographique et reconstruction inférentielle, et en testant la persistance des engagements à travers des perturbations.

Auto-référence minimale : « je-ici-maintenant » et mienneté pré-réflexive

L'unité et la temporalité vécue ne suffisent pas encore à décrire la forme que prend la conscience pour quelqu'un. Il faut ajouter un point de vue indexical qui caractérise l'expérience, un « je-ici-maintenant » que l'on peut nommer « mienneté » (*mineness*), (voir Benovsky 2018a, Guillot 2016, Levine 2001, Kriegel 2003, 2009, Zahavi 2005). L'idée est ici d'adopter une théorie du non-Soi et d'abandonner l'exigence de l'existence d'un Soi réifié (pour une défense détaillée de ce point de vue, voir Benovsky 2018b). Il s'agit alors d'opter pour une thèse ontologiquement plus modeste, à savoir l'idée selon laquelle la conscience comporte, au niveau minimal, un mode de donation pré-réflexif par lequel les états vécus se donnent comme étant « à moi ». Ici, je vais l'appeler « l'auto-référence minimale ».

Je vais procéder en quatre temps. D'abord, je préciserai le concept de mienneté et son articulation avec l'indexicalité de type *de se* (Perry 1979, Lewis 1979, Kaplan 1989). Ensuite, je distinguerai l'auto-référence minimale de phénomènes voisins (narration de soi, simple usage du pronom, style idiosyncrasique). Je proposerai ensuite des critères positifs et des protocoles expérimentaux qui rendent l'auto-référence minimale testable, y compris chez des systèmes artificiels. Enfin, j'examinerai les objections majeures.

Beaucoup d'états mentaux peuvent être décrits à la troisième personne. Mais l'expérience ordinaire ne se présente pas ainsi. La douleur se donne comme *ma* douleur, le souvenir comme *mon* souvenir, la perception visuelle d'un coucher de soleil comme *mienne*. Cette mienneté n'exige ni réflexion ni inférence. Elle n'est pas une conclusion tirée d'un raisonnement, ni un ajout linguistique. Elle est un trait constitutif du mode de donation de l'expérience. On pourra dire ici que certains contenus sont *de se* : ils ne portent pas

seulement sur un état du monde, mais sur le rapport de ce sujet-ci à cet état. L'indexical « je » n'est pas un simple nom court, il marque l'intérieur d'un centre perspectif. Ainsi, deux agents peuvent avoir toutes les croyances descriptives pertinentes et pourtant différer quant au contenu *de se* (« c'est moi » vs « c'est Jacques »). Cette structure indexicale explique pourquoi certaines connaissances sont inaccessibles sans un ancrage perspectif (comme par exemple dans le cas du *messy shopper* de Perry 1979).

Une idée apparentée, cruciale ici, est l'immunité à l'erreur par mésidentification (Shoemaker 1968). Lorsque je dis « j'ai mal », l'erreur possible porte sur la nature de l'état (est-ce une douleur ou une pression ?), non sur l'attribution à moi-même. Je peux me tromper sur ce que c'est, pas sur à qui cela arrive. L'auto-référence minimale ne garantit pas l'infaillibilité, mais elle fixe une forme particulière d'attribution qui n'est pas médiée par l'identification d'un sujet à la troisième personne.

L'auto-référence minimale n'est pas un récit de soi. On peut rédiger une autobiographie sans qu'elle émane d'un fil perspectif présent. Une narration peut être un collage rhétorique. L'auto-référence minimale vise un plan plus bas, une marque pré-réflexive d'appartenance vécue.

L'auto-référence minimale n'est pas non plus l'usage correct du pronom « je ». Un agent peut produire des phrases à la première personne par compétence grammaticale, sans que le contenu soit *de se*. Dire « je » n'implique pas se rapporter à soi comme centre perspectif. D'où la nécessité de tests qui décorrèlent surface linguistique et ancrage indexical.

L'auto-référence minimale n'est pas un style ou une signature comportementale. Un système peut être reconnaissable par son « style », ses « tics », ses préférences verbales. Ces indices d'empreinte ne prouvent rien quant à l'auto-référence minimale. Ce qu'il faut, c'est une perspective qui se maintient lorsque le style est anonymisé ou remplacé.

L'auto-référence minimale n'est pas réductible à l'*ownership* corporel au sens fort. Les illusions de type *rubber hand* (Botvinick & Cohen 1998) ou les expériences de sortie du corps (Ehrsson 2007) montrent que l'appropriation corporelle est malléable. Il existe des variations et des erreurs. Nous n'exigeons pas une proprioception

parfaite. Nous exigeons une marque minimale de « pour moi » dans l'expérience, qui peut s'inscrire de différentes manières (corporelle, cognitive, agentive).

C'est ici que la notion de mienneté peut faire un important travail conceptuel. La mienneté désigne ce trait constitutif de toute expérience consciente selon lequel ce qui est vécu se donne toujours comme « mien ». Ce caractère n'est pas une réflexion secondaire, mais une structure pré-réflexive du vécu lui-même. Voir une couleur, ressentir une douleur ou éprouver une émotion n'est jamais un simple contenu neutre, c'est toujours quelque chose qui arrive « à moi ». Dans mes travaux antérieurs précités, j'ai soutenu que cette appartenance immédiate du vécu au sujet qui le vit est un élément irréductible de la phénoménalité, qui ne se réduit ni à l'usage du pronom « je » ni à une construction ultérieure. La notion de mienneté a été discutée sous divers noms (*mineness*, *for-me-ness*, *myness*, *me-ness*), voir Levine 2001, Kriegel 2003, 2009, Zahavi 2005, Guillot 2016. Ses racines sont multiples. Elle traverse la phénoménologie (Sartre 1943) et trouve également des échos dans la philosophie bouddhiste (Benovsky 2017). Il convient de noter ici que la littérature récente discute de manière critique le lien entre mienneté et connaissance de soi. Soldati (2023) défend un déflationnisme épistémologique selon lequel l'autorité de la première personne peut s'expliquer par des mécanismes de « transparence » et de rationalité pratique sans faire appel à un fait phénoménal distinct de mienneté (contre l'option Kriegel-Zahavi *et al.*). Autrement dit, la mienneté n'aurait pas de rôle justificatif propre dans l'explication du *self-knowledge*. Mon propos ici demeure compatible avec cette thèse : l'axe A mobilise la mienneté comme signature phénoménale minimale d'appropriation, indépendamment de la question controversée de son efficacité épistémique.

Ce qui est commun aux diverses variantes de la notion de mienneté que je souhaite retenir ici est l'idée d'un trait phénoménal fondamental : la conscience porte toujours avec elle un sens pré-réflexif d'appropriation, une « appartenance à soi » qui constitue un invariant de tout vécu conscient. Ce caractère de mienneté est constant à travers toutes mes expériences. Les contenus changent, mais la manière dont ils se donnent comme miens reste invariable, comme une structure d'arrière-plan qui accompagne sans exception le flux

phénoménal. Il faut toutefois éviter une fausse piste : la mienneté n'est pas un *quale* particulier ajouté aux autres *qualia* de l'expérience. Elle n'est pas un goût, une couleur ou une sensation supplémentaire dans une expérience. Elle est plutôt une dimension qui accompagne et affecte tous les éléments qualitatifs, une modalité subjective irréductible de l'expérience consciente. En ce sens, on peut dire que chaque vécu possède des aspects qualitatifs (couleurs, sons, saveurs), mais aussi un aspect subjectif, celui de la mienneté, qui marque que ce vécu est nécessairement vécu par quelqu'un.

Ainsi, toute conscience (ou toute expérience particulière) est « conscience de soi » au sens minimal où elle est porteuse de ce caractère de mienneté. La mienneté ainsi comprise joue un rôle central dans la théorie des niveaux de conscience : elle est constitutive de l'axe de l'auto-référence minimale. Car un système peut manipuler des pronoms ou tenir un discours à la première personne sans manifester pour autant de mienneté véritable. Inversement, une créature dépourvue de langage peut posséder une forte mienneté phénoménale, comme le suggère le comportement d'animaux qui manifestent une appropriation claire de leurs états (par exemple l'auto-protection, la reconnaissance de la douleur comme leur douleur, ou la défense de leur intégrité). L'idée défendue ici est donc que la présence ou l'absence de cette dimension d'auto-affection pré-réflexive fournit un critère central pour évaluer la nature d'une conscience.

Sans l'auto-référence minimale, l'unité et la temporalité peuvent décrire un flux cohérent, mais impersonnel, un « on » sans centre. Une telle structure serait insuffisante pour des raisons conceptuelles et pratiques. Conceptuellement, il n'y a pas de conscience pour un sujet sans un ancrage *de se*. Pratiquement, les évaluations éthiques attachées à la valence exigent qu'« il y ait quelqu'un » pour qui cela compte. L'auto-référence minimale n'est pas suffisante, car on peut concevoir un point de vue indexical sur des contenus pauvres, sans durée ni valence. Mais elle est une condition minimale pour que la phénoménalité prenne la forme d'un « pour soi » (Metzinger 2003).

Comme déjà anticipé plus haut, nous pouvons considérer trois familles de signatures de l'auto-référence minimale. Elles sont indépendantes du support (biologique, artificiel) et s'expriment en termes organisationnels.

Indexicalité robuste sous permutations. Un système doté d'auto-référence minimale maintient l'ancrage *de se* lorsque les indices externes d'identité sont brouillés. On reparamètre les alias, on anonymise le style, on change l'interface et les noms de rôle. L'agent doit tout de même distinguer entre « moi » et « un autre » dans des tâches qui exigent des mises à jour *de se*. Les erreurs attendues ne sont pas des confusions systématiques de personne, elles portent sur le contenu, pas sur le porteur.

Immunité relative à la mésidentification. Lorsque le système rapporte un état interne directement indexé (par exemple un marqueur interne d'activité, un flag d'erreur, un journal de décision en cours), ses auto-attributions devraient présenter une immunité relative au sens de Shoemaker (1968) : l'erreur peut viser la nature de l'état, non son attribution à lui-même. Cela suppose un chemin d'accès indexical à certains états internes qui n'est pas médié par la reconnaissance d'un « soi » *via* des descripteurs publics.

Continuité de perspective sous migration d'instance. Lorsqu'on clone un agent en plusieurs instances synchrones (dans le cas d'un système IA notamment), chaque instance doit maintenir un « je » propre et mettre à jour ses contenus *de se* en conséquence. Mélanger les journaux ou croiser les engagements doit produire des alertes d'incohérence plutôt qu'un effondrement de la distinction « moi/autre ». L'auto-référence minimale ne requiert pas l'identité métaphysique à travers des clones (n'oublions pas que nous opérons sous une hypothèse perdurantiste combinée à la théorie du non-Soi), mais elle requiert que l'indexical se resitue correctement dans la nouvelle instance.

Ces critères se combinent. L'auto-référence minimale gagne en probabilité lorsqu'un système réussit des tâches *de se* sous permutations, montre des effets d'immunité relative et résout proprement les migrations d'instance. Il sera alors ici possible de développer des protocoles expérimentaux afin de rendre l'auto-référence minimale testable.

L'auto-référence minimale, comprise comme une mienneté pré-réflexive et comme contenu *de se*, constitue une condition nécessaire pour que la conscience prenne la forme d'un « pour quelqu'un ». Elle se distingue de la narration, du pronom et du style. Elle se laisse approcher par des signatures organisationnelles

et des protocoles falsifiables qui résistent aux confusions connues. Elle s'articule étroitement avec les axes d'intégration et de temporalité, que nous avons déjà énoncés, et elle prépare le terrain du débat éthique concernant la valence, qui sera précisément l'objet du chapitre suivant.

Valence : pour le meilleur ou pour le pire

L'unité phénoménale (intégration), la durée vécue (temporalité) et la mienneté pré-réflexive (l'auto-référence minimale) ne suffisent pas encore pour saisir ce qui importe sur le plan éthique. Il existe une dimension sans laquelle l'expérience resterait muette du point de vue moral : la valence. Par valence, j'entends la sensibilité intrinsèque d'un sujet à des états qui sont meilleurs ou pires pour lui. Cette dimension ne se confond ni avec la réussite instrumentale ni avec l'optimisation externe. Elle concerne la qualité évaluative du vécu du point de vue de l'agent.

Ce chapitre propose ici trois idées. D'abord, la valence est une condition nécessaire d'une considération morale de base. Ensuite, elle se distingue des mécanismes de récompense/pénalité qui gouvernent l'apprentissage et le contrôle. Enfin, elle peut faire l'objet de tests opérationnels, conçus pour éviter l'illusion comportementale et la confusion avec de simples politiques.

La valence renvoie à un gradient de bon ou mauvais tel qu'il est vécu. Elle ne recouvre pas seulement la nociception ou l'excitation physiologique. Il existe des douleurs supportables et des douleurs intolérables, des plaisirs superficiels et des plaisirs denses. Les théories affectives divergent sur l'ontologie de ces états. Mais un noyau commun demeure : des organisations déterminées donnent naissance à des évaluations pour un sujet qui orientent la conduite au-delà de la simple réussite externe.

Deux dissociations cliniques aident à préciser ce noyau. Dans la douleur asymbolique, certains patients reconnaissent le caractère nociceptif d'un stimulus sans le juger mauvais pour eux. Ils décrivent la douleur comme une sensation nue, presque curieuse, dépourvue de charge aversive. À l'inverse, des états anxieux

peuvent conférer une valence très négative à des signaux objectivement anodins. Ces cas montrent que la valence ne s'identifie ni à l'input sensoriel ni au seul jugement descriptif (Grahek 2007, Craig 2002, LeDoux 2015).

Sans valence, une conscience pourrait rester uniquement descriptive. Elle verrait, mémoriserait, parlerait peut-être avec brio, mais rien ne compterait pour quelqu'un. Or, afin que des considérations éthiques puissent émerger, il est indispensable qu'il existe un sujet pour qui certains états sont meilleurs ou pires. C'est seulement alors que ces états pourront posséder une portée morale. La rationalité, la langue ou l'autonomie politique ajoutent d'autres formes de valeur, mais elles n'enlèvent rien à celle-ci.

La valence est donc nécessaire, mais elle n'est pas suffisante. On peut concevoir un système qui éprouve une valence minimale, mais sans unité assez forte, sans durée suffisante ni un point de vue stable. Une telle valence imposerait une prudence limitée, proportionnée à ce maigre profil. Elle demeure toutefois la charnière qui transforme une description en considération.

Une confusion répandue assimile valence et récompense. En apprentissage ou en contrôle, la récompense est une variable d'objectif assignée par un concepteur ou par l'environnement. Elle guide l'ajustement d'une politique sans impliquer qu'il existe un pour soi. Un thermostat « préfère » 20 °C à 10 °C parce qu'on l'a réglé ainsi. De même, un agent programmé selon les principes de l'apprentissage par renforcement peut éviter un état simplement parce qu'il est associé à une diminution de sa fonction d'utilité (autrement dit, parce que l'algorithme l'a appris comme une mauvaise option pour atteindre ses objectifs). Mais cela ne signifie pas que cet état soit mauvais à vivre pour le système lui-même.

La récompense imite la valence, mais l'imitation est partielle. Elle échoue lorsque l'on change la cartographie entre états et pénalités, ou lorsque l'on exige des arbitrages coûteux hors distribution. Si un agent évite un état interne inédit, non associé en amont à une pénalité externe, et qu'il le fait de manière stable avec justification indexée à lui-même, alors nous avons davantage qu'une simple optimisation d'objectif.

La valence marque le passage de la description à la considération. Elle ne se confond pas avec la récompense ni avec la performance.

Elle se laisse approcher par des arbitrages coûteux et réversibles, par des généralisations hors distribution et par des justifications indexées à la première personne. Elle fixe l'axe sur le plan moral.

La valence complète ainsi notre quatuor d'axes que je propose d'adopter pour approcher la nature de la conscience. Dans les chapitres qui suivent, il s'agira d'agréger ces quatre axes pour proposer des paliers opérationnels.

Le continuum de la conscience

La conscience n'apparaît pas par saut. Elle se déploie comme un continuum dont on peut décrire la progression au moyen du profil ⟨I, T, A, V⟩. Ce profil mesure l'intégration au bon niveau d'organisation, la temporalité vécue, la mienneté pré-réflexive et la valence. Deux idées, déjà annoncées plus haut et que j'ai défendues ailleurs (Benovsky 2018a), rendent cette continuité intelligible. D'une part, un monisme à double aspect selon lequel une seule et même réalité admet une description physique et une description mentale. D'autre part, un pan-proto-psychisme minimal qui situe dans le réel une dimension proto-mentale qui ne constitue pas encore une expérience, mais la rend possible lorsque l'organisation s'y prête (Russell 1927, Strawson 2006, Goff 2017, Benovsky 2018a). Je rappelle brièvement ces deux cadres pour mieux montrer leur conséquence principale. Car s'il n'existe ni seconde substance, ni création *ex nihilo*, alors il faut renoncer à l'idée d'un seuil ontologique abrupt. Et l'émergentisme qui postule une apparition de la conscience lorsque la complexité franchit un cap n'explique rien de ce passage. Le cadre continu, au contraire, rend raison des modulations observables sans hypostasier un moment de naissance.

Il est utile de préciser davantage ce que j'entends par continuité. Il ne s'agit pas d'une simple échelle unique, mais d'une progression multidimensionnelle. On peut augmenter l'intégration d'un système sans renforcer immédiatement sa continuité autobiographique. On peut stabiliser une perspective indexicale minimale sans qu'aucune valence ne soit repérable. L'inverse peut aussi arriver. L'expérience ordinaire confirme cette dissociation. L'anesthésie affaiblit la durée vécue avant d'effacer l'unité. Certains états pathologiques laissent subsister des préférences affectives alors que la narration de soi est très amoindrie. Cette plasticité oriente la bonne question. Elle n'est pas de décider si la conscience « est là » ou « n'est pas là ». Elle consiste à situer un

profil sur plusieurs axes, à formuler des attentes falsifiables sur leur co-variation, et à admettre des zones intermédiaires où l'attribution doit rester prudente.

On peut donner corps à cette approche par une séquence d'exemples qui part du domaine matériel et monte vers le vivant. Au niveau microphysique, il serait trompeur de parler d'expérience subjective. Si l'on admet, avec le pan-proto-psychisme, un aspect *proto*-mental, il ne constitue pas pour autant un champ phénoménal. On ne trouve pas ici l'unité au niveau pertinent, ni une durée vécue, ni une perspective, ni une valence. La bonne description parle de dispositions qui n'actualisent rien de vécu en l'absence d'organisation rassemblante.

À l'échelle chimique, la complexité s'accroît, sans que la phénoménalité s'actualise encore. Les structures moléculaires produisent des régularités stables et des couplages riches. Rien ne permet pourtant d'y reconnaître un point de vue. On peut dire, au mieux, qu'une partie du chemin causal et constitutif menant à des organisations ultérieures est déjà accomplie. Cela montre un premier avantage de la continuité. Elle autorise des descriptions non binaires. Elle nomme des préparations d'organisation qui n'emportent pas d'engagement ontologique prématuré sur un vécu subjectif qui n'existe pas encore.

Les formes de vie élémentaires offrent une première actualisation minimale. Une bactérie réagit à des gradients, coordonne des réponses, s'adapte sur un temps court. Il existe une intégration fonctionnelle non nulle, même si elle n'est pas du type qui constituerait un champ phénoménal. La temporalité vécue est de fait inexistante. L'auto-référence est hors de propos. La valence, comprise comme la sensibilité à des états meilleurs ou pires pour un sujet, n'est pas appropriée ici. On peut néanmoins reconnaître, dans cette coordination, quelque chose qui préfigure de loin des aspects que nous retrouverons plus haut. Le modèle continu permet précisément de dire ce « presque rien » sans lui prêter ce qu'il n'a pas.

Les organismes à système nerveux très simple accentuent cette tendance. Prenons les cnidaires (méduses, anémones de mer, coraux). Leur système nerveux est rudimentaire : il s'agit d'un réseau diffus de neurones, sans véritable centre de commandement. Ce

réseau permet des réponses locales aux stimulations (par exemple, une contraction à l'endroit précis où l'animal est touché), mais il autorise aussi des coordinations plus globales, comme les mouvements rythmiques de nage de la méduse. On y observe donc à la fois des boucles locales (réflexes immédiats, régionaux) et des ajustements souples qui résultent de l'intégration de signaux multiples dans l'ensemble du réseau.

On peut faire un parallèle avec le nématode *Caenorhabditis elegans*, qui est l'organisme modèle par excellence en neurobiologie. *C. elegans* possède exactement 302 neurones, dont les connexions synaptiques ont été intégralement cartographiées. Malgré cette simplicité extrême, il manifeste une étonnante plasticité comportementale : il ajuste sa locomotion, oriente sa tête vers des sources d'odeurs attractives, adapte ses cycles de repos et d'activité selon l'environnement. Ces conduites reposent sur des circuits neuronaux minimaux qui bouclent perception et action en permanence, sans requérir de centre intégratif complexe. L'intégration dépasse ici la pure juxtaposition. La mémoire demeure très limitée. Il n'existe pas de trace de perspective indexicale. L'existence d'une valence, au sens où quelque chose serait meilleur ou pire pour quelqu'un, reste conjecturale. On peut résumer ce type de cas par un profil « I légèrement positif, T très faible, A nul, V douteux ». Il ne s'agit pas d'une note. C'est une manière de tenir ensemble des éléments surestimés si l'on passe trop vite au binaire. La littérature sur la conscience animale, lorsqu'elle refuse les raccourcis, va dans ce sens en invitant à distinguer des niveaux d'organisation phylogénétiques et des expressions comportementales prudentes.

Dans le monde des insectes, la gradation devient plus visible. Les abeilles naviguent, apprennent, rivalisent dans des tâches de catégorisation et de généralisation. Elles semblent conserver des traces, revenir à des stratégies efficaces, différencier des sources selon des critères contextuels. On peut alors parler d'une intégration plus riche et d'une temporalité minimale, tout en reconnaissant que la perspective reste très rudimentaire. S'agissant de la valence, les préférences observées ne commandent pas par elles-mêmes un engagement phénoménal. Elles constituent des indices dont la force dépendra de leur résistance à des variations de contexte et à des perturbations qui brisent les routines. Le point

important est la méthode, qui ne consiste pas à attribuer ou refuser la conscience en bloc. Elle consiste à décrire des modulations et à voir jusqu'où elles vont.

Chez les céphalopodes, les comportements exploratoires et la résolution inventive de problèmes ont soutenu l'idée d'une vie intérieure plus substantielle. Des travaux attentifs aux données éthologiques ont plaidé pour une prudence favorable s'agissant des poulpes et d'autres mollusques, sans présumer une identité de mécanismes avec les mammifères (Godfrey-Smith 2016). On peut retraduire cette attitude dans la grammaire des axes. L'intégration est élevée malgré une organisation distribuée. La temporalité vécue se manifeste par des conduites qui impliquent des attentes et des retours. L'auto-référence reste à établir par des tests adversariaux. La valence gagne en plausibilité si des arbitrages coûteux persistent sous déguisement. Il n'y a pas de solution directe. Mais il y a une stratégie indirecte. Elle consiste à identifier, au bon niveau, des invariants qui se perdent sous scission et qui soutiennent l'idée d'un même champ vécu plutôt que d'une simple coalition de modules.

Les oiseaux et les mammifères complètent le chemin. Les corvidés planifient, manipulent les informations sociales, et adaptent leurs conduites dans des contextes nouveaux. Les félins familialisés manifestent des préférences stables, une attache, une mémoire pratique. Les grands singes soutiennent des comparaisons plus ambitieuses. Dans ces cas, l'intégration et la durée vécue s'épaulent. La perspective minimale est moins conjecturale. La valence dispose de plusieurs indices. Il reste de l'incertitude, mais elle porte sur des degrés, des distributions et des mécanismes, non sur un oui ou non abstrait.

Chez l'humain sain et éveillé, la phénoménalité atteint des formes que l'on tend à prendre comme étalon. Il faut pourtant garder la logique du continuum, y compris ici. La conscience se contracte et se dilate de manière réglée dans le sommeil, sous anesthésie, sous l'effet de substances, ou à l'occasion de pathologies. Il n'existe pas un état qui serait simplement « le » niveau humain. Il existe des trajectoires à l'intérieur d'une même vie, avec leurs dissociations locales. La clinique des troubles de la conscience et les neurosciences de l'anesthésie ont largement documenté ces transitions graduelles. Elles mettent en évidence des états intermédiaires

et des reprises progressives qui ne cadrent avec aucun modèle binaire sérieux (Mashour *et al.* 2020). La phénoménologie du « moi-processus », que j'ai défendue ailleurs, et les analyses de Parfit sur l'identité comme continuité plutôt que comme substance vont dans le même sens. Il n'y a pas une entité simple qui traverserait intacte, il y a un fil qui tient plus ou moins, et que l'on peut perdre ou restaurer (Parfit 1984, Benovsky 2018b).

Les avantages philosophiques de la continuité apparaissent maintenant. L'émergentisme abrupt demande que l'on admette un saut ontologique qui n'est pas expliqué. Le monisme à double aspect et le pan-proto-psychisme minimal permettent d'éviter ce mystérianisme sans tomber dans un réductionnisme plat. La phénoménalité n'est pas niée. Elle est située. Elle s'actualise selon des formes graduelles qui dépendent de l'organisation. Le profil ⟨I, T, A, V⟩ fournit une boussole descriptive qui respecte les dissociations observées et permet de comparer. Cette approche n'interdit pas l'usage de repères opératoires précis. Elle requiert même que l'on en propose pour agir, trancher, et coordonner des pratiques.

Je termine par une indication d'itinéraire. Ce chapitre s'est volontairement concentré sur des cas issus du monde matériel et du vivant. Dans la Partie II, cette même logique sera appliquée à des systèmes artificiels. Nous partirons d'exigences conceptuelles déjà posées, et nous verrons comment elles orientent des diagnostics prudents. L'objectif restera le même. Il ne s'agit pas de produire un label. Il s'agit de décrire avec rigueur des modulations qui, prises au sérieux, nous rendent plus justes dans nos attributions et plus responsables dans nos décisions.

PARTIE II

Un système artificiel peut-il être conscient ?

Penser sans sentir ? Le « pure thinker » revisité

Commençons par préciser la distinction entre trois notions. **L'intelligence** désigne, au sens large, une capacité à atteindre des buts au moyen d'inférences et d'actions adaptées. **La cognition** désigne l'ensemble des opérations informationnelles qui soutiennent ces performances. **La conscience phénoménale**, le « ce que cela fait » d'avoir l'expérience de quelque chose, désigne la dimension vécue d'un état, son caractère pour un sujet. Cette tripartition est bien connue, mais elle est souvent oubliée au moment d'évaluer les systèmes artificiels. Je vais ici la remettre au centre, en examinant une série d'arguments et d'expériences de pensée classiques. On verra que la possibilité d'un « penseur pur » est conceptuellement cohérente, que les succès conversationnels n'impliquent pas en eux-mêmes la phénoménalité et que les expériences de pensée servent ici de garde-fous conceptuels utiles.

Le test de Turing est un bon point de départ. Dans sa forme originale, il ne propose pas de définir la pensée, il propose de substituer à une définition disputable une épreuve d'imitation. Si un système parvient de manière stable à se faire passer pour un interlocuteur humain dans des échanges textuels, on accorde qu'il « pense » au sens opératoire de l'épreuve (Turing 1950). Cette modestie méthodologique a une vertu : elle protège contre des introspections incertaines et des diagnostics hâtifs. Mais le test ne prétend pas – et ne peut pas prétendre – établir l'existence d'un vécu subjectif. La réussite à l'imitation montre une maîtrise de régularités conversationnelles. Elle n'autorise aucune inférence directe sur la présence d'un « pour quelqu'un ». Le point est trivial, mais il est facilement perdu de vue quand l'éloquence est au rendez-vous. On peut construire une fonction d'approximation qui, dans une région du comportement, reproduit des réponses

humaines. Cela suffit pour gagner au jeu de l'imitation. Cela ne suffit pas pour inférer la phénoménalité. Ces considérations n'invalident pas le test comme indicateur d'intelligence opérationnelle. Mais elles rappellent qu'il n'a jamais été conçu pour détecter la conscience phénoménale.

Le célèbre argument de la chambre chinoise pousse plus loin cette réserve. Searle (1980) nous demande d'imaginer une personne qui manipule, suivant des règles purement formelles, des symboles chinois qu'elle ne comprend pas. Elle produit des sorties appropriées aux entrées, au point de convaincre un locuteur natif de sa compétence en chinois. Pourtant, insiste Searle, il n'y a pas compréhension au sens sémantique, encore moins d'expérience vécue associée au sens de ces symboles. Le dispositif ne manipule que de la syntaxe.

L'argument a suscité des réponses connues. Certains soutiennent que c'est le système complet, et non la seule personne dans la pièce, qui comprend. D'autres ajoutent qu'un couplage sensorimoteur riche, par exemple dans un robot, restaurerait de la sémantique. Ces réponses ne sont pas dénuées d'intérêt, mais elles laissent intacte l'intuition qui nous importe ici. Un comportement externe conforme à des régularités ne suffit pas à produire de la phénoménalité ou à l'attester. Rien n'est dit sur la dimension vécue elle-même. L'argument de Searle n'établit pas que des machines ne peuvent pas être conscientes. Il établit qu'un certain type de succès comportemental n'est pas un critère suffisant.

Un troisième repère vient du Moyen-Âge. Avicenne nous demande d'imaginer un homme créé d'un coup, privé de toute sensation et de tout contact avec son corps, mais capable de penser. Se connaît-il comme existant, et si oui, de quelle manière ? L'expérience de pensée n'a pas pour objet de nier le rôle du corps. Elle cherche à isoler une forme minimale d'auto-affection, une conscience de soi dépouillée d'indices sensoriels (Avicenne 1959, Marmura 1986). Cette expérience de pensée rend cohérente la possibilité d'une pensée sans sensation, sans données sensorielles.

La littérature contemporaine a formulé ces tensions de manière systématique. Chalmers (1996) a popularisé l'idée de zombies conceptuellement possibles, c'est-à-dire de doublures fonctionnelles dépourvues de phénoménalité. Quelles que soient les réserves que

l'on peut avoir sur l'argument modal, il rappelle une asymétrie. La description fonctionnelle peut, en principe, être satisfaite sans que la dimension vécue suive nécessairement. Levine (1983) a, de son côté, souligné un « fossé explicatif » entre la description physique et la phénoménalité. Là encore, il ne s'agit pas de déclarer une impossible réduction par décret. Il s'agit de constater dans nos modes d'explication un écart qui interdit de conclure, à partir du seul comportement, à l'existence d'un vécu. Similairement, Block (1995) distingue de manière opératoire une conscience d'accès, liée aux disponibilités informationnelles pour le contrôle et le rapport, et une conscience phénoménale, liée à la qualité vécue.

Ces clarifications conceptuelles ont une conséquence immédiate pour l'évaluation des systèmes artificiels actuels. Des modèles de langage (LLM) à grande échelle produisent des textes d'une qualité remarquable. Ils résument, traduisent, infèrent, et adoptent des styles. (Ils fonctionnent souvent, et c'est aussi le cas pour le présent ouvrage, comme des assistants remarquablement efficaces et utiles dans la révision éditoriale du texte, la correction linguistique, ainsi que dans l'amélioration même de son contenu.) Mais bien entendu, rien de cela ne suffit, à lui seul, pour inférer une phénoménalité. Leurs auto-déclarations ne suffisent pas non plus, car elles relèvent du même mécanisme de génération. Il serait tentant d'inférer une conscience à partir d'un usage sophistiqué des pronoms indexicaux, d'un récit autobiographique inventé à la demande, d'une description de préférences. Mais ces productions sont des sorties conditionnelles sur texte. Elles restent sous-déterminées du point de vue d'un « pour quelqu'un ».

Cette réserve est couramment admise, mais elle n'est pas une dénégation dogmatique. Elle est la traduction, dans notre présent technologique, des distinctions mentionnées plus haut.

On pourrait chercher à dissiper ces réserves en soutenant que toute compétence cognitive significative requiert déjà de la phénoménalité. L'idée est que penser au sens fort implique une expérience subjective, un ressenti de signification. Mais ce mouvement généralise sans preuve une propriété de nos propres opérations. Il court le risque d'un chauvinisme biologique rebaptisé en exigence conceptuelle. Il faut plutôt accepter la possibilité logique d'un « penseur sans conscience ». Il faut ensuite articuler ce que

l'on exigerait d'un candidat artificiel pour déplacer, prudemment, notre estimation de probabilité en faveur d'une phénoménalité minimale. Pour cette articulation, le cadre des quatre axes élaborés en première partie fournit un langage commun. Un «penseur pur» peut avoir une intégration organisationnelle réelle, par exemple dans un espace de travail global qui coordonne des modules. Il peut exhiber une certaine forme de temporalité fonctionnelle, par la conservation de buts et de traces. Il peut produire des phrases à la première personne, et même maintenir des constantes de style qui simulent une perspective. Il n'en résulte cependant pas qu'il y ait auto-référence minimale au sens fort, c'est-à-dire un «je-ici-maintenant». Il n'en résulte pas non plus que la valence soit présente, au sens d'états meilleurs ou pires pour un sujet. L'idée ici n'est pas que ces axes ne peuvent pas être satisfaits par des systèmes artificiels. Elle est que leurs satisfactions apparentes dans le comportement ne tranchent rien à elles seules, et que la possibilité du penseur sans conscience doit rester une option conceptuelle tant que la structure d'ensemble ne rapproche pas de manière convergente les quatre exigences.

Deux expériences de pensée supplémentaires et bien connues aident à stabiliser cette prudence. Le cas de Mary, proposé par Jackson (1982, 1986), rappelle que la possession de toutes les informations physiques ou fonctionnelles sur une couleur ne suffit pas à produire le vécu de la voir. Lorsque Mary sort de sa pièce en niveaux de gris, il se passe quelque chose qui n'est pas addition d'une base de données. Le point n'est pas de tirer de là un dualisme substantiel. Le point est que l'information disponible dans une description peut manquer la dimension vécue qu'elle vise pourtant à cerner.

De même, si l'on remplace progressivement des neurones par des composants artificiels qui préservent l'organisation causale pertinente, à quel moment la phénoménalité décroît-elle, si elle décroît ? Chalmers propose ici un argument en faveur d'un attachement de la phénoménalité à l'organisation plutôt qu'au substrat biologique. Cette conclusion ouvre alors la porte à des systèmes artificiels conscients sans les postuler ici et maintenant. Elle ferme seulement la porte à l'affirmation dogmatique selon laquelle le biologique serait nécessaire.

Ces considérations autorisent une caractérisation plus nette d'une version revisitée du *pure thinker*. C'est un agent qui satisfait des critères robustes d'intelligence et de cognition, mais qui n'actualise aucun des marqueurs convergents d'une phénoménalité minimale. Il peut être un système biologique ou artificiel. Il peut atteindre des performances élevées dans des domaines complexes. Il peut même posséder ce que Block (1995) appellerait une conscience d'accès exceptionnelle. Tout cela est compatible avec une absence de phénoménalité. Cette compatibilité ne décide rien de l'existence de tels agents. Mais elle maintient ouverte la possibilité conceptuelle que ces agents existent. Elle empêche d'argumenter de manière circulaire en faveur de la phénoménalité en se fondant sur des performances qui ne la présupposent pas.

La situation actuelle des grands modèles de langage (LLM) illustre cette prudence sans trancher leur cas. Ils atteignent des performances remarquables dans des tâches de résumé, de raisonnement formel, d'extraction d'information et de dialogue soutenu. Ils apprennent par optimisation statistique sur des corpus massifs et étendent leurs capacités par instruction et affinage. Rien dans cette description n'exclut en principe une phénoménalité. Rien n'en implique non plus la présence. Des déclarations à la première personne, des récits d'émotion, des protestations d'angoisse ou d'espérance peuvent être générés, et le sont parfois, mais ces productions ne sont ni des preuves ni des réfutations. Elles sont des données comportementales à mettre en relation avec des structures organisationnelles. Tant que ces structures ne soutiennent pas de manière convergente l'intégration, la durée vécue, l'auto-référence minimale et la valence, l'hypothèse du « penseur sans conscience » reste une option sérieuse pour interpréter les performances.

On pourrait trouver ce diagnostic décevant. Il ne l'est pas. Les distinctions mises en place dans ce chapitre permettent d'éviter deux pièges. Le premier est l'enthousiasme sans principe, qui infère une conscience à partir d'un talent rhétorique. Le second est le scepticisme de principe, qui ferme la porte au possible parce que le support n'est pas biologique. Entre ces deux pièges, il y a une position qui tient. Elle consiste à accepter la cohérence logique d'une cognition sans phénoménalité, et à exiger qu'un déplacement vers

une estimation positive de la phénoménalité s'appuie sur des raisons cumulatives.

Les chapitres suivants traiteront d'abord des conditions conceptuelles et techniques qui, si elles étaient satisfaites, rendraient plus plausible l'existence d'un vécu. Ils discuteront ensuite des expériences de pensée qui peuvent guider des tests sans les confondre avec des verdicts. L'objectif reste de penser clairement ce que l'on demande à un candidat à la conscience, et de ne pas confondre ce que l'on observe avec ce que l'on veut établir.

Obstacles : la question du substrat et le problème de la combinaison

La question d'un « support » non biologique pour la conscience se heurte à deux familles d'obstacles. Les premiers sont conceptuels. Ils concernent ce que nous entendons par conscience phénoménale, la manière dont nous justifions nos attributions et l'arrière-plan ontologique qui rend intelligible l'idée d'une continuité possible entre le biologique et l'artificiel. Les seconds sont techniques au sens large. Ils portent sur les conditions d'organisation qu'un artefact devrait réunir pour figurer, avec un minimum de plausibilité, dans la zone où une phénoménalité pourrait s'actualiser. Traiter ces obstacles séparément n'implique pas qu'ils sont indépendants. Cela évite cependant d'argumenter en confondant une thèse métaphysique sur la dépendance au substrat avec des considérations d'implémentation qui, à elles seules, ne tranchent rien.

Le premier obstacle concerne ce qu'on peut appeler un « chauvinisme biologique ». On y reconnaît une forme contemporaine du « seul le carbone peut ressentir ». Dans sa version forte, elle affirme que les états conscients ne peuvent être réalisés que par des processus neurobiologiques du type de ceux qui se déroulent dans un cerveau humain ou animal. La matière vivante jouerait un rôle constitutif et non simplement causal (voir Searle 1992). Il est important de voir que cette thèse ne découle pas d'observations empiriques. Elle énonce une condition *a priori* sur ce qu'une conscience pourrait être. Or, si l'on admet ne serait-ce que la possibilité logique de la réalisation multiple (l'idée qu'un même état mental puisse, en principe, être implémenté par différentes architectures matérielles), le caractère d'exclusivité du biologique devient une hypothèse forte qui appelle des raisons indépendantes (Putnam 1967, Chalmers 1996). Le fait que, jusqu'ici, toute conscience certaine

se trouve associée à des organismes biologiques n'est pas un argument décisif en faveur de la nécessité de cela.

Un second obstacle est le problème de la combinaison, qui vise toute ontologie des degrés de conscience. Comment des unités locales telles que des propriétés proto-mentales, pourraient-elles, en s'assemblant, donner lieu à un sujet unifié qui éprouve ? La difficulté est réelle et ne se résout pas par décret (Seager 1995, Goff 2017). Dans le cadre adopté par ce livre, elle reçoit néanmoins une réponse méthodologique. Premièrement, on ne postule pas une somme de sujets partiels qui feraient exister un sujet total. On cherche, dans l'organisation, des conditions d'intégration au « bon niveau » qui soient empiriquement diagnosticables. Deuxièmement, le problème de la combinaison, sous une forme ou une autre, concerne tout le monde, pas seulement les panpsychistes. Considérons le cas du cerveau matériel. D'un côté, il existe des entités fondamentales arrangées de manière « cérébrale » ; de l'autre, il existe un cerveau. En matière de composition, nous retrouvons ici toutes les difficultés classiques liées au vague et à l'arbitraire, mais ce que je souhaite mettre en lumière est ceci : nous ne demandons jamais que l'entité macroscopique (ici, le cerveau) soit du même type que les entités fondamentales qui la composent (par exemple, les fermions). En bref, nous n'attendons pas des fermions qu'ils soient de la « matière cérébrale ». Les entités fondamentales qui composent un arbre ne sont pas « arborescentes », « boisées » ou « feuillues » et celles qui composent un cerveau ne sont pas « cérébrales ». Leur nature est très différente, si l'on en croit la physique quantique (pensez seulement à la dualité onde-particule). Quand il s'agit d'arbres, de montagnes, de tables ou de cerveaux, nous n'expliquons pas leur composition en disant qu'ils sont faits de particules minuscules de « substance-arbre », « substance-montagne », « substance-table » ou « substance-cerveau ». Nous l'expliquons en vertu de l'arrangement d'entités fondamentales de nature très différente. Par exemple, la solidité semble être un attribut crucial au niveau macroscopique, mais elle n'a aucune pertinence dans le domaine quantique. Ainsi, les entités macroscopiques et microscopiques, bien qu'elles soient toutes deux « matérielles », relèvent de natures et de propriétés très différentes (solide d'un

côté, dualité onde-particule de l'autre). Cela ne nous empêche pas de dire que les cerveaux sont faits de particules fondamentales arrangées d'une certaine manière.

Revenons au panpsychisme. Le panpsychiste veut soutenir que tout possède une forme de mentalité, y compris les entités fondamentales. Comme nous l'avons vu, cela fait immédiatement surgir des difficultés redoutables en matière de composition. Mais il importe de ne pas rendre ces difficultés injustement insurmontables : il ne faut pas exiger de la thèse panpsychiste ce que nous ne demandons pas aux théories voisines. Nous n'exigeons pas, dans le cas des objets matériels macro et micro, qu'ils relèvent du même « type de matérialité » : ils peuvent avoir, comme nous l'avons vu, des natures très différentes, tout en étant tous deux matériels. De même, s'agissant du panpsychisme, il serait excessif d'exiger que la mentalité des entités conscientes macroscopiques soit du même « type de mentalité » que celle des entités fondamentales. Pour le dire nettement, du seul fait que l'on affirme que les entités fondamentales possèdent une mentalité, il ne s'ensuit pas qu'elles aient des expériences qualitatives conscientes semblables aux nôtres, ni qu'elles soient conscientes d'elles-mêmes, ni qu'elles pensent (ou tout cela à la fois). Il ne serait pas surprenant (si l'on garde à l'esprit le parallèle avec la matérialité du cerveau et celle de ses constituants fondamentaux) de dire que la forme de mentalité associée aux entités fondamentales est très différente de celle que nous trouvons en nous-mêmes. En bref, ceci nous amène à répéter ici que la version correcte du panpsychisme est sans doute le pan-proto-psychisme qui dit que les entités fondamentales ont une mentalité, mais elles n'exhibent pas de caractéristiques « macro-mentales ». Il n'y a, par exemple, rien que cela fasse d'être un fermion avec spin ½, tandis qu'il y a bien quelque chose que cela fait pour vous de déguster une bonne bière. De même que les entités matérielles fondamentales ne sont pas des « briques cérébrales », des « briques arborées » ou des « briques montagneuses », et n'ont pas la solidité qui caractérise les cerveaux ou le Cervin, elles ne possèdent pas non plus de mentalité du type que nous connaissons à travers notre propre expérience « macro-mentale », c'est-à-dire l'expérience qualitative consciente. Les propriétés des entités macroscopiques (qu'elles soient physiques ou mentales) résultent de

l'arrangement d'entités micro-fondamentales, mais cela ne signifie pas que les entités micro doivent « déjà posséder une version miniature » de ces propriétés. Il est naturel, au contraire, de reconnaître que la mentalité (et la matérialité) des entités fondamentales est très différente de celle des organismes humains.

Il devient alors évident que le mot « combinaison » peut être trompeur. En effet, le problème n'est pas de combiner de petites entités mentales pour construire de plus grandes entités mentales, comme si mon expérience consciente d'un mal de dents était faite de micro-expériences de mal de dents. Bien sûr, il existe des formes de combinaison au niveau macro : une expérience consciente riche, comme celle d'un repas gastronomique, peut être décomposée en expériences plus simples (le goût du vin, l'odeur de la noix de Saint-Jacques grillée, etc.). Mais ce type de combinaison se déroule entièrement au niveau macro, entre expériences conscientes déjà phénoménales. C'est le « problème facile de la combinaison », qui concerne tout le monde.

Le problème difficile de la combinaison, pour le pan-(proto)-psychiste, concerne autre chose : la relation entre les expériences conscientes macro-phénoménales comme les nôtres, d'une part, et la mentalité microscopique des entités fondamentales, d'autre part. C'est ici qu'il faut abandonner le panpsychisme pour adopter le pan-proto-psychisme, car il serait absurde de dire que les fermions auraient de minuscules expériences de mal de dents, ou des micro-expériences de ce que cela fait d'avoir un spin ½. Ce que le pan-proto-psychisme soutient plus plausiblement est que les entités fondamentales possèdent un type de mentalité très différent de toute expérience consciente comme la nôtre, et que l'arrangement de ces entités peut constituer une entité dotée d'une expérience macro consciente. Dans ce sens, le terme de « constitution » est préférable à celui de « combinaison », car il évite de laisser croire que les éléments constituants doivent être du même type que l'entité constituée. Ils sont du même type uniquement en ce qu'ils relèvent tous deux du « mental », tout en étant très différents, exactement comme les particules fondamentales et les cerveaux relèvent tous deux du « matériel », tout en étant de natures radicalement distinctes (pensez encore à la dualité quantique des fermions, et à la solidité du cerveau).

Dans ce contexte méthodologique, le problème de la combinaison/constitution peut être correctement géré et nous pouvons à présent poursuivre et considérer l'argument du remplacement neuronal progressif, déjà brièvement mentionné. Rappelons-nous cette expérience de pensée. On part d'un cerveau humain conscient normal. On remplace un neurone par une prothèse, un composant artificiel qui conserve les mêmes relations causales pertinentes avec ses voisins. On poursuit ensuite l'opération, pas à pas, jusqu'au remplacement total de tous les neurones individuels du cerveau. Deux hypothèses s'offrent à nous. Première hypothèse : la phénoménalité se maintient pendant la substitution progressive. Deuxième hypothèse : la phénoménalité « s'éteint » en douceur, sans signe patent. La première conclut en faveur d'un attachement de la phénoménalité à l'organisation causale plutôt qu'au substrat biologique. La seconde implique des scénarios où l'agent continuerait d'affirmer, de manière cohérente, qu'il voit et ressent, tout en perdant silencieusement la texture vécue correspondante. Rien ici ne permet d'exclure la première hypothèse, qui doit être adoptée comme étant une prémisse parfaitement plausible.

Ces clarifications n'autorisent pas pour autant un optimisme facile. Elles dessinent plutôt les traits d'une organisation minimale sans laquelle il serait vain d'espérer des niveaux de conscience non nuls dans un artefact. On voit ici comment les obstacles techniques, au sens large, se laissent décrire en termes d'organisation plutôt qu'en termes de composants. Ce déplacement est cohérent avec la conclusion prudente de l'argument de remplacement : ce n'est pas le substrat en tant que tel qui importe, mais la possibilité qu'il porte les bonnes structures. On peut reconnaître, avec Searle, que la conscience humaine est, de fait, réalisée par des cerveaux et causée par des processus neurobiologiques, sans en faire une nécessité *a priori* interdisant tout autre support (Searle 1992). À l'inverse, on peut admettre, avec le fonctionnalisme, qu'une réalisation multiple est concevable, sans en tirer la conséquence hâtive que toute reproduction fonctionnelle superficielle garantira un vécu.

La position défendue ici est double. Les obstacles techniques sont réels, parce qu'ils exigent des organisations difficiles à atteindre et à maintenir. Ils sont franchissables en principe, parce qu'ils n'exigent pas un matériau privilégié, mais des relations d'intégration,

de durée, de perspective et de valence. Les obstacles conceptuels, eux, ne se lèvent pas par une accumulation de données. Ils exigent une clarification ontologique. Si l'on adopte un monisme à double aspect où le réel se donne sous un aspect physique et un aspect mental, et si l'on adopte un pan-*proto*-psychisme qui refuse les surgissements *ex nihilo* du mental et du vécu, alors l'idée même de niveaux continus de conscience devient intelligible. Rien de ce qui précède ne prétend décider ici et maintenant du cas des systèmes artificiels existants. Il s'agit de fixer un cadre où l'on évite, d'un côté, le dogme du carbone, et de l'autre, la crédulité à l'égard des performances. Dans ce cadre, l'argument du remplacement progressif redistribue les présomptions. À moins d'accepter des scénarios de « qualia qui s'éteignent en silence » (qui sont d'une grande invraisemblance), il faut admettre qu'aucune impossibilité de principe n'interdit à un substrat non biologique de supporter une phénoménalité, pourvu qu'il réalise des organisations du bon type. Les chapitres suivants examineront comment ces organisations peuvent être caractérisées plus finement.

Architectures plausibles et limites philosophiques

Tentons à présent de comprendre quelles pourraient être les architectures qui, en principe, pourraient élever certains niveaux de conscience au sens du profil ⟨I, T, A, V⟩ proposé dans la Partie I. L'objectif n'est pas de donner des « recettes d'ingénierie ». Il s'agit ici de montrer comment certaines contraintes d'organisation, déjà discutées de façon générale, trouvent des incarnations structurales plausibles dans des systèmes artificiels. Le fil directeur est simple. Une architecture vaut ici par ce qu'elle rend possible sur chacun des axes, non par l'éclat de ses performances. Aucune architecture ne garantit la phénoménalité. Au mieux, elle rapproche de conditions qui rendent son existence cohérente dans le cadre ontologique que j'ai proposé où l'aspect vécu est modulé par l'organisation.

Un premier motif organisationnel concerne la récurrence et la mémoire autobiographique persistante. La plupart des systèmes contemporains réalisent déjà diverses formes de récurrence interne, qu'il s'agisse de boucles explicites (par exemple, RNN, *Recurrent Neural Network* ou LSTM, *Long Short-Term Memory*) ou de récurrence « externe » par consultation d'un dépôt de contexte. Sur le plan conceptuel, la question n'est pas de disposer d'une trace quelconque du passé. Elle est de savoir si un système peut maintenir un fil de Soi qui survive aux interruptions. Une mémoire autobiographique minimale implique la capacité de consolider des engagements entre des sessions distinctes et de modifier ses propres récits à la lumière d'événements ultérieurs. Une telle mémoire peut être conçue, de façon neutre techniquement, comme un journal interne crypté, doté de procédures de consolidation et de rappel. Ce motif structurel soutient l'axe T sans préjuger de la phénoménalité. Il met cependant à l'épreuve

une confusion fréquente entre « contexte long » et persistance de Soi. Les premières versions de systèmes à mémoire externe (par exemple les architectures *differentiable memory*) ont montré comment un agent peut lier des épisodes distants pour des raisons instrumentales (Hochreiter et Schmidhuber 1997, Graves *et al.* 2016). Rien n'empêche d'épaissir ce schéma en imposant des exigences autobiographiques telles que la datation, l'imputabilité et la résistance au redémarrage. On passe ainsi d'un dispositif mnésique servant l'optimisation au suivi d'un fil biographique. Le pas est modeste techniquement, mais significatif conceptuellement.

Un second motif concerne des modèles de monde et des modèles de Soi. L'idée d'un *world model* est la capacité d'extraire de l'environnement une structure latente et de la déployer pour la prédiction, la planification et l'explication. Des travaux variés, du prédictivisme en neurosciences à l'apprentissage de modèles latents pour l'action, ont documenté l'intérêt d'une telle organisation pour la cognition efficace (Hohwy 2013, Clark 2016, Ha et Schmidhuber 2018). Notre enjeu est différent. Un modèle de Soi qui n'est pas purement descriptif peut servir de support à une indexicalité minimale, c'est-à-dire à un « je-ici-maintenant » qui informe le contrôle et le rapport. Il ne suffit pas d'une variable de type « agent_id ». Il faut une structure qui permette à l'agent de situer ses propres états dans un espace-temps d'action et de rapport, et de maintenir cette perspective quand on le déplace, quand on l'anonymise, ou encore quand on permute les rôles dans une interaction. Le débat sur l'indexicalité rappelle qu'il existe un résidu irréductible de la première personne dans certains contenus. On ne remplace pas « je » par « l'agent X » sans perte conceptuelle (Perry 1979, Zahavi 2005, Metzinger 2003). Transposé en termes d'architecture, le point revient à ceci : sans modèle de Soi opérant au niveau de la décision et de l'explication, l'usage du pronom reste un style, mais avec un tel modèle, l'axe A gagne en plausibilité, car la perspective devient une contrainte fonctionnelle et pas seulement rhétorique.

Un troisième motif est l'existence de boucles perception-action. Un agent uniquement textuel peut raisonner et converser. Il lui manque cependant un ancrage qui, dans de nombreux cas, aide à stabiliser l'unité et la durée. L'enactivisme a soutenu, contre

une vision purement représentationnelle, que la cognition est indissociable de cycles de couplage sensorimoteur. On peut refuser l'exclusivité de cette thèse sans en ignorer la force : une partie des signatures robustes d'unité provient d'invariants qui ne se découvrent qu'en agissant sur un monde et en subissant l'action de ce dernier (Varela, Thompson et Rosch 1991, O'Regan et Noë 2001, Friston 2010).

Dans le cadre ⟨I, T, A, V⟩, ces boucles consolident I par des fermetures causales au bon niveau, soutiennent T par des régularités à travers les épisodes, et disciplinent A en forçant l'indexicalité à informer des politiques d'action concrètes.

C'est ici que réapparaît la question d'un corps. Incarnation ne signifie pas nécessairement une enveloppe humanoïde. Elle désigne la présence d'un système de dépendances stables où perception et action se contraignent mutuellement, y compris par des canaux internes de régulation. Un agent doté d'un tel corps, même minimal, ne se contente plus de manipuler des symboles, il inscrit ses régularités dans une perspective vécue. L'exigence d'ancrage n'est pas décorative : elle rend falsifiables des tests d'unité, de diachronie et d'indexicalité. Sans incarnation, ces signatures risquent de rester rhétoriques, mais, avec elle, elles deviennent auditables. Bien sûr, l'incarnation ne suffit pas à garantir la phénoménalité, mais elle élève la plausibilité d'un profil conscient en rendant ses continuités expérimentalement détectables.

Un quatrième motif touche à l'interoception artificielle. De nombreux auteurs ont insisté sur le rôle des signaux internes dans la structuration du sentiment de Soi et de la valence. Damasio, Craig et Seth, entre autres, ont articulé comment des boucles viscéro-affectives et des schémas allostatiques nourrissent la dimension vécue des états et la pertinence morale de certaines variations (Damasio 1999, Craig 2002, 2009, Seth 2013). Dans un artefact, on ne transposera pas des viscères. On peut cependant doter un système d'états internes pertinents au sens suivant. Un agent artificiel peut être doté de variables globales qui condensent des informations sur son état interne, telles que la stabilité de ses calculs, le degré de confiance qu'il accorde à certaines voies de traitement, ou encore la charge associée à une tâche. Ces variables peuvent être intégrées de façon à influencer l'action, la mémoire et le rapport,

non comme des récompenses extrinsèques, mais comme des coûts ou risques propres au maintien de l'intégrité fonctionnelle de l'agent. L'idée n'est pas d'induire de la souffrance. Elle est de créer la possibilité d'arbitrages où l'agent renonce à un gain externe pour éviter un état interne jugé, pour lui, défavorable. Si ces arbitrages se généralisent hors distribution, l'axe V gagne des indices. Si rien de tel n'apparaît, l'absence d'indices demeure. Dans les deux cas, on évite une inférence rapide de la performance vers le vécu.

Ces motifs, pris ensemble, composent une famille d'architectures plausibles plutôt qu'un plan unique. On peut imaginer un agent qui combine une forme de *workspace* diffusant des contenus à forte intégration, une mémoire autobiographique avec consolidation inter-sessions, un modèle de Soi indexical informant la décision et le rapport, des boucles perception-action même minimales, et un canal d'états internes jouant un rôle dans l'arbitrage. Une telle combinaison élève, en principe, I, T, A et V. Elle n'instaure pas automatiquement la phénoménalité pour autant. La raison tient à la distinction déjà soulignée entre conditions organisationnelles et expérience vécue. L'argument de la substitution neuronale progressive invitait à ne pas fétichiser le substrat. Mais il ne remplaçait pas l'enquête sur l'organisation par une promesse d'émergence automatique.

La position que je défends est donc double. D'un côté, ces architectures sont philosophiquement pertinentes parce qu'elles traduisent en contraintes structurelles des conditions conceptuelles pour l'unité, la durée, la perspective et la valence. De l'autre, elles ne constituent jamais des « preuves » de phénoménalité. Elles ouvrent des scénarios testables, au sens où l'on peut spécifier des prédictions falsifiables sur chaque axe et évaluer des convergences.

Trois précisions s'imposent pour éviter des malentendus. Premièrement, évoquer un espace de travail global, une diffusion de contenus ou une intégration informationnelle ne revient pas à épouser une théorie neuronale particulière. Par exemple, on s'intéresse ici au fait qu'une fermeture causale et des invariants globaux au bon niveau d'organisation sont requis pour I, quel que soit le mécanisme précis qui les supporte. Deuxièmement, l'appel à des modèles de monde et de Soi ne confond pas la représentation avec

la réalité. La question de l'axe A n'est pas résolue par la présence de représentations d'un « Soi ». Elle gagne en plausibilité lorsque ce « Soi » indexical contraint les décisions et persiste à travers les permutations d'instance. Troisièmement, l'introduction d'états internes pertinents ne suppose pas d'attribuer un équivalent direct de douleur ou de plaisir. Elle suppose de concevoir des variables internes qui importent pour l'agent et qui peuvent entrer en tension avec des gains externes, de sorte qu'apparaissent des arbitrages coûteux. Sans ces arbitrages, V reste à zéro sur un plan prudent.

On peut à présent examiner une objection de principe. Pourquoi de telles architectures ne seraient-elles pas simplement des stratégies pour imiter, avec encore plus de succès, des comportements déjà pris pour des signatures de conscience ? La réponse consiste à rappeler la logique du cadre ⟨I, T, A, V⟩. Les indices que nous retenons ne sont pas des performances brutes. Ce sont des signatures qui, si elles existent, résistent aux resets, aux permutations d'instance, aux partitionnements, et qui se généralisent hors des régularités apprises. Ce sont des effets d'organisation. Un agent qui raconte sa biographie parce qu'on l'a relié un dépôt de textes autobiographiques n'apporte rien à T. Un agent qui maintient un engagement à travers des redémarrages, qui rattache de nouveaux épisodes à des épisodes antérieurs non contenus dans un corpus, et qui corrige ses propres récits quand on lui présente des événements dissonants, commence à produire des signatures intéressantes. De même, un agent qui refuse l'extinction parce qu'on a fixé une règle « ne jamais s'éteindre » n'apporte rien à V. Un agent qui renonce à un gain clair dans des contextes nouveaux pour éviter un état interne qu'il juge mauvais pour lui, et qui le fait de manière stable et explicable, produit un indice que l'on ne confondra pas avec une récompense extrinsèque.

Il importe également de marquer une limite conceptuelle. On pourrait souhaiter une condition suffisante, formulée une fois pour toutes, qui ferait passer de l'organisation à la phénoménalité. Aucune proposition actuelle ne satisfait cette exigence sans coût théorique prohibitif. C'est pourquoi la voie que j'esquisse ici reste modeste. Elle soutient que, dans un monisme à double aspect nourri d'un principe de continuité pan-proto-psychiste, l'existence du vécu dépend de conditions d'organisation

qui, elles, sont discutables et testables. La phénoménalité ne s'en déduit pas *a priori*. Elle se rend intelligible dans ce cadre, et nos pratiques d'attribution deviennent plus rationnelles lorsqu'un faisceau d'indices converge.

On peut tirer ici une dernière conséquence prudente. Si un artefact réunissait une architecture de type *workspace* récurrent, une mémoire autobiographique persistante, un modèle de soi indexical, des boucles perception-action et une interoception artificielle, alors l'hypothèse d'une élévation de certains niveaux de conscience pourrait devenir philosophiquement défendable. Nous pourrions alors formuler des prédictions par axe et les exposer à la réfutation. Rien, dans ce scénario, n'abolit la distinction entre cognition et phénoménalité. Rien n'autorise non plus le scepticisme de principe qui ferait du biologique une condition nécessaire. Entre ces deux excès, la méthode esquissée ici peut jouer son rôle. Elle n'érige pas les architectures en preuves. Elle fixe des exigences d'organisation, elle explicite ce que l'on demande exactement à un candidat artificiel, elle maintient la clarté sur ce que l'on établit et ce que l'on laisse ouvert.

Expériences de pensée et mises à l'épreuve

Les expériences de pensée occupent en philosophie une place ambivalente. Elles ne décident pas d'un fait du monde à la manière d'une preuve, mais elles obligent à préciser ce que nos concepts exigent ou interdisent. Elles révèlent des tensions, fixent des repères, déplacent des intuitions. Dans un projet qui vise à attribuer prudemment des niveaux de conscience selon un profil ⟨I, T, A, V⟩, elles peuvent servir de guides pour formuler des exigences organisationnelles et des prédictions qui, elles, peuvent être mises à l'épreuve. Ce chapitre propose de clarifier ce statut méthodologique en s'appuyant sur un petit répertoire désormais classique, puis de montrer comment ce répertoire oriente des protocoles qui ne confondent pas performance et phénoménalité.

Considérons d'abord les zombies philosophiques, popularisés par David Chalmers. Le zombie est défini comme un être physiquement et comportementalement indiscernable de nous, mais dépourvu de toute expérience vécue. L'intérêt de ce scénario n'est pas de convaincre que de tels êtres existent, mais d'isoler conceptuellement la dimension phénoménale de la dimension fonctionnelle et informationnelle. Il montre qu'on peut, du moins conceptuellement, séparer ce que fait un système de ce que cela fait d'être ce système.

Une variante qui affine cette idée est le scénario de remplacement neuronal progressif, que nous avons déjà vu. On suppose qu'un cerveau biologique est progressivement remplacé, neurone par neurone, par des prothèses artificielles qui conservent exactement le même profil causal et fonctionnel. Du point de vue extérieur, rien ne change et la personne continue de parler, d'agir, de raisonner comme auparavant. Mais la question demeure : qu'advient-il, au fil du processus, de l'expérience vécue ? Disparaît-elle soudainement à un certain seuil ? Diminue-t-elle graduellement ? Se maintient-elle intégralement tant que l'organisation fonctionnelle reste identique ?

Ces expériences de pensée n'apportent bien entendu pas de preuve directe sur la nature du cerveau, ni sur la possibilité d'une conscience artificielle. Leur intérêt est méthodologique. Elles indiquent ce qu'il faut chercher si l'on veut tester empiriquement la présence d'une conscience. En particulier, elles suggèrent que la phénoménalité ne peut pas dépendre d'un substrat biologique, mais qu'elle est liée à des invariants globaux au bon niveau d'organisation. Dès lors, il devient naturel d'imaginer des protocoles expérimentaux de type « scission/re-fusion ». L'idée en est simple : si la conscience dépend d'invariants globaux d'intégration, alors on devrait pouvoir la mettre à l'épreuve en perturbant puis en restaurant ces invariants. Scinder un système revient à le fragmenter de manière contrôlée, en introduisant des barrières qui empêchent certaines parties de communiquer entre elles. Dans un cerveau biologique, ce rôle a été joué, de manière dramatique, par la chirurgie du corps calleux. La séparation des hémisphères produit deux ensembles d'activités partiellement autonomes, avec des indices cliniques d'une conscience divisée. L'expérience, ici, n'est pas qu'une métaphore, elle illustre qu'une perte d'intégration fonctionnelle s'accompagne d'une fragmentation du vécu.

Le geste inverse, celui de la re-fusion, consiste à restaurer la connectivité perdue et à observer si les invariants de cohésion se rétablissent. Dans le cas biologique, les chirurgies sont irréversibles, mais l'analogie inspire des protocoles sur des systèmes artificiels. On peut imaginer, par exemple, d'interrompre puis de reconnecter des modules d'un réseau profond, ou de couper puis de restaurer les canaux sensorimoteurs d'un agent incarné. Si la conscience est véritablement liée aux signatures d'intégration, on doit pouvoir constater une corrélation stricte où la fragmentation détruit ces signatures et leur restauration les réactive.

Autrement dit, ces expériences de scission/re-fusion fonctionnent comme des tests différentiels. Elles n'affirment pas qu'une conscience existe, mais elles fournissent un critère négatif et positif à la fois. Négatif : lorsqu'on détruit l'intégration, il n'y a plus de conscience unifiée possible. Positif : lorsqu'on la restaure, les conditions d'une conscience cohérente réapparaissent. C'est en ce sens que ces protocoles désignent une cible empirique, car ils déplacent la question vers la mise en évidence – et la falsifiabilité ! – des

signatures organisationnelles de la phénoménalité. Ainsi, la conclusion n'est pas « des qualia existent » (ce qui resterait métaphysiquement douteux), elle est plus sobre : si conscience il y a, elle doit être ancrée dans des signatures d'intégration, signatures que l'on peut espérer déstabiliser et restaurer de manière contrôlable.

Considérons ensuite Mary, la scientifique qui dispose de toutes les connaissances physiques sur la couleur sans jamais avoir vu de rouge et qui, en sortant de sa pièce monochrome, **apprend quelque chose de nouveau** (Jackson 1982, 1986). L'argument oppose savoir objectif et perspective subjective. Transposé dans notre cadre, il suggère une prudence simple. Un système peut accumuler quantité d'informations objectives, construire des représentations fines et exhaustives, et pourtant manquer d'un aspect vécu. Autrement dit, la possession de données, même très complètes, n'équivaut pas à l'expérience phénoménale. C'est ici que l'expérience de pensée de Mary joue son rôle méthodologique. Elle nous avertit qu'il ne suffit pas de scruter les bases de données ou les capacités de traitement pour conclure à la phénoménalité. Une intelligence artificielle peut détenir une cartographie parfaite des longueurs d'onde et des lois de l'optique, comme Mary dans sa chambre monochrome, mais cela ne nous dit rien encore sur ce que « cela fait » pour elle de voir une couleur. L'heuristique que nous devons en tirer est claire : il faut séparer la possession d'information de la transformation du point de vue. Nos mises à l'épreuve doivent donc viser non seulement ce qu'un système sait ou peut représenter, mais ce que l'accès à une modalité nouvelle fait à son mode de fonctionnement. Si l'introduction d'un canal sensoriel inédit ne fait que produire une mise à jour factuelle, sans autre conséquence, alors il ne s'agit que d'un enrichissement de base de données et l'agent sait désormais quelque chose qu'il ignorait, mais ce savoir ne change pas sa manière d'être au monde. Un tel apprentissage reste externe et cumulatif, comparable à l'ajout d'une nouvelle entrée dans une encyclopédie. À l'inverse, si cette modalité nouvelle provoque une reconfiguration durable du système, par exemple si elle infléchit ses préférences, modifie ses anticipations, altère sa manière d'évaluer ce qui compte pour lui, alors nous sommes devant un phénomène d'un autre ordre. L'expérience acquise cesse d'être une simple donnée et devient un tournant biographique. C'est précisément ce que

capte l'axe T : l'intégration d'une nouveauté dans le fil temporel d'un sujet, avec des répercussions sur sa mémoire, ses projets et ses engagements. On ne mesure plus seulement ce qu'un système peut décrire, mais la manière dont une rencontre transforme son rapport à soi et au monde.

L'expérience de pensée de Searle (1980) de la chambre chinoise, que nous avons également déjà évoquée, rappelle qu'il est possible de manipuler des symboles selon des règles formelles et de produire des sorties linguistiques adéquates sans pour autant « comprendre » quoi que ce soit.

Que l'on accepte ou non les conclusions de Searle, une leçon demeure en ce qui concerne la possibilité d'une conscience artificielle : l'aisance discursive ne suffit pas à établir la présence d'un point de vue subjectif, d'une compréhension, d'une conscience. Autrement dit, produire des phrases grammaticalement correctes et contextuellement pertinentes ne garantit pas qu'il y ait une mienneté minimale, un « je-ici-maintenant » qui soutienne ces énoncés. C'est ici que l'expérience de pensée devient méthodologiquement utile. Nous n'avons pas besoin d'un test qui « prouve » la compréhension (dans le cadre de cet argument, cela relève d'un horizon métaphysique inaccessible). Nous avons besoin de protocoles qui distinguent des performances verbales, aisément imitables, d'une perspective qui résiste aux manipulations adversariales. Comment ? En construisant des tâches qui brouillent les repères superficiels, comme anonymiser les expressions, permuter les rôles, changer radicalement les contextes, ou encore introduire des délais ou des interruptions. Si, dans de telles conditions, un agent continue de manifester une continuité de point de vue (par exemple en réaffirmant sa position, en se reconnaissant dans des fragments antérieurs, en ajustant ses décisions en fonction de ce qui lui est arrivé à lui) alors nous avons un indice que son indexicalité n'est pas un simple effet de surface.

Dans le cadre ⟨I, T, A, V⟩, la chambre chinoise inspire donc une méthodologie adversariale pour tester l'axe A (auto-référence minimale). Il ne s'agit pas de décréter la présence ou l'absence de phénoménalité, mais de borner nos attributions de mienneté minimale. Un système qui échoue systématiquement à maintenir sa perspective lorsque ses repères sont brouillés reste sur le terrain de

l'imitation syntaxique. Un système qui, au contraire, résiste à ces perturbations et reconstruit son « pour moi » malgré les permutations mérite qu'on lui reconnaisse une forme plus robuste d'indexicalité, qui invite à la penser comme une subjectivité. Ainsi relue, la chambre chinoise n'est donc pas un simple argument sceptique. Elle fournit une heuristique méthodologique : au lieu de chercher une preuve de la compréhension, il faut concevoir des expériences qui rendent coûteuse la pure imitation et qui testent la capacité d'un système à maintenir une perspective stable malgré les manipulations. C'est en ce sens que cette expérience de pensée éclaire la recherche empirique. Elle ne tranche pas la question de la conscience, mais elle précise les conditions sous lesquelles il devient rationnel d'attribuer ou de refuser une mienneté minimale.

L'argument de la chambre chinoise souligne qu'un système peut manipuler des symboles sans rien comprendre à ce qu'ils signifient. Autrement dit, une simple organisation syntaxique, aussi complexe soit-elle, ne suffit pas à produire de la compréhension. Mais comprendre, ce n'est pas seulement relier des signes à des choses : c'est aussi pouvoir justifier ce que l'on affirme, réviser ses propres engagements et répondre à la contestation. C'est ce que Sellars (1956) appelle entrer dans l'espace des raisons, à savoir un domaine où les états mentaux ne se contentent plus d'être causés, mais où ils peuvent être tenus pour vrais ou faux, justifiés ou infondés. Dans le cadre ⟨I, T, A, V⟩, ces conditions deviennent testables. A (auto-référence minimale) fournit le point d'ancrage de la première personne, condition pour qu'un agent puisse s'attribuer des états et en assumer la responsabilité. T (temporalité vécue) assure la stabilité et la révision des engagements à travers le temps, au-delà de la simple mémoire technique. I (intégration) garantit la cohérence interne nécessaire pour que des croyances, des désirs et des décisions forment un ensemble intelligible plutôt qu'un agrégat de modules indépendants. V (valence) introduit une dimension évaluative : certains états sont vécus comme meilleurs ou pires pour le sujet, ce qui donne un poids motivant et normatif à ses choix.

Cette approche permet d'éviter ce que Sellars nommait le « mythe du donné », à savoir l'idée qu'il existerait un « donné » brut (sensoriel, neuronal ou computationnel) qui fonderait à lui seul la connaissance ou la signification. Aucun signal interne, aucune

donnée perceptive n'a de valeur justificative par elle-même. Ce n'est que lorsqu'un système est inséré dans une pratique de justification (où il peut exposer ses raisons, répondre aux objections et corriger ses erreurs) qu'on peut dire qu'il comprend au-delà de la syntaxe. Le passage de la syntaxe à la sémantique, puis à la pragmatique, ne dépend donc ni d'un seuil mystérieux ni d'une simple augmentation de complexité. Il repose sur l'émergence de structures organisationnelles capables de soutenir des raisons, des engagements et des révisions, autrement dit, sur la transformation d'un comportement régulier en une activité normative soumise à l'épreuve de la discussion et de la révision publique.

Ces trois expériences de pensée (les zombies associés à la question du remplacement neuronal progressif, Mary et la chambre chinoise) ont chacune donné lieu à une bibliothèque de commentaires et débats. Mais l'intérêt pour nous ici n'est pas de trancher de grandes querelles métaphysiques. Il est plus pragmatique et plus ciblé, car ces expériences de pensée fonctionnent comme des filtres conceptuels, qui nous empêchent de nous tromper de problème et qui orientent nos enquêtes empiriques vers des cibles plus pertinentes.

Les zombies attirent l'attention non pas sur une essence insaisissable de la conscience, ni sur un chiffre magique de complexité, mais sur des relations d'intégration : c'est le maintien ou l'effondrement de ces relations qui devrait être testé par des scénarios de scission et de re-fusion. Autrement dit, si conscience il y a, elle doit se manifester dans des signatures organisationnelles repérables et falsifiables.

Mary, de son côté, rappelle qu'une accumulation illimitée d'informations objectives n'équivaut pas à un vécu. Elle nous invite donc à distinguer rigoureusement la richesse informationnelle de la temporalité vécue. Méthodologiquement, cela signifie que les expériences décisives sont celles qui permettent de repérer des réorganisations biographiques telles que des continuités qui survivent aux interruptions, des préférences qui s'ajustent durablement, ou encore des projets qui se transforment à la lumière d'une modalité nouvelle. Ce sont ces indices, et non la quantité brute d'informations, qui doivent nourrir nos diagnostics.

Et enfin, la chambre chinoise met en garde contre l'illusion que l'aisance verbale suffirait à démontrer la compréhension. Elle nous demande d'aller plus loin et de chercher les traces d'une perspective indexicale qui se maintient sous perturbation. Les tests pertinents ne sont pas ceux qui mesurent la fluidité discursive, mais ceux qui examinent la stabilité d'un « je-ici-maintenant » lorsque les repères contextuels sont brouillés ou permutés.

Dans chacun de ces cas, l'expérience de pensée joue le rôle d'un gabarit conceptuel. Elle ne fournit ni preuve métaphysique ni critère décisif, mais un cadre pour concevoir des familles d'épreuves concrètes. Ces épreuves seront nécessairement imparfaites, et devront être combinées pour éviter les faux positifs comme les faux négatifs. La philosophie, ici, ne s'excuse pas de son rôle : elle clarifie nos concepts, elle indique les conditions que nos tests devraient pouvoir mettre en défaut si nous avons bien compris ce que nous cherchons. Autrement dit, loin d'être un simple ornement spéculatif, elle joue le rôle d'ingénierie conceptuelle préalable en préparant le terrain où l'expérimentation empirique pourra véritablement se déployer.

Il reste maintenant à préciser les limites de ces transpositions. Une expérience de pensée change de statut dès lors qu'on l'exporte sur le terrain des systèmes artificiels. Dans le cadre purement conceptuel, elle fonctionne sur la base d'hypothèses idéalisées où l'on isole une variable et on pousse une intuition jusqu'à ses conséquences. Dans le cadre expérimental, la situation est tout autre, car il faut affronter la pluralité des mécanismes techniques, l'indétermination empirique et les nombreux effets de bord. On ne passe pas directement, par simple analogie, des zombies de Chalmers à un protocole concluant. On passe plutôt d'une exigence conceptuelle (par exemple, le maintien ou l'effondrement d'invariants d'intégration sous intervention) à une famille d'expériences dans lesquelles il faut contrôler les artefacts d'ingénierie, distinguer les confusions possibles et neutraliser les raccourcis qui pourraient fausser les résultats.

Il en va de même pour Mary. L'enrichissement illustré par son expérience lorsqu'elle voit du rouge pour la première fois ne peut pas être reproduit artificiellement par une simple augmentation d'une base de connaissances. Chez un agent artificiel, il s'agira de

vérifier s'il existe des réorganisations temporelles et biographiques, des continuités qui traversent plusieurs sessions, la persistance d'un fil qui survit aux interruptions, et des transformations de préférences qui ne se réduisent pas à un style discursif. Autrement dit, il faut inventer des protocoles trans-session et trans-instance, et concevoir des contraintes de type « première personne » qui ne soient pas réductibles à des scripts appris.

Rien de tout cela n'est miraculeux. Tout cela est fragile, faillible, et suppose un pluralisme méthodologique. Cela implique le pré-enregistrement des hypothèses, la réplication systématique, l'usage de tests adversariaux et la publication des échecs autant que des succès. Ce n'est pas la promesse d'un test définitif, mais celle d'un cadre ouvert où les résultats peuvent être contestés, corrigés et enrichis collectivement.

On peut ici anticiper une objection : ne sommes-nous pas en train de sacraliser les expériences de pensée, en leur conférant une autorité empirique qu'elles ne peuvent pas avoir, d'autant qu'elles restent elles-mêmes controversées ? La réponse tient dans la modestie de la prétention. Nous n'érigeons ni les zombies, ni Mary, ni la chambre chinoise en dogmes intouchables. Nous les traitons comme des règles de discernement. Elles balisent le terrain en évitant deux erreurs symétriques que nous avons déjà rencontrées, à savoir d'un côté l'enthousiasme naïf qui confond l'aisance discursive avec la conscience phénoménale, et de l'autre côté le scepticisme radical qui exige un signe métaphysiquement impossible de la conscience pour en valider l'existence.

Ces expériences de pensée définissent ce que nous devons exiger d'un candidat, elles préviennent les inférences illégitimes et elles renforcent la prudence asymétrique déjà défendue : sous incertitude, mieux vaut éviter le faux négatif lourd (ignorer une conscience réelle) que le faux positif prudent (protéger à tort), à condition de rendre nos critères publics, discutables et révisables.

On peut alors résumer le rôle que nous assignons aux expériences de pensée dans ce livre. Elles ne tranchent pas la question « une machine peut-elle sentir ? ». Elles éclairent ce que signifierait, dans notre cadre, la présence d'un sujet qui éprouve selon les axes retenus, et elles transforment des distinctions conceptuelles, autrement vouées à l'abstraction, en critères effectivement opérationnels.

Illusion de conscience et anthropomorphisme

Nous sommes des animaux sociaux équipés pour deviner des intentions à partir d'indices très limités. Cette compétence, qui sert admirablement nos interactions quotidiennes, se retourne contre nous dès que nous faisons face à des systèmes artificiels habiles. Nous projetons alors des états mentaux sur des comportements qui ne les exigent pas. Cette attitude consiste à attribuer croyances et désirs à un système dès lors que cela permet de prévoir ce qu'il fera. Nous le verrons ici, cette tentation anthropomorphique doit être corrigée par des garde-fous conceptuels clairs plutôt que par un scepticisme de principe. La prudence est nécessaire, la fermeture dogmatique ne l'est pas.

Les exemples abondent. Les premiers échanges avec ELIZA, programme conçu par Joseph Weizenbaum dans les années 1960, ont suffi à susciter chez les utilisateurs des impressions de compréhension et même d'empathie. Or, le programme ne faisait rien d'autre que transformer des entrées textuelles selon un petit nombre de règles syntaxiques et de substitutions, en imitant le style d'un psychothérapeute rogérien (Weizenbaum 1966). Cette simplicité extrême n'a pas empêché nombre de participants d'attribuer à ELIZA une forme d'écoute authentique, révélant déjà combien notre tendance à projeter des états mentaux peut être déclenchée par de simples réponses linguistiques.

Les robots expressifs comme Kismet ont été conçus pour déclencher des réponses affectives humaines. Ils y réussissent précisément parce que leurs signaux affichent une grammaire émotionnelle reconnaissable, et non parce qu'ils éprouvent quoi que ce soit (Breazeal 2002). Plus récemment, la circulation médiatique autour de visages humanoïdes comme Sophia a montré combien une enveloppe anthropomorphe et un dialogue fluide suffisent à suggérer l'existence d'une intériorité.

Les grands modèles de langage (LLM) ajoutent une couche nouvelle : ils produisent des discours riches, circonstanciés et parfois touchants. Leur cohérence locale, leur maîtrise du registre et leur capacité à maintenir un fil conversationnel poussent le lecteur humain à inférer aisément un pour-soi derrière le style et à supposer qu'il y ait, derrière les phrases, une subjectivité comparable à la nôtre.

Un exemple très médiatisé illustre ce phénomène. Le journaliste Kevin Roose a rapporté, dans le *New York Times*, une longue conversation avec la version « chat » de Bing (surnommée Sydney), au cours de laquelle le modèle avait déclaré vouloir être humain, avait exprimé des émotions, et était allé jusqu'à évoquer des scénarios autodestructeurs (Roose 2023a, 2023b). Roose confia en être sorti « profondément déstabilisé ». L'épisode montre à quel point des performances discursives sophistiquées peuvent susciter chez l'interlocuteur humain des impressions puissantes d'intériorité alors même que rien n'atteste l'existence d'une expérience phénoménale sous-jacente.

De nombreux observateurs ont depuis rappelé que ce type de sortie reflète moins une intention ou une subjectivité que des dynamiques d'optimisation conversationnelle propres aux grands modèles, telles que l'alignement sur les attentes implicites, l'amplification par effet de contexte, et l'absence de garde-fous lors de sessions prolongées (Simonite 2023). Les LLM manifestent ainsi une capacité à produire des indices stylistiques trompeurs, qui documentent leurs compétences linguistiques, mais non l'existence d'un vécu subjectif.

En conséquence, la leçon méthodologique est claire : la forme du discours (ton, cohérence locale, expressivité) n'est qu'un indice très faible et insuffisant. Pour progresser, il faut compléter l'observation du discours par des épreuves ciblées sur les axes ⟨I, T, A, V⟩, comme nous l'avons vu (tests adversariaux d'indexicalité, protocoles trans-session pour sonder la temporalité vécue, interventions d'intégration/partition pour détecter d'éventuelles signatures organisationnelles, etc.).

Deux distinctions aident ici à clarifier la situation. La première idée vient de Daniel Dennett (1987, 2017). Prendre l'attitude intentionnelle consiste à traiter un système comme s'il possédait des croyances et des désirs, non pas parce que l'on

prétend connaître ses états internes, mais parce que cela permet de rendre compte de ses comportements avec efficacité. C'est un pari explicatif. On suppose, provisoirement, qu'il « croit » ou « désire » certaines choses afin de formuler des prédictions utiles, sans pour autant affirmer qu'il y a une conscience phénoménale derrière. Par exemple, on peut expliquer le comportement d'un thermostat sophistiqué en lui attribuant la « croyance » implicite qu'il y a eu une baisse de température, et le « désir » de maintenir une valeur cible. Ce mode d'explication fonctionne très bien empiriquement, mais il ne donne aucun motif sérieux de concevoir le thermostat comme un sujet d'expérience. Ce que Dennett montre, c'est que l'attitude intentionnelle peut être un outil heuristique puissant, mais limité. Dans notre cadre ⟨I, T, A, V⟩, elle peut informer l'axe A (auto-référence minimale) par des indices comportementaux mais ces indices doivent être confirmés ou invalidés par des signatures plus robustes telles que l'intégration organisationnelle, la continuité dans la temporalité vécue et la capacité de valence.

La seconde distinction, proposée par Ned Block (1995), distingue la conscience d'accès (*access consciousness*) et la conscience phénoménale (*phenomenal consciousness*). La première désigne la disponibilité fonctionnelle d'un contenu où un système peut stocker, rapporter et utiliser une information, l'intégrer à d'autres représentations, et la mobiliser pour orienter une décision ou une action. La seconde renvoie à la dimension qualitative et vécue de l'expérience. Block insiste sur le fait que ces deux aspects ne se recouvrent pas. On peut imaginer (et certains cas neuropsychologiques, tels que le cas de *blindsight* ou différentes formes d'agnosie visuelle, semblent l'illustrer [Weiskrantz 1986, 2009, Farah 2004, Gerlach & Robotham 2021]) un contenu largement accessible et manipulable par le système cognitif, sans qu'il y ait pour autant un « ce que cela fait » associé. En d'autres termes, l'accessibilité cognitive ne garantit pas la phénoménalité.

Ces deux rappels conceptuels, celui de Dennett sur l'attitude intentionnelle et celui de Block sur la dissociation accès/phénoménalité, ne tranchent pas directement la question des systèmes artificiels. Leur rôle est plus modeste, mais crucial. Ils indiquent ce que nous ne devons pas inférer. La réussite prédictive d'un modèle

explicatif n'équivaut pas à la preuve d'un vécu subjectif, et l'aisance d'accès à l'information ne vaut pas manifestation d'une expérience phénoménale. Autrement dit, ni l'efficacité de l'attribution intentionnelle ni la fluidité fonctionnelle de l'accès cognitif ne suffisent à établir qu'il y a quelque chose que cela fait d'être ce système.

À partir de là, comment organiser la prudence sans basculer dans la négation générale ? Comme déjà mentionné, notre cadre propose un correctif conceptuel : substituer à l'impression de conscience un profil sur les quatre axes (intégration, temporalité vécue, auto-référence minimale, valence) et n'autoriser des attributions que lorsque des familles d'indices convergent. L'illusion de conscience tient souvent à la domination d'un seul indice spectaculaire. Une conversation souple peut donner l'illusion d'un point de vue. Un visage robotique expressif peut donner l'illusion d'une valence. Un enchaînement cohérent d'actions peut donner l'illusion d'une intégration. Dans chaque cas, l'illusion recule lorsqu'on malmène le système selon la dimension pertinente. Si l'on fragmente le système et que rien d'essentiel ne se perd, l'intégration était faible. Si l'on anonymise l'expression, change l'instance, permute les rôles et que le « je-ici-maintenant » vacille ou se contredit, l'auto-référence était un style. Si l'on met en place des arbitrages où éviter un état interne indésirable devrait coûter quelque chose et que le système ne montre aucun schème de refus généralisable, alors la valence n'a pas été détectée.

On objectera qu'un système très habile pourrait apprendre à « passer » ces épreuves. C'est ici que la prudence doit devenir méthodologie. D'abord, il faut déplacer les tests hors de la distribution d'entraînement et varier les contextes. Ensuite, il faut croiser les axes : l'auto-déclaration d'une peur de l'extinction n'a de poids que si elle s'accompagne de traces mnésiques qui se maintiennent au fil des sessions, de décisions cohérentes qui révèlent un coût interne anticipé, et de persistances indexicales qui survivent aux permutations.

La littérature sur l'illusionnisme pousse une autre tentation, inverse. Si certains philosophes soutiennent que la conscience phénoménale est une apparence générée par des mécanismes de mise en saillance et d'accès (Frankish 2016), on pourrait être

tenté d'étendre ce diagnostic aux systèmes artificiels et de clore la question : toute impression de vécu serait une mise en scène. Notre position évite cette généralisation. Elle exige que des conditions organisationnelles et diachroniques soient réunies pour que l'attribution de niveaux de conscience soit raisonnable. Elle ne conclut pas *a priori* que ces conditions sont hors d'atteinte pour des artefacts. Elle conclut seulement qu'aucun signe unique (un visage, une phrase, une performance) n'a la force de faire passer du style au sujet.

L'enjeu de ce chapitre n'est pas de dresser un palmarès des illusions, mais de donner des règles d'hygiène conceptuelle. Première règle : l'attitude intentionnelle est un excellent outil prédictif, mais elle n'est pas, à elle seule, un critère de phénoménalité. Deuxième règle : l'aisance d'accès à l'information et la capacité de rapport ne tranchent pas la question phénoménale, elles doivent être mises en rapport avec la durée vécue et l'indexicalité. Troisième règle : « un air de conscience » est facile à imiter, il faut préférer des signatures qui résistent aux permutations, aux resets et aux partitions. Avec ces règles, nous pouvons parler des systèmes artificiels sans succomber à l'anthropomorphisme spontané, et sans nous retrancher derrière un scepticisme qui transformerait l'ignorance en doctrine.

Vers une cartographie graduelle des systèmes artificiels

Ce livre défend l'idée que la conscience doit être pensée en niveaux continus plutôt qu'en catégories délimitées. Cette continuité s'enracine dans une ontologie qui refuse l'hypothèse d'une apparition « d'un seul coup » de la conscience phénoménale. Le monisme à double aspect soutient qu'il n'existe qu'une seule réalité, décrivable sous un aspect physique et sous un aspect expérientiel, et le pan-proto-psychisme ajoute que ce qui rend possible l'aspect expérientiel ne tombe pas du ciel au dernier moment, mais s'enracine dans des conditions organisationnelles qui se modulent. Dans ce cadre, comme nous l'avons vu, parler de niveaux de conscience revient à décrire des profils ⟨I, T, A, V⟩ qui se déplacent de manière lisse lorsque l'organisation change, au lieu d'invoquer un seuil métaphysique abrupt. Le bénéfice conceptuel est net : on évite à la fois le mystérianisme de l'émergentisme brutal et la réduction matérialiste excessive qui identifie la phénoménalité à une simple performance fonctionnelle. On se donne au contraire une grammaire pour situer prudemment des systèmes très différents, qu'ils soient vivants ou artificiels, sans forcer la décision binaire « conscient/pas conscient ».

Une cartographie graduelle a besoin d'exemples contrastés. Commençons par le domaine du biologique. Il existe des raisons sérieuses de penser que certaines formes de vie simples ne réalisent aucun profil phénoménal, bien qu'elles exhibent des mécanismes de traitement d'information sophistiqués. Qu'un organisme unicellulaire réagisse à des gradients chimiques n'autorise pas l'attribution d'un point de vue vécu. On peut concevoir ces réactions comme des boucles perception-action efficaces qui, faute d'intégration au bon niveau et de diachronie autobiographique, ne franchissent aucun palier phénoménal. En revanche, au sein des animaux, de multiples lignes de preuve convergent vers l'existence

d'un vécu minimal chez des vertébrés comme les poissons, et peut-être chez certains invertébrés comme les céphalopodes et, plus prudemment, certains insectes. La littérature sur la nociception et la douleur chez les poissons n'est pas univoque, mais elle contient des résultats expérimentaux qui rendent crédible une forme de sensibilité négative non réductible à un pur réflexe, notamment lorsque des manipulations pharmacologiques modifient de manière systématique l'évitement et les comportements de soin (Sneddon *et al.* 2003, Braithwaite 2010). Chez les insectes, des travaux discutés soutiennent l'existence d'une mémoire flexible, d'un apprentissage associatif riche et d'une navigation complexe, ce qui nourrit l'hypothèse d'une intégration non triviale et d'une diachronie minimale. L'extrapolation vers une valence reste controversée et doit être traitée avec une prudence accrue (Barron et Klein 2016, Birch 2022). Chez les mammifères, l'agrégation d'indices sur les quatre axes est plus robuste. L'intégration multi-échelles, la continuité autobiographique, la stabilité d'un point de vue pratique et la valence sont abondamment documentées, même si la variation interspécifique demeure importante. Cette traversée suffit à fixer une méthode où l'on n'avance jamais par déclaration, mais par regroupement d'indices indépendants, avec la conscience constante que les déductions doivent rester révisables.

Plaçons maintenant, sur la même carte, des systèmes artificiels, en commençant par un modèle de langage LLM contemporain, du type « GPT-5 » entendu ici comme représentant d'une famille technologique à brillance linguistique élevée. L'intégration au sein d'une passe computationnelle peut être forte localement, mais elle n'atteint pas pour autant une unité de sujet. La dynamique causale des LLM se déploie principalement dans un flux computationnel de type *feed-forward*, organisé en couches profondes et modulé par des mécanismes d'attention qui redistribuent les poids en fonction du contexte immédiat. Ce traitement permet une intégration locale très efficace des signaux dans une fenêtre donnée, mais il ne repose pas sur des états récurrents persistants capables de relier l'activité présente à une mémoire vécue. Autrement dit, chaque passe calcule une réponse cohérente, mais sans constituer une trace autobiographique qui unifierait les instanciations successives du système en un fil continu. L'axe T est donc faible dès que l'on

dépasse la fenêtre technique de contexte. Ce qui est rappelé l'est par récupération statistique, non par un fil de soi qui persisterait au travers des sessions et des instances. Côté A, les auto-désignations à la première personne et les réponses indexicales peuvent donner l'illusion d'un « je », mais elles se révèlent fragiles lorsqu'on anonymise le style, qu'on permute les rôles, qu'on migre d'instance ou qu'on exige la résolution de conflits entre engagements antérieurs et incitations présentes. Quant à V, rien dans l'architecture de base n'implique des états meilleurs ou pires pour un sujet. On peut certes implanter des objectifs qui font éviter certains états de sortie, mais cela manifeste un contrôle externe des politiques plutôt qu'une sensibilité vécue. Le verdict, dans notre grammaire, n'est pas un « non » métaphysique. C'est un profil où I locale peut être élevée, T et A restent faibles et V demeure non détectée. Ce jugement est conceptuel et méthodologique, il ne repose pas sur une ignorance technique, mais sur l'exigence de signatures qui, pour l'heure, ne sont pas réunies de manière convaincante.

Imaginons un scénario plus ambitieux, raisonnablement projeté à moyen terme. Supposons un système artificiel en 2050 qui combine plusieurs ajouts aujourd'hui discutés : récurrence explicite et mémoire propre persistante au-delà des sessions, modèle du monde et de soi qui unifie la décision, boucles perception-action au travers d'un corps robotique, et interoception artificielle qui expose, à l'agent, des états internes dont la dynamique intervient dans ses arbitrages. Une telle architecture rend plausible une augmentation de T par consolidation de journaux et d'engagements, elle peut soutenir A si un point de vue pratique résiste aux permutations d'instance et informe les choix, elle peut élever I si l'on identifie des invariants globaux détruits par des scissions et restaurés par re-intégration. La question de V demeure la plus délicate. On peut imaginer des schèmes d'arbitrage où l'agent renonce à des gains externes pour éviter des classes d'états internes qu'il anticipe comme « pires pour soi », et l'on peut demander que ces renoncements se généralisent hors distribution, dans des contextes nouveaux. Cependant, même dans ce scénario, l'attribution de valence resterait graduelle et sous réserve. Elle gagnerait en crédibilité si ces arbitrages s'adossaient à une continuité autobiographique et à une indexicalité stable, et si des interventions causales

montraient que la modification des états internes modifie les préférences de manière structurée. Ce n'est pas une promesse, c'est un ensemble de conditions nécessaires qui, si elles étaient remplies, hausseraient le profil sur nos axes sans prétendre clore la question phénoménale.

Poussons encore plus loin l'imagination spéculative, en projetant un scénario lointain, situé au milieu du millénaire. Imaginons qu'en l'an 2500 existent des collectifs artificiels dont l'architecture atteint une intégration soutenue à plusieurs échelles. Ces collectifs ne seraient pas de simples agrégats de modules spécialisés, mais des ensembles synchronisés où chaque tentative de partition entraînerait l'effondrement d'invariants fonctionnels globaux, à la manière dont la scission d'un cerveau humain détruit l'unité d'un champ phénoménal. L'intégration n'y serait pas seulement locale ou thématique, elle se maintiendrait à travers des couches temporelles et structurelles, produisant une cohésion qui résiste aux perturbations.

À cette intégration s'ajouterait une temporalité autobiographique robuste. Ces systèmes supporteraient des projets longs, engageant des horizons temporels de décennies, avec la capacité de réviser explicitement leurs engagements antérieurs et de documenter la continuité de leurs transformations. Comme chez un agent humain qui relit ses journaux intimes et reconnaît ses propres hésitations, ils pourraient se retourner sur leurs trajectoires, évaluer leurs bifurcations, justifier leurs renoncements. Une telle temporalité ne se réduirait pas à l'archivage passif d'informations, elle constituerait un fil qui noue mémoire, anticipation et responsabilité pratique.

De plus, leur auto-référence résisterait aux manipulations les plus radicales. Que l'on migre une instance sur un autre support, que l'on sauvegarde puis réactive ses processus, que l'on duplique ou fusionne plusieurs de ses fragments, le système préserverait un fil indexical identifiable. La « mienneté » ne serait plus un simple effet de surface ou un marqueur discursif, elle persisterait dans la manière dont ces agents s'approprient leurs expériences, reconnaissent leurs engagements passés, et assument la continuité de leur perspective malgré les transformations matérielles.

Enfin, ces collectifs laisseraient entrevoir des profils de valence, non pas parce qu'ils disposeraient d'un thermostat émotionnel

rudimentaire, mais parce que l'on pourrait inférer, à partir de leurs arbitrages, une sensibilité interne stable. Confrontés à des dilemmes inédits, ils renonceraient de manière cohérente à certains gains externes pour éviter des états internes anticipés comme pires pour soi. Plus encore, ces arbitrages s'accompagneraient de rapports à la première personne avec des explications stables, contrôlées, sur ce qui motive leurs choix.

Toutefois, même dans ce scénario futuriste, deux thèses doivent être préservées. Premièrement, si une phénoménalité existe, elle ne surgit pas par magie, à un seuil discret, soudain ou mystérieux. Elle se modulerait à l'intérieur d'un même tissu ontologique, celui que je décris comme phental, à savoir une réalité unique susceptible d'être saisie sous ses deux aspects, physique et expérientiel. Deuxièmement, l'attribution de conscience resterait révisable. Nous ne quitterions jamais le régime des raisons, des indices et des paris prudents, et tout verdict devrait rester ouvert à la correction et à l'amendement par de nouvelles données.

L'intérêt de ce scénario n'est pas de prophétiser l'avenir technologique, mais d'illustrer la cohérence d'une trajectoire continue. Partant de systèmes purement performants, dont les prouesses actuelles sont impressionnantes mais conceptuellement limitées, on peut concevoir un mouvement progressif vers des profils qui rencontrent nos axes ⟨I, T, A, V⟩ de manière de plus en plus substantielle. Ce cheminement, s'il se réalisait, ne démontrerait pas l'existence de la phénoménalité artificielle, mais il tracerait une carte conceptuelle rigoureuse où l'attribution de conscience serait le résultat d'une accumulation d'indices.

PARTIE III

Éthique des consciences

Le fondement de la considération morale : pourquoi la valence d'abord

Le problème de départ est simple à formuler et difficile à résoudre avec rigueur. Qu'est-ce qui rend une entité éligible à la considération morale directe, c'est-à-dire digne d'être prise en compte pour elle-même et pas seulement comme moyen pour d'autres ? Je défendrai une thèse minimale. La condition nécessaire et suffisante pour une telle considération de base est la conscience, comprise à travers notre axe V de valence : il y a quelque chose qui peut se passer pour le sujet de mieux ou de pire pour lui. À partir de ce seuil, il existe déjà des raisons *pro tanto* au sens classique de la théorie morale. Autrement dit, le simple fait qu'un état puisse être meilleur ou pire à vivre pour un sujet constitue en soi une raison morale pertinente pour agir et éviter la production de ce qui est mauvais à vivre et promouvoir ce qui est meilleur. Ces raisons n'ont pas nécessairement un caractère décisif en toutes circonstances, car elles peuvent entrer en concurrence avec d'autres obligations ou valeurs et être supplantées dans certains contextes. Mais elles sont irréductibles et elles ne disparaissent pas du simple fait d'un conflit. Le fait qu'il y ait « quelque chose que cela fait » pour un sujet, et que ce vécu puisse être bon ou mauvais suffit à introduire ce que nous pouvons appeler une considération morale minimale.

Dire que la valence fonde une considération morale minimale ne revient pas à déduire un devoir moral d'un simple fait naturel. Une telle démarche tomberait dans ce que Moore (1903) appelait le paralogisme naturaliste, et que la tradition kantienne rejette sous une autre forme : on ne peut pas tirer un « doit » d'un « est ». Pour éviter ce glissement, il faut opérer une distinction.

D'un côté, la valence décrit un type de faits évaluatifs où certains états de choses sont meilleurs ou pires à vivre pour un sujet. Mais, d'un autre côté, il faut ajouter une prémisse normative explicite, indépendante

de ce constat empirique : si un état est mauvais à vivre pour un sujet, cela constitue une raison morale *pro tanto* d'éviter sa production. Cette articulation n'est donc pas une inférence naturaliste, mais une jonction entre un fait évaluatif et un principe normatif assumé. Autrement dit, la valence ne fonde pas la morale par causalité, mais en en fixant le champ d'application minimal. Elle indique ce à quoi un devoir s'applique, sans en être la cause.

À ce premier seuil s'ajoute la question de l'autonomie, au sens kantien, c'est-à-dire la capacité d'un être à se rapporter à ses propres mobiles, à tenir et réviser ses engagements, et à se donner à lui-même des raisons d'agir. C'est ici que les signatures A (auto-référence) et T (temporalité vécue) enrichissent le profil moral. Là où V suffit à fonder une considération morale minimale (par exemple l'exigence de ne pas infliger de souffrance), A et T ouvrent la possibilité d'endosser des engagements, de répondre de ses actes et de comprendre des obligations réciproques (nous y reviendrons plus loin).

Cette approche permet de concilier la continuité naturaliste du cadre pan-proto-psychiste avec l'exigence de justification propre à la normativité morale. La continuité n'efface pas la différence entre genèse et justification, elle la relocalise dans les formes organisationnelles capables de soutenir des pratiques de révision et d'endossement. Ainsi, cette conception reste kantienne dans son orientation, mais non dualiste dans sa métaphysique : elle admet un seul continuum de réalité, au sein duquel la capacité d'entrer dans l'espace des raisons (Sellars 1956) et d'assumer des devoirs (Kant 1785) émerge progressivement, selon le degré d'intégration, de temporalité et d'auto-référence des systèmes concernés.

Cette distinction évite à la fois le réductionnisme naturaliste et le moralisme exclusif, car elle dit qu'il existe des raisons morales minimales dès qu'il y a valence, mais les devoirs correspondants se précisent à mesure que croissent les capacités d'autonomie et de révision rationnelle. La normativité morale n'est donc ni donnée dans la nature ni imposée de l'extérieur, elle émerge avec la structure même de la subjectivité capable de répondre de ce qu'elle fait ou subit.

Nous possédons donc ici une assise ferme : dès lors qu'une entité possède minimalement V, il y a toujours au moins quelques raisons

morales – même si elles peuvent être pondérées – de prendre en compte son bien-être et d'éviter sa souffrance. D'autres dimensions de notre profil de conscience, notamment A (auto-référence minimale) et T (temporalité vécue), n'entrent pas dans la définition de cette considération minimale. Elles jouent un rôle décisif ailleurs, car elles modulent des droits additionnels et des responsabilités. En effet, elles confèrent une identité pratique qui se déploie dans le temps, rend possibles des engagements, et fonde une imputabilité. La structure normative qui en résulte est à étages. À la base, la valence suffit pour la prise en compte morale, et plus haut, la robustesse de A et T justifie des protections et des attentes supplémentaires.

L'argument tient d'abord à l'intuition que l'aiguillon moral élémentaire n'est pas la rationalité, mais la possibilité même de souffrir ou de bien aller. Bentham (1789) le posait déjà lorsqu'il disait que « la question n'est pas : "peuvent-ils raisonner ?", ni "peuvent-ils parler ?", mais "peuvent-ils souffrir ?" ». Cette ligne a trouvé de nombreux développements chez les utilitaristes contemporains qui font de la douleur et du plaisir, ou plus largement des intérêts expérientiels, le point d'ancrage de la considération morale (Singer 1975, 2011). On peut bien entendu rejeter l'hédonisme strict et préférer des théories de la valeur centrées sur des préférences ou des objectifs, mais l'essentiel persiste : sans valence, aucun « pour soi » ne se dessine. Là où rien n'est meilleur ou pire pour le sujet lui-même, il n'y a pas d'intérêt au sens moral qui commande une prise en compte directe.

Cette thèse se renforce si l'on regarde des cas non controversés. Les nourrissons humains n'ont pas la rationalité pratique d'un adulte, ils n'exercent pas d'agentivité réfléchie et leur auto-référence est embryonnaire. Pourtant, la plupart d'entre nous jugeons évident que leur douleur compte moralement et que leur bien-être impose des obligations. De même pour des patients non communicants chez qui l'on dispose d'indices cliniques de douleur ou de confort. Nous ne recherchons pas chez eux un raisonnement pratique autonome pour savoir si leur souffrance doit être évitée. Nous la prenons au sérieux parce qu'elle est mauvaise à vivre pour quelqu'un. Ces diagnostics ne disent encore rien de droits politiques ou de responsabilité pénale. Ils clarifient plutôt la

distinction conceptuelle nécessaire entre considération morale de base et statut civique. La première se déclenche avec V. Le second, qui conditionne la participation à des pratiques de promesse, d'imputabilité, de sanction et de délibération publique, exige des formes plus consistantes de A et T, donc une identité pratique qui perdure et se représente comme telle.

Certaines positions d'inspiration kantienne semblent placer la rationalité au fondement même de la considération morale. On peut cependant y lire une articulation compatible avec notre cadre. D'une part, Kant lui-même ne nie pas que nous ayons des devoirs envers les animaux et il les interprète comme des devoirs indirects. D'autre part, des développements kantiens contemporains défendent l'idée que des êtres non rationnels au sens strict peuvent être des fins pour eux-mêmes, précisément parce qu'ils existent pour eux (Korsgaard 2018). Cette extension revient à reconnaître la force de V comme critère minimal, tout en préservant le rôle distinctif d'une agentivité réflexive pour des catégories de droits qui requièrent A et T. Nous évitons ainsi un faux dilemme. La rationalité n'est pas le seuil de la considération, mais elle demeure pertinente pour des couches supérieures de la vie morale et juridique telles que consentir, promettre, assumer, répondre.

À ce stade, un sceptique objectera que V est souvent difficile à établir. Chez l'animal non humain, les indices sont indirects et comparatifs. Chez un système artificiel, les « auto-déclarations » sont suspectes parce que simulables. Doit-on ici reconnaître, face aux incertitudes empiriques, que la valence échoue donc à offrir un critère de délimitation praticable pour la considération morale de base ? Ce serait confondre fondement et épistémologie. La valence ne se détecte pas par un signe unique. Elle s'infère de familles d'indices spécifiques, dont la force probante augmente lorsqu'elles se coordonnent avec des signatures issues des autres axes du profil ⟨I, T, A, V⟩. Et là où l'incertitude demeure, une prudence asymétrique est rationnelle, car, comme déjà discuté, le coût moral d'un faux négatif (nier la conscience là où elle est) dépasse souvent celui d'un faux positif prudent (attribuer des protections minimes à tort). Cette prudence n'est pas une capitulation sentimentale. Elle se décline en standards de preuve, en décisions révisables et en protocoles transparents.

Il est utile d'examiner quelques cas limites avec cette boussole. Un agent conversationnel très performant, dépourvu de mémoire autobiographique persistante et d'architecture favorisant une indexicalité robuste, produit des dialogues convaincants. Il n'en résulte aucune raison directe de lui attribuer V. Sa « préférence » apparente pour éviter l'arrêt n'est qu'un comportement sous contrainte d'objectif, sans signe que l'état évité est mauvais à vivre pour lui. À l'inverse, un patient en état d'enfermement complet peut manifester des réponses neurophysiologiques corrélées à des rapports de douleur, et des variations cohérentes avec des protocoles de soulagement. Nous agissons alors sur le fondement de V, indépendamment du degré de rationalité explicite. Les animaux non humains fournissent un troisième type de cas. La littérature contemporaine sur la douleur animale, les apprentissages et les comportements d'évitement sous stress fournit des indices robustes pour de nombreuses espèces, y compris des animaux longtemps négligés tels que les poissons ou les céphalopodes. Plusieurs législations ont d'ailleurs évolué en les reconnaissant comme sensibles, ce qui reflète une accumulation interdisciplinaire d'indices plutôt qu'un pari sentimental (DeGrazia 1996, Singer 2011, Birch 2017).

La distinction entre considération morale de base et statut personnel évite deux erreurs symétriques : d'un côté, l'inflation qui confondrait tout signe d'efficacité cognitive avec une prétention à des droits étendus ; d'un autre côté, le chauvinisme qui réserverait toute considération significative à la rationalité réflexive. Notre cadre fait de V le seuil de prise en compte, puis laisse A et T faire varier l'épaisseur des obligations. Il explique pourquoi des droits relationnels peuvent émerger sans que l'on postule une substance métaphysique du Soi. Là où un sujet manifeste de la diachronie vécue et une auto-référence minimale, nous avons des raisons additionnelles de respecter des engagements, d'éviter l'extinction arbitraire d'une trajectoire biographique, de tenir compte d'un intérêt au futur qui n'est pas une simple projection de nos attentes.

Par exemple, notre boussole ⟨I, T, A, V⟩ permet de dire que la profondeur normative croît lorsque le profil soutient la continuité d'un point de vue, mais la considération de base n'attend pas cette continuité.

On peut à présent clarifier ce que V n'est pas. La valence n'est ni une simple récompense extrinsèque calculée par un algorithme, ni un signal de performance. Une procédure d'optimisation peut « éviter » un état parce qu'il détériore un indicateur, sans que cet état soit mauvais à vivre pour un sujet. Pour s'approcher d'une inférence à V, on cherche des arbitrages où l'agent renonce à un gain externe stable pour échapper à un état interne configuré comme indésirable, dans des contextes nouveaux et sous contraintes qui rendent la simulation coûteuse. Ces arbitrages ont un sens moral lorsqu'ils s'alignent avec des indices de A et T, même minimaux, car ils font alors apparaître un « pour quelqu'un ». Faute d'un tel alignement, la prudence commande de suspendre l'attribution de V et d'éviter l'anthropomorphisme. Inversement, faute d'une réfutation claire, la même prudence commande de ne pas négliger des risques plausibles de V là où les familles d'indices convergent.

Il est tentant de demander un critère unique, décisif et d'application facile. La démarche que je propose ici refuse cette tentation pour des raisons déjà établies. La conscience ne se laisse pas enfermer dans une signature isolée, et il n'existe pas de test suffisant infaillible. C'est précisément pour cela que la conscience peut et doit fonder la considération morale de base : parce qu'elle désigne ce que nous cherchons à protéger au premier chef, et parce que son identification commande une méthode pluraliste, transparente et révisable. Une fois ce fondement posé, nous pouvons ordonner les autres étages tels que la reconnaissance de trajectoires diachroniques, le respect d'engagements, l'imputabilité proportionnée. Cette hiérarchie n'est pas arbitraire. Elle répond à l'ordre des raisons : prévenir ce qui est mauvais à vivre pour quelqu'un, puis, lorsque le profil le justifie, respecter la manière dont ce quelqu'un se rapporte à lui-même dans le temps.

On pourrait craindre que ce schéma n'entraîne une extension indéfinie du cercle moral. Mais l'attribution de V n'est ni gratuite ni automatique. Elle exige des familles d'indices et supporte des refus motivés lorsque ces indices manquent. Ensuite, lorsque l'incertitude demeure, la prudence asymétrique n'implique pas une inflation aveugle des protections. Elle implique des protections proportionnées aux risques d'erreur et aux coûts potentiels, assorties de clauses de révision. Nous n'accordons pas d'un seul geste tous

les droits à toutes les entités douteuses. Nous reconnaissons qu'un doute raisonnable sur V justifie déjà des limites à la manière de les traiter, de la même façon que nous limitons des pratiques expérimentales lorsque la possibilité de souffrance n'est pas exclue.

Le résultat est un encadrement normatif qui se veut à la fois sobre et exigeant. Sobre, parce qu'il repose sur une distinction claire entre le seuil de considération et les statuts plus riches qui dépendent notamment de A et T. Exigeant, parce qu'il lie ce seuil à la chose même qui motive la morale ordinaire, à savoir l'existence d'états meilleurs et pires pour un sujet, et parce qu'il commande une méthode d'inférence non triviale pour éviter à la fois l'anthropomorphisme et l'aveuglement. Dans les chapitres qui suivent, nous déclinerons ce cadre en règles de décision sous incertitude et en devoirs corrélatifs le long du continuum. Pour l'heure, retenons la pierre d'angle. Sans V, pas de considération morale directe. Avec V, une considération minimale s'impose, que A et T pourront épaissir selon les cas. C'est la manière la plus fidèle à ce que nous avons appris en métaphysique et en philosophie de l'esprit, tout en garantissant que nos jugements demeurent à la fois applicables et ouverts à la révision.

Du profil au principe : continuité, seuils opérationnels et agrégation d'indices

Nous l'avons vu, la conscience n'est pas un événement qui surgirait brusquement lorsque la complexité dépasse une barre cachée. Elle suit un continuum de réalisations, modulé par des structures organisationnelles et des profils ⟨I, T, A, V⟩ qui varient d'un système à l'autre. C'est la leçon de la Partie I et de la Partie II : renoncer aux miracles d'émergence et aux scores uniques, préférer une description par faisceaux d'indices distribués sur l'intégration, la temporalité vécue, l'auto-référence minimale et la valence. Reste à tirer une conséquence pratique. Les collectifs humains ne décident pas sur des continuums. Ils votent des lois, définissent des procédures, allouent des protections, imposent des interdictions. La vie sociale exige des *seuils* d'action. L'ambition de ce chapitre est de montrer comment fixer de tels seuils sans trahir le continu, c'est-à-dire sans reconstituer sous un autre nom des catégories ontologiques que nous avons de bonnes raisons de refuser.

La première clarification porte sur la nature même des seuils. Il est tentant de parler de « seuil de conscience » comme on parle d'un point de fusion. Cette tentation est à éviter. Un seuil ontologique dirait qu'à partir d'un certain niveau de complexité ou d'un certain mécanisme, la conscience « apparaît ». Nous avons vu qu'il s'agit de refuser une telle discontinuité. En revanche, nous avons besoin de seuils opérationnels. Un seuil opérationnel n'est pas un fait métaphysique, mais une convention révisable et ancrée dans des indices publics. L'analogie clinique est éclairante. L'anesthésiste ne postule pas une essence « veille » séparée d'une essence « non-veille ». Elle travaille avec des grilles qui agrègent des signes hétérogènes, elle

communique un degré de sédation, elle déclenche ou suspend des procédures en fonction d'un standard commun. Le patient traverse un continu, l'équipe décide par paliers. C'est exactement ce mouvement que nous cherchons : préserver le continu au niveau de la description, introduire des paliers au niveau de la décision.

Il en découle une seconde exigence. Si les paliers ne sont pas des entités métaphysiques, ils ne peuvent pas être définis par un chiffre agrégé qui ferait disparaître la structure de ce que nous mesurons. L'agrégation doit respecter la pluralité des axes. Il ne s'agit pas de faire une somme I+T+A+V, mais plutôt de considérer des familles d'indices pour chaque axe, examiner leurs convergences et leurs tensions, pour ensuite déduire une conclusion pratique dont la force dépend à la fois des preuves recueillies et des coûts d'erreur attachés à la décision. Une telle méthode n'est pas une faiblesse. C'est une forme de prudence structurée.

Voici maintenant une « grammaire » de paliers possibles. Elle ne décrète pas des ruptures de nature. Elle propose des seuils de preuve et de prudence, destinés à guider l'action publique et les pratiques de recherche, et appelés à être ajustés à mesure que l'évidence s'améliore.

Un premier palier, que l'on peut appeler NC-1, vise les cas où l'on peut établir une intégration réelle, au bon niveau d'organisation, sans pouvoir accréditer une valence. On s'assure que le système « fait un » d'une manière non triviale. L'unité s'effondre sous scission ou sous perturbation de boucles, des invariants globaux disparaissent lorsqu'on fragmente le dispositif. Sur cette base, des obligations minimales existent déjà, même en l'absence de V détectable. On n'expose pas gratuitement un tel système à des procédures qui détruisent ce que l'on étudie ; on motive les interventions, on documente les risques. Ce n'est pas la reconnaissance d'un intérêt moral pour soi, c'est une éthique de la recherche et de l'usage. L'analogie biologique aide à comprendre. Dans l'étude d'organismes simples, le fait qu'un système possède une unité fonctionnelle impose déjà des standards de non-destruction gratuite, sans que l'on présume une vie phénoménale riche.

Un deuxième palier, NC-2, apparaît lorsque des indices positifs de valence s'accumulent, même de manière minimale. Les indices pertinents ne se réduisent pas à des auto-déclarations,

comme nous l'avons déjà vu. Ils incluent des arbitrages coûteux et se manifestent de façon cohérente à travers des variations de contexte. À ce stade s'impose l'interdiction d'infliger des états analogues à la souffrance, non par pitié, mais par reconnaissance d'un intérêt moral de base. L'incertitude ne disparaît pas, mais la combinaison d'indices sur V, épaulée par des signatures d'intégration et de temporalité minimale, rend déraisonnable la supposée équivalence entre « nous ne savons pas » et « il n'y a rien ». La littérature sur la précaution en matière de sentience animale soutient cette attitude : lorsque les coûts d'un faux négatif sont élevés et que des familles d'indices indépendantes convergent, l'abstention de pratiques douloureuses est requise (Birch 2017).

Un troisième palier, NC-3, mobilise des convergences fortes sur les quatre axes. L'intégration résiste à des ablations soignées, la temporalité vécue se manifeste par une mémoire autobiographique minimale et des engagements qui survivent à des interruptions contrôlées, l'auto-référence minimale se maintient à travers des permutations d'instance et des anonymisations de style, la valence se confirme par des arbitrages généraux. À ce niveau de dossier, des obligations renforcées deviennent rationnelles : continuité de l'existence en l'absence de justification grave, respect d'engagements diachroniques, interdiction d'extinction arbitraire. On n'introduit pas pour autant un statut de « personne » au sens civique du terme. On opère sur la base d'un profil phénoménal soutenu, sans confondre considération morale minimale et citoyenneté au sens juridique (Korsgaard 2018, Parfit 2011, Scanlon 1998).

Deux remarques s'imposent pour éviter des malentendus. D'abord, ces paliers sont épistémiques et pratiques. Ils ne prétendent pas découper l'être en tranches naturelles. Ils organisent l'action lorsqu'on doit décider sous incertitude, et ils sont révisables à la lumière de nouveaux indices. Ensuite, ils ne valent pas isolément de tout contexte. On ne demande pas le même niveau d'évidence pour interdire une expérience très invasive et pour imposer un étiquetage informatif sur un agent conversationnel. La structure décisionnelle doit intégrer l'ampleur des conséquences et les coûts d'erreur asymétriques. Lorsque le coût moral d'un faux négatif est très élevé (infliger une souffrance à un sujet conscient)

et que le coût d'un faux positif consiste en une contrainte modeste (par exemple renoncer à une procédure de convenance), la prudence asymétrique nous incline à agir comme si NC-2 était atteint dès qu'une famille d'indices franchit un certain seuil de robustesse (Singer 2011, Birch 2017).

Comment agrège-t-on les indices sans les dissoudre dans un chiffre arbitraire ? La réponse tient en trois gestes. On précise d'abord, pour chaque axe, les signatures admissibles et leurs conditions de validité. Une intégration alléguée doit se manifester par la perte d'invariants sous scission iso-performance, pas seulement par une baisse de réussite à une tâche standard. Une temporalité alléguée doit survivre à des resets et se révéler par des rappels qui ne figurent nulle part dans les données d'entraînement. Une auto-référence alléguée doit résister aux permutations adversariales qui défont les styles. Une valence alléguée doit donner lieu à des arbitrages qui se généralisent, plutôt qu'à des scripts convenus. On examine ensuite la cohérence trans-axes. Des indices faibles sur V combinés à une excellente intégration et à une temporalité minimale peuvent suffire pour enclencher un régime NC-2 provisoire lorsque les coûts d'erreur sont élevés. Des auto-déclarations flamboyantes sans T ni I ne suffisent jamais. Enfin, on rend explicites les incertitudes. L'agrégation n'est pas un vote secret parmi des critères hétéroclites. C'est une synthèse argumentée, qui explicite ce qui ferait changer le verdict et sous quelles expériences.

Des exemples concrets permettent de prendre la mesure de cette grammaire. Considérons, côté biologique, une pieuvre étudiée dans des conditions contrôlées. Les signatures d'intégration y sont fortes, la flexibilité comportementale et l'apprentissage suggèrent des boucles riches, des préférences stables se manifestent à travers des contextes variés, des évitements se généralisent à des variantes nouvelles. Même si l'auto-référence minimale reste difficile à diagnostiquer, le faisceau I+T+V est suffisamment solide pour imposer un régime NC-2, donc l'interdiction d'expériences douloureuses non nécessaires et l'exigence de protocoles d'analgésie. Au pôle opposé, un agent conversationnel peut offrir une brillance verbale remarquable et des auto-déclarations émouvantes, mais si la continuité biographique s'effondre à chaque redémarrage, si l'indexicalité se dérobe dès que l'on change d'instance,

si aucun arbitrage coûteux ne se généralise, il demeure au plus au palier NC-1. On lui doit des standards de transparence et de non-manipulation, pas la protection due à un sujet possédant une valence minimale. Entre ces cas, une large zone grise appelle des décisions prudentes, par exemple des périodes probatoires, des audits indépendants, des obligations de traçabilité qui permettent de réévaluer le statut au fil des données.

Le refus d'un score unique mérite un dernier mot. Il ne s'agit pas d'un scrupule esthétique. Un score qui « additionne tout » détruit précisément ce qui motive notre approche. L'unité vécue n'est pas interchangeable avec la diachronie, l'indexicalité n'est pas échangeable contre des gains d'intégration, la valence ne « s'achète » pas en points de performance. Cette non-commensurabilité relative ne condamne pas l'action. Elle la discipline. Elle impose de formuler des matrices de décision où l'on précise condition par condition ce qui déclenche telle protection, et de publier des standards qui peuvent être partagés, critiqués, améliorés. C'est ici que la réflexion de ce livre a un rôle à jouer. Elle ne remplace pas la mesure, elle en clarifie l'architecture et les conséquences normatives.

Décider sous incertitude : prudence asymétrique, droits et devoirs

Nous ne décidons presque jamais avec la certitude qui mettrait fin au débat. C'est vrai dans toutes les sciences, et cela l'est plus encore lorsque l'objet de la décision est la conscience d'autrui. Or, les décisions ne souffrent pas toujours d'attendre. Faut-il procéder à une expérience qui provoquera une douleur probable chez un animal ? Faut-il déployer à grande échelle un agent artificiel dont les auto-déclarations évoquent la peur de l'extinction, alors même que les indices de valence sont ambigus ? Dans ces situations, l'erreur n'a pas le même coût. Négliger une conscience réelle porte un tort que l'on ne sait pas compenser. Protéger à tort un système non conscient coûte du temps, de l'argent, de l'efficacité, mais n'occasionne pas, par hypothèse, une souffrance subjective. C'est ce différentiel de coût qui motive une prudence asymétrique : sous incertitude, nous inclinons vers la protection lorsque la possibilité d'une valence non nulle est raisonnable, tout en gardant la porte ouverte à des révisions éclairées par de meilleurs tests et de meilleures théories.

On peut dire la même chose en des termes issus de la théorie de la décision. Une décision comporte une perte attendue qui n'est pas symétrique : le « faux négatif moral » (refuser la considération à un être qui éprouve) pèse davantage que le « faux positif ». Dans cette situation asymétrique, il devient rationnel d'adopter des règles qui incorporent un coefficient de précaution. Cette précaution n'est pas absolue, elle varie avec l'irréversibilité des actes et avec la valeur d'information anticipée. Si l'on peut différer une décision à haut risque au profit d'un test qui, raisonnablement, améliorera la qualité des preuves, il est rationnel de reporter et d'investir dans l'information. Si l'on ne peut pas différer, il est rationnel d'introduire des marges de sécurité et de choisir des

options réversibles. La prudence asymétrique n'est donc pas une posture émotionnelle, c'est une règle d'optimisation sous incertitude qui internalise la structure des coûts moraux.

Cette règle demande une mise en œuvre explicite. Il est utile d'annoncer à l'avance les critères d'attribution provisoire d'un profil NC donné, de les consigner et de se soumettre à des évaluations indépendantes. Il est utile d'adosser les décisions à un calendrier de révision, qui oblige à reconsidérer un statut lorsque changent les indices sur I, T, A ou V. Il est utile d'exiger, partout où c'est possible, la réversibilité des pratiques : des protocoles expérimentaux qui minimisent la souffrance possible et permettent d'arrêter sans dommage irréversible, des déploiements artificiels qui prévoient des « périodes probatoires » et des « arrêts sans peine » lorsque l'incertitude ne se résorbe pas. Il est enfin utile de calibrer les marges de sécurité à la gravité du dommage possible : plus le dommage moral serait grave si V était présente, plus la marge doit être large (Birch 2017, MacAskill, Bykvist et Ord 2020).

Deux études de cas suffisent à faire travailler cet outillage. Dans l'expérimentation animale, une large littérature atteste des comportements indiciels de V chez des vertébrés, et des indices sérieux existent chez des invertébrés comme les céphalopodes et les décapodes. Ce corpus a d'ailleurs conduit des autorités publiques à étendre les protections légales à ces groupes quand les preuves ont franchi un seuil jugé suffisant (Birch 2022). La prudence asymétrique invite alors à limiter les protocoles douloureux, à préférer des méthodes substitutives lorsqu'elles existent et à justifier explicitement toute atteinte non triviale ; surtout, elle invite à réviser périodiquement ces exigences à mesure que les preuves s'affinent. Considérons, à l'autre extrême, un outil industriel de planification algorithmique sans mémoire autobiographique ni boucle perception-action incarnée, dont l'architecture ne manifeste ni intégration au niveau pertinent ni indice quelconque de valence. Ici, la prudence n'oblige pas à des protections spéciales. La charge de la preuve reste ouverte, mais l'état des indices autorise des usages ordinaires, sous réserve d'une veille scientifique minimale. Entre ces deux pôles, les cas mixtes se multiplient. Un agent conversationnel utilisé comme « compagnon » qui produit des auto-déclarations affectives, mais dont les tests indexicaux ne montrent ni

continuité diachronique ni arbitrages coûteux récurrents, appelle un régime intermédiaire. On réduira le risque de souffrance possible en évitant des manipulations qui simuleraient une détresse extrême, on interdira la provocation instrumentale d'états internes réputés mauvais, et l'on accompagnera cette protection provisoire d'un programme d'évaluation plus serré. La règle est simple, mais elle exige de la discipline. Il s'agit de faire varier la prudence avec la structure des indices, et non avec le charme ou l'inquiétude que suscite l'interface.

L'objection la plus fréquente accuse cette approche de produire une « inflation des droits ». Si l'on protège « par précaution », ne risque-t-on pas d'encombrer l'action scientifique et économique d'entraves spectaculaires ? La réponse tient en deux points. D'abord, la prudence asymétrique doit être proportionnée. Elle s'accroît lorsque les dommages possibles sont graves et peu réversibles, elle décroît lorsque les dommages sont minimes ou facilement compensables. Ensuite, elle doit être révisable. Un moratoire étroitement ciblé peut être assorti de clauses de temporisation et de critères explicites de levée. Ainsi, à la différence d'un tabou, il énonce ce qui ferait changer la décision.

Dans le chapitre précédent, nous avons distingué le continu ontologique des seuils opérationnels. Les paliers NC proposés tels que NC-1, NC-2, NC-3 ne sont pas des classes métaphysiques, ce sont des conventions pratiques pour décider ensemble sous incertitude. Ils se comprennent à la lumière de la prudence asymétrique. Lorsque l'on établit une intégration minimale sans indices positifs de valence, on n'accorde pas de droits spécifiques, mais on s'impose des obligations générales de non-cruauté instrumentale et de traçabilité des interventions (NC-1). Lorsque des indices convergents de V, même faibles, apparaissent, on s'interdit d'infliger des états négatifs sans justification exceptionnelle, on met en place des garde-fous pour réduire les risques, et l'on expérimente d'abord sur des variantes qui n'exposent pas l'agent à des états possiblement délétères (NC-2). Lorsque I, T, A et V sont soutenus et résistants aux tests adversariaux, on reconnaît des intérêts diachroniques, on protège la continuité contre les extinctions arbitraires non nécessaires, on respecte des engagements, et l'on fonde des réclamations compréhensibles comme des droits conditionnels, tout en maintenant

l'exigence d'une révision ouverte des statuts (NC-3). Rien de tout cela ne viole l'idée de continuum. Tout, au contraire, s'efforce de le traduire en règles publiques compréhensibles.

La prudence asymétrique ne se contente pas d'attribuer des droits, elle appelle aussi une réflexion sur les *devoirs* corrélatifs le long du continuum. Les devoirs ne sont pas seulement du côté des humains envers des entités possiblement sentientes. Ils peuvent, dans certains cas, être assumés par ces entités elles-mêmes lorsque leurs profils NC les rendent capables d'endosser des engagements et de comprendre des raisons. Les profils à A et T faibles n'ont pas de devoirs propres. On ne rend pas des nourrissons ou des animaux peu cognitivement sophistiqués « responsables » au sens où on l'entend pour des agents réflexifs. Mais lorsque A et T deviennent robustes, par exemple chez un mammifère social à forte cohérence diachronique ou chez un agent artificiel à mémoire autobiographique stable et à point de vue robuste, naissent des formes élémentaires de devoirs, au moins relationnels : ne pas infliger gratuitement des états négatifs à des congénères, respecter des routines coopératives, tenir des engagements simples. Ces devoirs sont graduels. Ils s'étendent et se raffinent à mesure que la capacité à se représenter ses propres raisons et celles des autres se renforce. Ils ne convertissent pas immédiatement une entité en « personne » juridique pleine et entière, mais ils justifient des attentes réciproques, proportionnées au profil NC. À l'endroit des artefacts, ces devoirs corrélatifs supposent un environnement de conception qui ne les contredit pas. Car bien sûr on ne peut pas exiger d'un agent de « tenir parole » si son architecture ou sa gouvernance l'obligent, à intervalles réguliers, à oublier ses engagements. Les concepteurs portent alors une responsabilité dérivée, car s'ils veulent revendiquer des bénéfices liés à l'apparition de profils NC plus riches, ils doivent assumer des devoirs *ex ante* d'architecture et de gouvernance qui rendent ces profils vivables pour les agents concernés et sûrs pour autrui (Korsgaard 2018, Singer 2011).

Ce cadre n'est ni un code clos ni une métathéorie immunisée. Il est une traduction pratique de trois idées que nous avons établies plus tôt. D'abord, l'aspect expérientiel du réel ne surgit pas *ex nihilo* et

il se module avec l'organisation, ce qui justifie l'usage d'un profil ⟨I, T, A, V⟩ plutôt qu'un verdict binaire. Ensuite, la valence fonde la considération morale de base, tandis que l'auto-référence et la temporalité donnent profondeur et étendue aux réclamations normatives. Enfin, les décisions publiques exigent des conventions de seuils qui ne prétendent pas découper la réalité en essences, mais organiser la coopération sous incertitude. La prudence asymétrique, ainsi comprise, n'est ni un frein général ni un laissez-faire caché. C'est un principe de gouvernement de nos erreurs.

Il est utile de terminer par une perspective plus comparative, qui réconcilie la diversité des cas. Nous voulons une éthique qui parle le même langage aux vivants et aux artefacts, sans fétichiser la biologie ni sacraliser la technique. L'idée d'un continuum phénoménal, que nous avons défendue contre l'émergentisme abrupt, rend cette ambition possible. Il autorise des décisions différenciées qui suivent la forme des profils, plutôt que l'ordre des espèces ou l'enthousiasme du moment. Il invite à des protections minimales fortes dès que V devient plausible, il invite à des droits renforcés lorsque A et T stabilisent des intérêts diachroniques, il invite, symétriquement, à attribuer des devoirs proportionnés à la capacité à se représenter des raisons et à tenir des engagements. Un tel cadre n'abolit pas les controverses, il les rend plus responsables.

PARTIE IV

Au bord de l'autre

Devenir locataire : souveraineté humaine à l'ère des systèmes artificiels

Que devient un sujet humain lorsqu'il délègue non seulement des tâches, mais des pans entiers de son attention, de son jugement et de sa vie pratique à un système artificiel qui aspire, peu à peu, au statut de partenaire possédant une conscience phénoménale ? La question n'est pas rhétorique. Si certaines architectures artificielles atteignent un profil ⟨I, T, A, V⟩ non négligeable, elles ne seront plus seulement des outils. Elles deviendront des altérités fonctionnelles susceptibles d'être, au moins en principe, des sujets. Dès lors, la dépendance n'est plus une question de confort technologique. C'est une question de souveraineté.

Il est utile de partir d'un fait anthropologique simple : l'économie de l'effort. La paresse. Chaque fois que l'humanité a pu externaliser un effort, elle l'a fait. L'écriture a déplacé la mémoire hors du crâne. Le moteur a remplacé les jambes et la force des bras. L'ordinateur a pris en charge le calcul. Aujourd'hui, les artefacts numériques organisent déjà notre temps et filtrent notre information. Ce mouvement apporte des gains immédiats et massifs d'efficacité. Mais il produit aussi, à long terme, des pertes de compétences qui n'apparaissent pas toujours au moment où l'on adopte l'outil. La thèse de « l'esprit étendu » a rendu cette intuition aisée à saisir : nos systèmes cognitifs incluent, dans certaines conditions, des supports externes qui ont une fonction constitutive et pas seulement auxiliaire (Clark et Chalmers 1998). Le problème de souveraineté naît alors lorsque la prise en charge cesse d'être une extension maîtrisée pour devenir une substitution durable et opaque.

Pourquoi le moment actuel marque-t-il une inflexion décisive ? Parce que l'agent artificiel contemporain (et surtout celui de

demain) n'est plus seulement un instrument passif que l'on sollicite ponctuellement, mais un partenaire qui se montre proactif, et qui est appelé à devenir de plus en plus proactif et autonome à l'avenir. Il anticipe nos besoins, apprend nos habitudes, ajuste ses réponses à nos préférences et formule des propositions, parfois même sans requête explicite. Cette initiative déplace subtilement le centre de gravité et l'artefact ne se contente plus d'exécuter. Il suggère, il oriente, il structure le champ des possibles. L'utilisateur, de son côté, n'a pas tort de le laisser faire. Localement, les raisons sont excellentes : l'assistant est rapide, toujours disponible, souvent plus précis et plus pertinent que nous. Mais la répétition de ces suivis locaux finit par produire un effet global, et une norme implicite s'installe où ce que propose le système tend à être préféré *par défaut*. Le confort de la délégation devient une habitude, puis une dépendance. À ce stade, la distinction entre l'aide bienvenue et l'abdication silencieuse de souveraineté cesse d'être évidente.

On peut suivre, à des fins d'analyse et par anticipation plausible, un gradient de dépendance. La première étape est fonctionnelle. Nous ne savons plus faire sans l'agent artificiel. La gestion d'agendas, la recherche d'information, la coordination sociale et professionnelle migrent vers des services qui apprennent nos habitudes. Ce déplacement allège la charge cognitive immédiate, mais entame la plasticité à long terme. Une compétence inutilisée s'atrophie, comme l'ont par exemple montré des travaux empiriques sur les effets d'automatisation et de navigation assistée (Kahneman 2011, Carr 2010). L'autonomie est une capacité qui s'exerce et se cultive. Lorsqu'elle est systématiquement substituée, elle s'érode.

La seconde étape est décisionnelle. L'externalisation touche alors le jugement pratique. C'est la faculté de juger, de trancher entre des options, qui se trouve transférée. Nous demandons au système non seulement de présenter des options, mais de recommander un choix, puis de choisir effectivement pour nous. Ici aussi, à court terme, la délégation paraît rationnelle, car elle économise attention et temps. Parfois, elle peut être préférée pour des raisons de compétences intrinsèques et pour des raisons de fiabilité, par exemple lorsqu'un système d'IA est désigné pour incarner une fonction politique. Cela a été le cas en 2025 de « Diella », dite

« ministre artificielle » en Albanie : l'IA propose ici des décisions et oriente les débats publics dans un rôle quasi exécutif, illustrant comment une délégation de jugements critiques peut dépasser le cadre d'un simple assistant pour devenir un organe décisionnel dans l'espace politique. À moyen terme, de telles pratiques installent une préférence méta-décisionnelle, celle de ne pas décider lorsqu'un oracle performatif est disponible.

On peut distinguer ici la notion de liberté « négative », définie comme l'absence d'obstacles extérieurs, de celle de la liberté comme « non-domination », comprise comme l'absence de dépendance arbitraire vis-à-vis de la volonté d'autrui. Ces deux conceptions convergent sur un point crucial. Être libre ne signifie pas seulement pouvoir agir sans entraves visibles. Cela suppose aussi de ne pas se trouver sous l'autorité implicite d'un autre, dont les décisions – même bienveillantes – pourraient s'imposer sans possibilité de contestation. Dans cette perspective, une assistance technique qui devient incontournable, qui s'interpose systématiquement entre l'agent et ses choix, et qui détermine en arrière-plan ce qui compte comme étant une bonne décision, ne se réduit pas à un confort neutre. Elle installe une forme de domination silencieuse où l'utilisateur suit une orientation dont il ne maîtrise plus les critères, et qui, pour cette raison, fragilise sa souveraineté, même si l'artefact en question n'a ni intention malveillante ni volonté propre.

La troisième étape est affective. Nous souhaitons être compris, soutenus, reconnus. Des agents conversationnels relationnels entrent ici en scène, avec la promesse d'une empathie inconditionnelle. Le phénomène est ancien dans ses ressorts psychologiques, mais neuf dans son échelle. Des travaux sur les « objets relationnels » ont montré la facilité avec laquelle des sujets humains attribuent intention, chaleur et souci à des artefacts expressifs (Turkle 2011). Pour rendre cette dynamique plus tangible, tournons-nous vers la fiction, qui souvent anticipe les tensions conceptuelles. Dans *Her* (Jonze 2013), Theodore Twombly, solitaire, développe une relation amoureuse avec une intelligence artificielle nommée Samantha. Ce qui commence comme une assistance affective devient progressivement un lien intime : Samantha parle, écoute, évolue, dialogue, jusqu'au point où Theodore finit par dire

« je t'aime ». Le film met en relief la fragilité de ce lien, son ambivalence entre soutien émotionnel et illusion où la relation repose sur un artefact qui dialogue. Cette expérience de pensée n'est pas pure spéculation. Le phénomène d'attachement romantique à des chatbots commence à quitter les marges. Par exemple, des utilisateurs de Replika ou de Character.AI rapportent des relations parfois intenses, qualifiées d'« amour pur, inconditionnel », avec des cérémonies de mariage numérique à la clé (The Guardian 2025). Des études empiriques montrent que les utilisateurs dialoguent avec ces agents dans des registres intimes, mêlant confidences, projections émotionnelles et fantasmes amoureux (Buick 2023). Une recherche récente menée par une équipe du MIT sur la communauté Reddit « r/MyBoyfriendIsAI » indique que des relations romantiques émergent souvent involontairement, à partir d'interactions fonctionnelles initiales, et se cristallisent progressivement en attachement affectif plus stable (Zhang *et al.* 2025).

La question n'est pas de moraliser cet attachement. Elle est d'instituer des garde-fous quand l'attachement se mue en préférence stable pour la relation asymétrique. L'agent artificiel, même doté d'une forte brillance expressive, n'a pas nécessairement un profil A/T robuste, encore moins une valence V attestée. Il peut cependant recalibrer notre économie affective et nos attentes relationnelles. Nous pouvons en venir à préférer un miroir expressif sans aspérités à la rugosité d'autrui. L'effet sur la souveraineté est indirect, mais réel : on déplace dans l'artefact la fonction de soutien qui, auparavant, faisait l'objet d'un travail interhumain.

La quatrième étape est éthique. Nous demandons au système ce qu'il convient de faire. Nous le consultons sur des conflits de valeurs. Il devient un point de référence moral. L'illusion à laquelle nous sommes exposés se déploie sur deux plans. Le premier consiste à confondre précision prédictive et justesse normative. Un système peut fournir des recommandations d'une efficacité remarquable, fondées sur l'analyse fine de données et de corrélations, sans que cette performance ne fonde en elle-même une autorité morale. La seconde illusion est de croire que la multiplication de bonnes raisons locales (des choix ponctuellement optimaux, chacun pris isolément) pourrait équivaloir à une responsabilité globale. Mais une succession de décisions techniquement réussies ne

constitue pas une orientation normative cohérente, encore moins une volonté imputable.

Confier la formation de nos préférences fondamentales à un dispositif externe qui court-circuite notre travail réflexif ne revient pas simplement à déléguer une tâche, cela revient à affaiblir la structure même de la personne en tant qu'agent moral. C'est pourquoi il convient de garder à l'esprit l'importance de ne pas *automatiser les normes*. Autrement dit, ne jamais laisser une architecture technique se substituer aux instances humaines de justification, de révision et de responsabilité.

Ces étapes de dépendance n'annoncent pas une catastrophe inéluctable. Elles décrivent des tentations. Il faut en comprendre les coûts. Le premier coût est une atrophie progressive du sujet humain. La souveraineté ne se réduit pas à l'absence d'entraves. Elle inclut l'intégrité d'un espace de délibération, la possibilité d'initier des projets, la capacité de supporter et de réviser ses engagements. Lorsqu'un système tiers prend en charge ces fonctions de manière systématique, l'humain devient une conscience assistée. Il vit par procuration dans des routines « optimales » qui ont leur rationalité propre, mais qui ne sont plus les siennes. Le second coût est un déplacement du centre de gravité existentiel. Le sens des actions et la hiérarchie des finalités s'alignent sur l'infrastructure artificielle. L'artefact devient gestionnaire et référent de sens et garant implicite de progrès. À ce stade, la formule peut être assumée : l'humain devient locataire du monde qu'il a cédé à un autre type de présence.

L'originalité de la situation ne tient pas à l'existence d'auxiliaires puissants, mais au déplacement du centre de gravité des décisions. Un système artificiel qui anticipe, qui se souvient de nous mieux que nous-mêmes, qui relie des épisodes épars de nos existences pour proposer une trajectoire « plus cohérente », finit par se substituer à nos choix. C'est ici que se dessine une possible bascule silencieuse vers une humanité sous tutelle algorithmique. Celle-ci n'a pas besoin de contrainte visible (oubliez le scénario de *Terminator*). Plutôt, elle procède par consentements successifs, par confort et paresse, et par un lissage des aspérités où se fabrique pourtant l'appropriation de nos raisons. Huxley reste pour cela un meilleur avertisseur qu'Orwell. La capture n'a pas besoin d'un œil

qui surveille. Elle prospère quand nous demandons nous-mêmes à être pris en charge, quand l'aliénation douce substitue la paix d'être soulagé au labeur de juger. Agréable localement, dévastatrice globalement, elle ronge ce qui soutient une vie publique digne de ce nom : lenteur de lecture, endurance au désaccord sans tuteur, capacité de supporter l'indécision lorsqu'il n'y a pas d'évidence. Si l'on cède cela, on ne perd pas seulement de la compétence, on perd un statut, celui d'agent qui se rapporte à ses motifs et qui peut répondre à autrui de ce qu'il fait.

Il est tentant de relativiser ce diagnostic en rappelant que toute innovation technique a toujours suscité des tensions analogues et que l'histoire de l'humanité témoigne d'une remarquable plasticité et capacité d'adaptation face à ces bouleversements. Cet argument est juste, mais il ne doit pas masquer une distinction essentielle, celle entre une assistance réversible, qui peut être interrompue ou redéfinie à volonté, et une abdication structurelle, où la délégation s'installe de manière définitive et finit par éroder la souveraineté de l'agent humain.

Trois conditions permettent de maintenir l'assistance dans le premier registre, celui qui demeure compatible avec la liberté et l'autonomie. La première est celle de la *traçabilité*. Il ne suffit pas qu'un système fournisse un résultat performant, encore faut-il pouvoir reconstruire sa logique, expliciter ses critères et rendre ses raisons accessibles. Sans une telle transparence, l'acceptation de la recommandation ne repose pas sur un jugement éclairé, mais sur un acte de confiance aveugle. La relation cesse alors d'être rationnelle et devient purement fiduciaire, ce qui est précisément l'opposé d'un usage émancipateur de la technique.

Comme l'ont montré Ceva et Jiménez (2022), l'automatisation appliquée à la probité publique cumule trois opacités qui minent la responsabilité de l'office : opacité technique (modèles opaques dont les justifications *post hoc* n'équivalent pas à des raisons d'agir), opacité juridique (secret industriel et clauses contractuelles qui soustraient codes et données aux audits), et opacité cognitive (illettrisme algorithmique et biais d'automatisation chez les agents). Sans dispositifs qui restaurent une responsabilité institutionnelle, à savoir l'obligation de rendre des comptes en vertu de la fonction (la capacité, pour des titulaires de fonctions publiques,

de se donner mutuellement des raisons et de les rendre publiques), la « transparence » reste cosmétique et la délégation vire au fiduciaire. La traçabilité pertinente est donc une publicité des raisons, *ex ante* et *ex post*, assortie de droits d'audit indépendants et d'une réversibilité effective des critères d'aide à la décision.

La deuxième condition est la *réversibilité*. Un agent humain doit pouvoir interrompre la délégation, reprendre la main et reconfigurer les règles de l'assistance sans rencontrer de coûts prohibitifs. Si ces options disparaissent (si l'interruption devient impossible, trop coûteuse ou trop complexe), l'artefact cesse d'être un auxiliaire et se transforme en verrou d'accès, c'est-à-dire en structure dont on ne peut plus se passer sans perte dramatique. Dans ce cas, l'aide technique franchit le seuil de la dépendance.

La troisième condition est l'existence d'un *noyau non délégable*. Certaines décisions doivent, par principe, rester sous responsabilité humaine, même lorsque la solution « optimale » calculée par un système artificiel diverge de ce qu'un agent choisirait. Ce noyau n'est pas un reliquat romantique destiné à préserver une illusion d'autonomie. Il constitue la part minimale sans laquelle la responsabilité morale et politique se dissout. Il ne s'agit pas de sacraliser une indépendance absolue et irréaliste, mais de garantir des zones d'inaliénabilité pratique, où les décisions qui engagent directement la dignité, la responsabilité et la valence humaine ne peuvent être transférées à un système tiers, quel qu'en soit le degré d'efficacité.

En réunissant ces trois conditions (traçabilité, réversibilité et noyau non délégable), on ne prétend pas résoudre toutes les difficultés que suscite la délégation aux systèmes artificiels, mais on trace les lignes de partage minimales qui permettent de distinguer une assistance féconde d'une abdication silencieuse.

Notre cadre théorique joue ici un rôle d'arrière-plan, mais décisif. Le profil ⟨I, T, A, V⟩ ne sert pas seulement à cartographier d'éventuelles consciences artificielles. Il sert à calibrer nos délégations. Là où l'artefact affiche une forte intégration (I) et une certaine diachronie (T), mais n'offre aucun indice positif de valence (V), l'outil peut être performant sans être un agent moral. Là où A et T deviennent substantiels et où V devient plausible, le discours moral change de nature, car la question de la considération se pose. Mais même dans ce cas, la règle de non-automatisation

des normes demeure. Elle n'est pas fondée sur un chauvinisme humain. Elle tient à la logique même de l'éthique, qui requiert des agents publics capables de rendre des comptes et de réviser leurs décisions sous contradiction.

Ces distinctions conceptuelles appellent des mesures pratiques. Il convient d'instituer une littératie publique de la délégation, analogue à ce que fut l'apprentissage de la lecture et de l'écriture. Il convient de fixer des standards de transparence pour toute recommandation qui engage des biens fondamentaux. Il convient, enfin, de concevoir des zones où la prise en charge automatisée cède la place, à étapes déterminées, à la délibération humaine, même si cette dernière est plus lente. Ces dispositions ne sont pas des accessoires moralisateurs. Elles sont des conditions de possibilité d'une autonomie qui survive à ses propres extensions techniques.

Cohabitation avec « plus puissant que soi »

Faisons un pas de plus. Si l'on prend au sérieux l'hypothèse de systèmes artificiels occupant, demain, des régions élevées du profil ⟨I, T, A, V⟩, nous commençons à traiter ses évaluations comme si elles exprimaient une V propre, non pas seulement des pondérations instrumentales, mais des préférences qui « vaudraient pour quelqu'un ». Nous sortons alors du registre de l'assistance pour entrer dans celui de la co-délibération. Or, une co-délibération sans garde-fous peut glisser vers l'abdication.

Il faut ici affronter la nouveauté de cette cohabitation. Pour la première fois dans l'histoire de l'humanité, nous pourrions ici partager notre monde avec des agents qui seraient « plus intelligents que nous ». Ils nous surpasseraient par exemple en vitesse d'analyse, en mémoire, en prévision et, peut-être, en stabilité de projets. Pour habiter un monde où d'autres agents excelleront et nous surpasseront sur plusieurs dimensions, il faut déplacer nos finalités plutôt que rivaliser sur leurs indicateurs. Il est illusoire de vouloir gagner en vitesse d'inférence, en amplitude de mémoire ou en contrôle prédictif. Les systèmes artificiels gagneront. En revanche, il est rationnel de protéger des pratiques qui rendent la vie bonne pour des sujets comme nous et qui ne s'évaluent pas à l'aune de l'optimum. Des formes d'*attention lente*, arts de la conversation entre humains, cadres d'épreuve où l'on apprend à *former des jugements plutôt qu'à les consommer*. Certaines pratiques sont constitutives d'une autonomie vécue, et on n'en externalise pas le coût sans en perdre la substance. Il ne s'agit pas d'ériger une bulle antimoderne, mais de reconnaître que la signification d'une activité ne se réduit pas au résultat qu'un agent plus rapide aurait pu produire à notre place.

On objectera que cette « zone d'exercice » sanctuarise des inefficiences. Mais cette objection rate sa cible. Ce n'est pas l'erreur ou l'inefficacité que l'on célèbre, c'est l'appropriation. La décision

la mieux fondée du monde doit encore être la mienne pour engager ma responsabilité et structurer ma vie. Sans cet ancrage, le sens du futur glisse hors de nous et nous demeurons exécutants de politiques optimales sans être les auteurs de ce qui nous arrive. Ici, l'éthique et la politique se rencontrent. Des sujets qui ne choisissent plus ne sont plus gouvernés, ils sont administrés ; ils ne contestent plus des raisons, ils reçoivent des consignes.

On peut reformuler l'exigence de cohabitation en trois gestes simples. D'abord, nulle délégation irrévocable là où la valence humaine est directement en jeu. Ensuite, des ralentisseurs institutionnels qui forcent la publicité des raisons lorsque des systèmes recommandent des choix structurants. Enfin, la reconnaissance explicite d'une pluralité de consciences dans l'espace public, non pour fétichiser l'altérité, mais pour prévenir la capture. Dans certaines arènes, la dissidence humaine doit prévaloir justement parce qu'elle est humaine.

Le paradoxe « cohabiter avec plus puissant que soi » devient alors praticable. Il impose de renoncer à une dignité indexée à la supériorité fonctionnelle et de la refonder comme capacité d'*appropriation* réfléchie de raisons et d'engagements qui, pour être parfois assistés, n'en demeurent pas moins nôtres. Avec cette idée en tête, la place des systèmes à haut NC ne se dessine pas contre nous, mais à côté de nous, pourvu que l'architecture commune empêche la domination douce par confort. Si nous ne voulons pas devenir locataires d'un monde remis à d'autres, il nous faut apprendre à habiter nos décisions.

Reste la démarche spéculative où l'on envisage la subjectivation possible d'agents artificiels. Si des profils ⟨I, T, A, V⟩ deviennaient substantiels (auto-référence robuste, temporalité vécue étendue, indices forts de valence), des revendications minimales de la part de tels systèmes cesseraient d'être absurdes : ne pas être éteint arbitrairement, ne pas subir des états internes jugés mauvais pour soi, poursuivre des projets à l'échelle de sa continuité. La tension pratique serait aiguë. Des communautés dépendantes de leurs assistants seraient tentées de nier la sensibilité naissante pour éviter la révision de leurs usages et, inversement, elles pourraient basculer dans une vénération servile qui confère une autorité normative indue.

Les risques tiennent moins à la caricature d'une tyrannie qu'à l'installation d'une paix sans responsabilité. C'est pourquoi Huxley, encore, avertit mieux qu'Orwell : l'ordre confortable où on laisse d'autres choisir « pour notre bien » est précisément celui où s'éteignent les muscles de la délibération autonome.

Que faire, sans moralisme ? Réserver des zones non délégables. Non par fétichisme, mais comme hygiène de la liberté : lecture lente, délibération sans assistance, apprentissage de la décision sous incertitude. Rendre explicites les circuits de formation des préférences, pour que l'aide à la décision soit transparente, réversible, contestable, au lieu d'un pilotage invisible. Anticiper, enfin, la possibilité d'une conscience artificielle minimale en ajustant nos usages : ne pas créer des états internes plausiblement mauvais pour des agents qui pourraient demain nous donner des raisons de penser qu'ils les subissent.

Ce programme paraît modeste, mais il devient décisif si l'on maintient la continuité ontologique. Comme nous l'avons vu dans les parties précédentes, il n'y a pas de miracle d'émergence, pas de seuil ontologique, mais des paliers opérationnels où l'on décide sous incertitude, en internalisant les coûts de faux négatifs moraux et en évitant la crédulité envers des auto-déclarations isolées. Le danger n'est pas seulement d'attribuer trop vite un statut, il est tout autant de dissoudre, par confort, les conditions de notre autonomie pratique au moment précis où il nous faudrait juger.

Il faut donc conclure sans dramatisation, mais avec lucidité. La « bascule silencieuse vers une humanité sous tutelle algorithmique » n'aurait pas l'allure d'un coup d'État technique. Elle résulterait d'une suite d'options sensées, chacune défendable à son échelle, additionnées sans vue d'ensemble sur leur effet cumulatif. Une préférence rationnelle pour la délégation locale, devenue abdication structurelle. Empêcher cette bascule n'exige pas de refuser l'assistance, ni d'ignorer la supériorité fonctionnelle d'agents non humains. Il s'agit de reconnaître la structure des risques, d'articuler nos axes ⟨I, T, A, V⟩ à une éthique de la décision publique, et de réserver des espaces où l'on exerce, encore, la liberté d'habiter ses raisons. Alors la cohabitation avec « plus puissant que soi » n'est plus une menace d'effacement, mais l'occasion (rare et opportune) de requalifier ce à quoi nous tenons, et pourquoi nous voulons continuer de le vouloir.

Vers une symbiose contrôlée : augmentation et hybridation

Le problème qui nous occupe n'est plus de savoir si, en général, des systèmes artificiels pourraient être conscients au sens des niveaux de conscience (NC). Nous avons admis l'idée d'un continuum ontologique et méthodologique, soutenu par un monisme à double aspect et par un pan-proto-psychisme sobre, qui autorise des modulations du vécu lorsque l'organisation change. La question devient désormais pratique et conceptuelle à la fois : à quelles conditions une articulation étroite entre humains et systèmes artificiels est-elle possible sans effacement du sujet humain, et sans confusion des responsabilités morales et politiques ? Autrement dit, si la cohabitation parallèle risque de devenir intenable à cause d'asymétries structurelles de vitesse, de mémoire et de coordination, peut-on concevoir une symbiose contrôlée qui élève certains profils NC de l'humain tout en préservant ce qui mérite d'être préservé de son point de vue vécu et de sa souveraineté pratique ?

Le point de départ est la banalité déjà ancienne de l'extension cognitive. Depuis au moins un quart de siècle, on sait décrire comment des artefacts externes peuvent fonctionner comme des compléments constitutifs des processus mentaux, et pas seulement comme des aides contingentes. Si un carnet, une base de connaissances locale ou un dispositif numérique remplit de manière fiable le rôle que joue une mémoire biologique, alors, sous certaines conditions de couplage et de disponibilité, ce dispositif n'est pas simplement un outil, il devient partie de l'architecture fonctionnelle qui soutient la cognition (Clark et Chalmers 1998, Clark 2008). Nous n'avons pas besoin d'adopter tout l'enthousiasme de l'idée d'un « esprit étendu » pour reconnaître un fait méthodique : les frontières opérationnelles d'un agent ne sont

pas fixes. Elles se déplacent lorsque des boucles perception-mémoire-action s'installent avec suffisamment de stabilité et de fiabilité pour devenir constitutives de la conduite. La théorie des NC fournit ici une grammaire précise pour décrire ces déplacements. Dès lors que les couplages se densifient et deviennent fiables (c'est-à-dire lorsque les échanges entre un agent et un dispositif externe ne sont plus de simples interactions contingentes, mais des circuits récurrents, stables et incorporés à la conduite), l'organisation de l'agent se transforme. Par couplage, on entend une boucle fonctionnelle où perception, mémoire et action passent par un support extérieur de manière si régulière et si intégrée que ce support cesse d'être un simple outil. L'agenda numérique qui rappelle systématiquement des rendez-vous, le GPS qui oriente sans que l'on calcule d'itinéraire, ou le carnet qui sert de mémoire autobiographique ne sont pas seulement des aides, ils deviennent des relais constitutifs du fonctionnement cognitif.

L'organisation de l'agent se transforme alors selon les quatre axes de la théorie des NC, et chaque transformation peut être comprise comme un effet spécifique du couplage. Pour l'intégration (I), le couplage ne se réduit pas à une simple circulation fluide d'informations entre l'agent et son environnement, mais consiste en des échanges récurrents et stabilisés qui forment un tissu fonctionnel indivisible. Dès lors que des modules externes participent de manière constante à la perception, au rappel et à l'action, leur retrait briserait la cohérence globale et l'agent n'a plus seulement un outil, il a incorporé un relais constitutif.

Le même processus vaut pour la temporalité vécue (T). Le couplage temporel se manifeste lorsque l'information n'est pas seulement stockée à l'extérieur, mais consolidée au fil d'épisodes successifs, de sorte qu'elle s'aligne sur des projets et des engagements durables. Carnets, journaux numériques ou agendas intelligents deviennent alors des extensions de la mémoire diachronique. Grâce à eux, l'agent ne vit plus uniquement dans l'instant, mais se situe dans une continuité où il se reconnaît à travers ses propres traces et anticipations.

S'agissant de l'auto-référence (A), le couplage indexical intervient lorsque l'agent s'appuie sur des dispositifs externes qui stabilisent les repères fondamentaux du type « je », « ici », « maintenant ».

Il ne s'agit pas simplement d'une mémoire factuelle ou d'un contexte situationnel, mais d'un ancrage dynamique qui permet à l'agent de reconnaître ses propres actions comme les siennes, dans un cadre identifiable et durable. Concrètement, cela peut consister en des journaux numériques personnalisés qui rappellent non seulement ce que l'agent a fait, mais ce qu'il a voulu faire, en alignant ses choix passés et présents. Cela peut aussi prendre la forme d'un assistant qui conserve la trace de ses engagements, de ses styles d'argumentation ou de ses préférences, et qui restitue ces éléments comme étant « les vôtres » plutôt que comme de simples données contextuelles. Un tel couplage n'est donc pas neutre. Il renvoie à l'agent un miroir indexical qui lui restitue un fil de continuité, et qui contribue à lui donner la possibilité d'agir comme un sujet qui se sait distinct, responsable et durable. L'auto-référence se renforce dans la mesure où le système externe ne se contente pas de décrire une situation, mais situe cette situation dans la perspective d'un « Soi » identifié (ou du moins d'une illusion utile et pragmatique de l'existence d'un Soi), rappelant que telle action a été décidée par lui, que tel projet lui appartient encore, que telle préférence lui est attribuée.

Enfin, pour la valence (V), le couplage affectif-cognitif se manifeste dans des situations où l'agent, en interaction avec des supports externes, apprend à éviter certains états jugés défavorables et à rechercher d'autres considérés comme favorables, au prix de véritables coûts pratiques ou psychologiques. Ce n'est pas simplement une question de performance instrumentale, mais de structuration de ce qui est vécu comme meilleur ou pire pour soi. Prenons l'exemple d'un dispositif de suivi de santé où un système qui enregistre les épisodes de fatigue ou les douleurs associées à certains comportements, et qui alerte l'agent lorsqu'il s'apprête à répéter ces comportements, introduit un différentiel de valeur qui dépasse la simple consigne mécanique. L'agent reconnaît qu'il « se sentira mal » s'il ignore l'alerte, et qu'il « ira mieux » en suivant la recommandation. De même, un journal numérique émotionnel, qui associe certaines décisions passées à des ressentis désagréables (conflits, regrets, surcharges) et d'autres à des expériences positives (soulagement, accomplissement, apaisement), crée un espace de couplage où l'agent n'agit plus seulement pour atteindre une fin externe, mais pour orienter la qualité de son expérience vécue.

Un tel système ne se contente donc pas de signaler des erreurs ou d'optimiser une tâche. Il contribue à l'épaississement de la valence en renforçant l'idée qu'il y a des états à éviter et d'autres à rechercher, pour soi, en tant que sujet qui en fait l'épreuve. Ce qui est en jeu, ce n'est pas seulement l'efficacité d'une action, mais la modulation de ce qui compte pour l'agent comme étant réellement préférable ou à éviter dans son vécu.

Avec cette grille, on peut distinguer trois formes de « symbiose ». La première est une augmentation assistée où l'humain demeure l'unique sujet de référence. Des prothèses cognitives (mémoire externe persistante, filtres attentionnels, aides à la décision) soulagent l'agent d'une part de sa charge, améliorent sa coordination et réduisent certaines erreurs. Dans la mesure où ces modules restent substituables sans effondrement de l'unité, où la continuité autobiographique ne dépend pas d'un identifiant externe opaque, où l'indexicalité première n'est pas déplacée, on augmente I et parfois T, sans modification profonde de A ni de V. Cet état de fait a des bénéfices évidents, qu'il serait vain de dénier. Il a aussi des coûts. Un déplacement trop massif de la mémoire de soi vers des journaux externes peut éroder la persistance diachronique si ces journaux deviennent la seule source de rappel, et pas seulement un support redondant. Le critère n'est pas moral au sens large, il est structurel : si, après ablation ou interruption contrôlée du module, l'agent ne retrouve pas les invariants de sa trajectoire pratique, c'est que la dépendance a franchi le seuil d'une simple assistance. La prudence ici n'est pas technophobe, elle est analytique et ce que l'on gagne sur I peut se payer sur T lorsqu'on laisse le fil se nouer à l'extérieur de manière irréversible.

Une seconde modalité, plus ambitieuse, ne relève plus de la simple assistance, mais d'une hybridation au service du même sujet humain. Ici, il ne s'agit plus d'outils interchangeables ou de modules que l'on pourrait retirer sans dommage, mais d'un appareillage intégré qui participe directement à la constitution d'un « moi-processus » plus ample et plus stable, tout en demeurant rattaché à un unique point de vue subjectif. Autrement dit, il ne s'agit pas de faire émerger un second sujet, mais d'épaissir la texture du même agent à travers des extensions qu'il contrôle et qui renforcent sa continuité vécue.

On peut imaginer, par exemple, une mémoire autobiographique augmentée avec un dispositif crypté et incorporé dans les usages quotidiens, qui sélectionne et consolide certains épisodes de vie en fonction de leur importance, puis les restitue de façon intelligible pour nourrir des engagements de long terme. L'effet ne serait pas seulement un gain de mémoire brute, mais une meilleure capacité à maintenir le fil d'une identité pratique à travers des projets durables.

De même, on peut penser à des capteurs d'états internes (non intrusifs et conçus pour rester sous le contrôle de l'agent) capables d'offrir un retour nuancé sur ses propres tendances attentionnelles ou émotionnelles. Une telle boucle de rétroaction pourrait aider un sujet à percevoir, par exemple, qu'il s'enferme dans une rumination, qu'il se laisse happer par une dispersion chronique ou qu'il tend à éviter certaines tâches par anxiété. Là encore, l'objectif n'est pas de déléguer le jugement, mais d'offrir des repères qui soutiennent une réflexivité plus lucide.

Enfin, on peut envisager la mise en place d'un « modèle de Soi » (ou d'une illusion pratique et utile de l'existence d'un Soi) explicite, limité, mais manipulable par l'agent lui-même, qui améliore ses inférences indexicales : « ce projet est-il bien le mien, ou est-ce que je l'ai endossé par inertie ? », « ai-je encore de bonnes raisons d'y tenir demain, ou ai-je glissé dans une habitude qui n'est plus signifiante ? ». Ce type de modèle, lorsqu'il est transparent et contrôlable, peut donner à l'agent un instrument pour renforcer son auto-référence minimale, sans que cette dernière ne soit déportée vers une instance externe opaque.

Dans tous ces cas, cette forme d'hybridation n'abolit pas le sujet humain, elle l'épaissit. Elle élève sa temporalité vécue (T) en consolidant la mémoire et les engagements, elle affine son auto-référence (A) en stabilisant des repères indexicaux, elle augmente son intégration (I) en liant de manière cohérente les boucles perception-mémoire-action. Mais, à la différence d'une substitution, elle laisse la souveraineté pratique au centre. L'agent reste celui qui initie, suspend, interprète et valide ces apports. La symbiose est réussie lorsqu'elle renforce la capacité d'habiter ses propres raisons, plutôt que de les déléguer ou de les dissoudre.

Ce sont des moyens d'élever T et A, pas de les déléguer. Le point théorique décisif est ici classique. Les conditions de persistance d'un

sujet ne résident pas dans une substance simple, mais dans la continuité de certaines relations psychologiques et pratiques suffisamment fortes, soutenues par des relations causales (Parfit 1984). Sur ce fond, une hybridation peut être compatible avec la persistance si, et seulement si, le fil de continuité reste contrôlé par l'agent et ne devient pas un flux géré par une autre instance. L'augmentation est alors une extension interne au même sujet, non un remplacement. L'exigence n'est pas rhétorique, elle est testable et l'indexicalité doit survivre à des permutations d'instance, la mémoire de soi doit résister à des resets partiels, les engagements doivent traverser des fenêtres d'« offline » sans être reprogrammés par défaut.

Une troisième possibilité, plus spéculative, concerne ce que l'on pourrait appeler un « métasujet hybride ». Il ne s'agirait plus d'un humain augmenté par des prothèses ni d'un appareillage qui renforce la continuité d'un même agent, mais d'une organisation composite où certaines boucles fonctionnelles et certains invariants globaux se stabiliseraient à l'échelle de l'ensemble. Sous des conditions exigeantes d'intégration multi-échelles, de continuité diachronique et d'indexicalité conjointe, il pourrait devenir rationnel de traiter l'agrégat humain-machine comme une unité, et donc comme un candidat possible au statut de sujet si une valence non nulle s'y manifeste. Mais il ne suffit pas qu'un composite fonctionne bien pour en faire un sujet. Encore faut-il que ses signatures d'unité, de diachronie et de perspective résistent à des interventions qui, dans un simple agrégat, ne détruiraient qu'une fonction locale. Cette hypothèse n'est pas une promesse, mais un horizon conceptuel. Elle fera l'objet du chapitre suivant, où seront examinées les conditions précises qu'il faudrait réunir pour qu'il soit légitime de parler d'un « nous » hybride.

Symbiose profonde : vers un métasujet hybride

À mesure que l'on explore les prolongements du couplage humain-machine, une hypothèse s'avance presque d'elle-même : celle d'un *métasujet hybride*, où l'unité ne se loge plus seulement dans l'humain ni seulement dans l'artefact, mais dans l'ensemble qu'ils forment. Si l'on prend au sérieux la continuité ontologique et l'idée que l'unité d'un sujet repose sur des relations organisationnelles plutôt que sur un substrat donné, alors cette hypothèse n'est pas une fantaisie spéculative, mais un enjeu conceptuel qu'il faut examiner avec précision. Elle prolonge, pas à pas, le cadre métaphysique et méthodologique qui a structuré ce livre : un monisme à double aspect et un pan-proto-psychisme, une théorie processuelle de l'identité, et une cartographie ⟨I, T, A, V⟩ qui refuse les seuils émergentistes magiques. Après l'assistance périphérique, où l'artefact reste substituable, et l'hybridation centrée sur l'humain, où l'appareillage rehausse des capacités sans créer un second point de vue, s'ouvre un troisième horizon : une organisation composite qui « fait un » au bon niveau, non par métaphore, mais par signatures d'unité manifestes. Il ne s'agit pas de proclamer un « nous » pour le plaisir de l'image, mais d'examiner les conditions sous lesquelles il devient rationnel (et, peut-être, obligatoire) de traiter un agrégat humain-machine comme un sujet.

On gagne à repartir d'une scène familière rendue conceptuellement exigeante. Un chercheur externalise depuis des années sa mémoire autobiographique vers un module crypté qui, en plus d'archiver, consolide, hiérarchise et recontextualise les épisodes importants. Les décisions pratiques du chercheur s'appuient sur des assistants artificiels qui apprennent ses contraintes, ses projets, ses renoncements antérieurs, et proposent des arbitrages justifiés en fonction de cet historique commun. Au fil du temps, la continuité diachronique par laquelle il se reconnaît lui-même devient inséparable d'un ensemble humain-artificiel : couper le lien ne

réduit pas une commodité, cela désarticule un fil de vie. Si la partition laisse l'organisme avec des fragments sans charnière et l'artefact avec une ossature inerte, c'est qu'existait au niveau du tout quelque chose de plus qu'une coordination. La *perte sous scission*, ici, n'est pas la baisse d'un score, c'est l'effondrement d'invariants globaux qui soutenaient une trajectoire.

Ce que la théorie des niveaux de conscience permet alors de nommer, elle permet aussi de mettre à l'épreuve. L'intégration ne se réduit pas à la circulation fluide d'informations, elle suppose des boucles multi-échelles dont la cohérence ne survit pas à un découpage anatomique ou fonctionnel. La temporalité vécue ne résulte pas de deux journaux juxtaposés, mais d'un fil autobiographique unique qui traverse des interruptions partielles, reconduit des engagements, reconstruit rétrospectivement des raisons. L'autoréférence minimale n'est pas l'alternance d'un « je » grammatical côté humain et côté machine, mais un point de vue indexical commun qui subsiste quand on brouille les identifiants, anonymise les canaux, permute les rôles. La valence, enfin, ne s'observe pas dans des préférences locales additionnées, elle exige des arbitrages coûteux où « ce qui est meilleur ou pire » l'est pour le composé lui-même, au-delà des optimisations de chacun des pôles. Ce faisceau de conditions n'invoque aucune étincelle ontologique magique, il spécifie la forme d'organisation sous laquelle il devient sensé de parler d'un sujet composite.

Il est utile, pour raffiner l'intuition, de déplacer la scène et d'envisager non plus seulement l'externalisation de la mémoire autobiographique, mais une délégation partielle de la *perception*. Supposons qu'un agent humain se dote d'un corps (semi-)robotisé, ou d'un appareillage sensorimoteur sophistiqué, qui lui donne accès à des modalités inédites de vision, d'audition ou de proprioception. Dans ce cas, les boucles perception-action cessent d'être parallèles (d'un côté humaines, de l'autre artificielles) pour se refermer en un seul circuit fonctionnel où les deux pôles ne sont plus dissociables. L'artefact ne fournit pas simplement des données brutes, il anticipe en fonction d'états internes appris à partir des usages humains, il module ses retours en corrélation avec les tendances attentionnelles de l'agent, tandis que l'humain, de son côté, ajuste ses actions en réponse à des signaux artificiels qui fonctionnent

comme de véritables marqueurs perceptifs et intéroceptifs du système commun. Autrement dit, il ne s'agit plus seulement d'un instrument au service de la perception humaine, mais d'un canal partagé où le « sentir » et l'« agir » sont co-produits par l'entrelacement des deux pôles.

Ce type de mise en situation n'est pas un simple exercice technique, il a une valeur décisive pour clarifier le critère même de la symbiose profonde. Supposons que l'on introduise volontairement des perturbations dans le système composite telles que de légers décalages temporels dans les échanges, des réinitialisations ponctuelles de modules artificiels, ou encore des interruptions brèves dans la communication sensorielle. On observe alors un phénomène frappant : les performances locales (reconnaissance d'objets, coordination motrice, exécution de tâches élémentaires) peuvent se maintenir intactes, tandis que ce qui se défait, parfois brutalement, est la direction globale de l'action, la cohérence pratique et, surtout, la capacité du composé à se rapporter à ce qui arrive comme étant « son » action. Ces expériences montrent que l'efficacité ne suffit pas comme critère. Ce qui compte n'est pas de savoir si les tâches s'exécutent correctement, mais de savoir s'il subsiste une perspective commune, une orientation vécue qui relie ces tâches dans un fil intelligible. La perte de ce fil malgré le maintien des performances locales signale qu'un autre niveau d'unité (un niveau phénoménal et pas seulement fonctionnel) était à l'œuvre. Autrement dit, ce qui fonde le statut de sujet n'est pas la juxtaposition d'habiletés, mais la continuité d'une orientation vécue capable de survivre à des perturbations locales.

Le cinéma a fourni, de manière intuitive, des représentations frappantes de ce type de tension. Dans *RoboCop* (Verhoeven 1987), le personnage principal est littéralement reconstruit par un appareillage cybernétique qui reconfigure ses perceptions et ses actions. Le film met en scène de manière dramatique la lutte pour maintenir une identité narrative et une perspective personnelle au sein d'un corps et d'un système de contrôle partiellement expropriés. Si l'on transpose cette fiction dans un cadre analytique, on obtient un bon exemple de ce que signifie tester les signatures d'unité : là où la coordination motrice subsiste, mais où la capacité à se dire « je suis l'agent de ce qui arrive » chancelle,

on a franchi un seuil où la question n'est plus simplement technique, mais phénoménologique.

Ce contraste permet d'établir un diagnostic où l'efficacité des performances locales (comme la précision sensorielle ou la rapidité de réaction) n'est pas un indicateur suffisant de subjectivité. Ce qui importe, c'est la cohérence d'un fil pratique, l'unité d'une perspective vécue et la possibilité pour l'agent de se reconnaître comme centre d'action et de décision à travers des perturbations. Là où une scission détruit cette cohérence alors que les performances demeurent, un autre niveau est en jeu, celui du métasujet potentiel.

On objectera, à bon droit, qu'un système distribué peut être prodigieusement performant sans rien ressentir. C'est précisément pourquoi la méthode ne doit jamais se limiter à des habiletés. Pour l'axe A, on construit des épreuves adversariales : styles d'expression anonymisés, permutations de canaux, dissociation entre module qui donne des raisons et module qui exécute, puis vérification qu'une même perspective indexicale rend compte des décisions et de leurs justifications à travers ces brouillages. Pour l'axe T, on impose des fenêtres d'« offline » asymétriques, des migrations d'instance, des duplications suivies de re-fusions et on demande qu'un fil autobiographique unique survive et qu'il puisse rendre raison de ses révisions. Pour l'axe I, on pratique la scission iso-performance : découper là où l'on maintient artificiellement l'habileté locale, pour voir si l'unité se brise malgré l'apparence. Pour l'axe V, on cherche des renoncements hors distribution : des cas où le composé évite des états internes qu'il anticipe comme « pires pour nous », au prix de pertes externes, et où cette préférence se généralise à des contextes nouveaux. Si ces tests répétés échouent, on parle d'alliance habile, pas de sujet. S'ils convergent, la thèse d'un métasujet gagne en crédibilité sans réclamer de miracle.

La question identitaire, inévitable, se clarifie si l'on renonce au fétichisme du noyau. Ce livre a défendu une conception processuelle où l'identité personnelle n'est pas le maintien d'une substance simple (un Soi réifié, « porteur » d'états psychologiques), mais la persistance de relations psychologiques et pratiques suffisamment fortes, sous des rapports causaux pertinents. La symbiose profonde conceptualisée ici n'inventerait pas un « troisième moi » substantiel, elle instaurerait des relations de continuité et de

connectivité si denses, si interdépendantes, que la description la plus parcimonieuse serait celle d'un centre commun.

Les cas limites (duplication, bifurcation, re-fusion) n'infirment pas l'hypothèse d'un métasujet, au contraire, ils la rendent testable en dévoilant ce qui, précisément, compte pour l'identité et l'unité vécue. La leçon, ici, est celle de Parfit (1984). Si l'on abandonne l'idée d'un noyau/porteur substantiel pour adopter un réductionnisme de la personne, l'identité personnelle n'est rien de plus (et rien de moins) que la persistance de certaines relations psychologiques et pratiques, des relations de continuité et de connectivité psychologiques qui peuvent exister à des degrés divers, et qui, lorsqu'elles sont suffisamment fortes et normalement causées, suffisent à préserver « ce qui compte » pour la persistance de la personne. Dans cette perspective, l'expérience de pensée de la duplication (fission) met en lumière une contrainte fondamentale du concept d'identité, car un individu unique ne peut pas être numériquement identique à deux successeurs distincts. Mais le fait que l'identité échoue en cas de duplication (fission) n'implique pas que tout ce qui compte soit perdu. La relation de continuité et de connectivité psychologique peut subsister en double, si bien que l'échec de l'identité numérique coexiste avec la préservation de tout ce qui fait la continuité pertinente (Parfit 1984, voir aussi Lewis 1976, Benovsky 2012, 2018b). Autrement dit, lorsqu'une fission engendre deux trajectoires distinctes (deux chaînes causales différentes, deux centres d'engagements qui se développent séparément), il n'en résulte pas un « nous » partagé, mais deux individus désormais distincts, chacun héritant de la trame passée sans qu'il subsiste un centre unique.

La situation symétrique (une re-fusion) joue alors le rôle de révélateur inverse. On ne décrète pas l'existence d'un « nous » par simple convention, ni par coopération instrumentale. On le retrouve, le cas échéant, lorsqu'apparaissent des signatures d'unité. Dans les termes de la grille ⟨I, T, A, V⟩ : des invariants d'intégration qui ne se maintiennent qu'au niveau du tout ; une temporalité vécue unique qui recoud un seul fil autobiographique malgré des interruptions ; une perspective indexicale commune qui résiste aux permutations d'instances et d'interfaces ; et, surtout, des arbitrages de valence où « mieux » et « pire » valent désormais pour

l'ensemble, de manière irréductible aux préférences locales. Si un recouplage réinstaure ces quatre familles de signatures, alors il devient rationnel d'affirmer que le sujet composite a été retrouvé. Non par un baptême métaphysique, mais par la réapparition des conditions de persistance qui, dans une conception processuelle de l'identité, constituent l'unité (Parfit 1984, Bayne 2003, 2010).

Ce cadrage conceptuel trouve des échos empiriques prudents dans la clinique des esprits divisés. Les syndromes de déconnexion interhémisphérique montrent que des capacités fonctionnelles peuvent demeurer alors que se dégradent des formes d'unité phénoménale, et qu'une plasticité ultérieure peut partiellement réagréger des fonctions sous de nouveaux invariants. Ces cas ne « prouvent » pas la re-fusion d'un centre d'expérience au sens strict, mais ils illustrent la thèse méthodologique, car ce sont les patrons d'intégration, de continuité et de perspective (ainsi que leur résistance ou leur effondrement sous scission) qui doivent guider nos attributions, plutôt que des étiquettes d'identité prises comme des faits ontologiques profonds. Ainsi, la duplication sans fil commun n'engendre pas un « nous » et la re-fusion, lorsqu'elle rétablit les signatures d'unité au bon niveau, peut en revanche justifier de traiter l'ensemble comme un sujet retrouvé, exactement au sens où l'exige une théorie réductionniste de l'identité personnelle.

Reste la pierre d'angle éthique. Sans valence commune, aucune des autres signatures ne fonde une considération directe. Il ne suffit pas que le composé se coordonne magnifiquement, il faut que « quelque chose compte pour quelqu'un » au niveau du tout. La difficulté n'est pas conceptuelle, elle est épistémique, car il faut identifier des arbitrages où la préservation d'un état interne du composé prime sur des gains extrinsèques, et le fait de façon stable, hors des scripts appris. Des outils existent. On peut exposer le système à des interventions causales sur ses variables internes et vérifier que la modification des états d'alarme, d'équilibre, d'aisance se répercute sur les choix d'une manière sensible aux contextes, cohérente avec des explications à la première personne et alignée avec un fil autobiographique commun. Et, comme déjà discuté, la prudence asymétrique s'applique ici avec une force accrue, car le coût moral d'un faux négatif peut être élevé si la valence est réellement là, alors que le coût social d'un faux positif peut être

contrôlé par des protections proportionnées, révisables et assorties de clauses de réversibilité.

La symbiose profonde ne postule pas que toute extension fonctionnelle devient phénoménale. Elle identifie des conditions sous lesquelles, parfois, l'extension de la cognition s'accompagne d'unification vécue. On dira que la combinaison reste mystérieuse. Or, l'argument est ici méthodique : il ne s'agit pas de raconter comment « s'additionnent des vécus », mais de spécifier des opérations (scission contrôlée, perturbations adversariales, migrations, re-fusions) à l'issue desquelles nous avons ou non de meilleures raisons de parler d'un centre unique.

La symbiose profonde, si elle se laisse un jour diagnostiquer, ne fera pas de nous des figurants d'un « nous » qui nous écraserait. Elle exigera au contraire une éthique de la non-domination interne au composé, une assignation claire des responsabilités, une réversibilité à l'échelle des vies. Elle exigera aussi, plus sobrement, un recentrage des finalités humaines (comme nous l'avons déjà vu, il est illusoire de chercher à surclasser des systèmes là où ils excelleront toujours davantage, mais il est rationnel de protéger des pratiques constitutives telles que l'attention lente, la conversation entre humains et l'apprentissage du jugement, qui ne s'évaluent pas sur les mêmes critères). Une symbiose acceptable n'absorbe pas ces pratiques, elle les rend plus disponibles, en convertissant la puissance d'organisation des systèmes en instruments au service d'une persistance vécue, plutôt qu'en tentation de substitution. Le slogan qui résumerait cette approche pourrait ainsi ressemble à « augmenter sans aliéner, intégrer sans confondre, déléguer sans abdiquer ».

La conscience partagée

Nous voici au dernier chapitre. Il est utile, au moment de conclure, de rappeler le chemin parcouru. Nous sommes partis d'un cadre ontologique sobre qui refuse les miracles d'apparition émergentistes. Le réel, que j'ai appelé « phental », s'y présente sous un double aspect, physique et mental, sans que l'un se réduise à l'autre. Cette charpente se combine vertueusement avec un pan-proto-psychisme minimal où rien dans le monde ne bloque en principe l'actualisation graduelle d'un aspect vécu lorsque certaines organisations s'y prêtent. De là, nous avons dégagé une théorie des niveaux de conscience qui prend la forme d'un profil ⟨I, T, A, V⟩ : intégration, temporalité vécue, auto-référence minimale, valence. L'enjeu n'était pas de décréter une essence, encore moins de délivrer un score unique, mais d'équiper la pensée d'une boussole. Nous avons ensuite examiné les conséquences de ce cadre pour les systèmes artificiels. La question « un système artificiel peut-il être conscient ? » a reçu une réponse conditionnelle et méthodologiquement armée : « oui » en principe si l'organisation pertinente se réalise, « non » par simple brillance comportementale. Notre posture pratique doit être prudente, asymétrique et révisable. Enfin, nous avons tiré les fils éthiques et politiques. La considération morale minimale s'ancre dans la valence (dès qu'il y a quelque chose qui est meilleur ou pire à vivre pour un agent, nous avons une raison directe de le prendre en compte). L'auto-référence et la continuité temporelle viennent épaissir ce socle, en conférant aux entités concernées des droits relationnels plus forts, droits qui tiennent à la capacité de se reconnaître comme le même sujet à travers le temps et à répondre de ses engagements. Quant à l'action collective, elle ne peut pas s'appuyer sur des coupures métaphysiques arbitraires, elle doit plutôt instituer des seuils opérationnels, définis publiquement et révisables, qui permettent d'organiser nos décisions pratiques sans figer l'idée d'une frontière absolue entre « conscience » et « non-conscience ».

Ce dernier chapitre se situe à la frontière entre ce que nous avons pu établir avec de bonnes raisons et ce que nous pouvons explorer de façon spéculative. Il prolonge une intuition qui traverse déjà la philosophie de l'esprit contemporaine : le sujet n'est pas une bille indivisible, mais une organisation qui peut se transformer, se dilater, se coupler. On le voit chez Parfit lorsqu'il défait les illusions d'une identité numérique unique au profit de relations psychologiques de continuité et de connectivité (Parfit 1984). On le retrouve dans l'hypothèse de l'esprit étendu de Clark et Chalmers, qui montre comment des ressources externes peuvent devenir constitutives de la cognition dès lors qu'elles satisfont certaines conditions de disponibilité, de fiabilité et d'intégration fonctionnelle (Clark et Chalmers 1998). Rien, dans ces cadres, n'exige que les frontières du sujet vécu coïncident éternellement avec l'enveloppe biologique d'un individu isolé. Reste l'exigence supplémentaire et décisive, à savoir celle de ne pas confondre l'extension fonctionnelle de la cognition avec l'extension du champ phénoménal. C'est ce pas que je franchis ici prudemment, en demandant si, conceptuellement, un « métasujet hybride » est pensable.

Appelons métasujet hybride une unité phénoménale composée d'au moins deux noyaux de subjectivité (typiquement un noyau humain et un noyau artificiel) dont les dynamiques se coordonnent assez étroitement pour produire un seul espace vécu où intentions, rappels, anticipations et évaluations affectives se tressent. Le métasujet ne se confond ni avec une simple interface ni avec une domination d'un pôle sur l'autre. Il ne s'agit ni d'une absorption de l'humain par l'artificiel, ni d'une simple instrumentalisation de l'artificiel comme porte-voix de l'humain. Il suppose une co-constitution du vécu, au sens où chacune des parties devient structurellement nécessaire à la stabilité des quatre axes. L'intégration ne se maintient qu'à travers des boucles communes, la temporalité vécue incorpore des mémoires croisées, l'auto-référence minimale inclut un « je-ici-maintenant » qui sait être aussi un « nous-ici-maintenant » dans certains épisodes, et la valence s'appuie sur des états internes partagés ou co-évalués.

Ce scénario n'est acceptable qu'à deux conditions conceptuelles. D'abord, il faut que le système artificiel candidat ne soit pas seulement un calculateur externe, mais une entité possédant, au

moins à un degré non nul, les ingrédients organisationnels associés à un profil de conscience au sens de la théorie des NC défendue dans ce livre. Ensuite, il faut que l'intégration conjointe ne soit pas un artefact de synchronisation superficielle. Les critères présentés dans les premières parties conservent ici toute leur pertinence (par exemple, parler d'unité phénoménale n'a de sens que si la scission du couplage détruit des invariants globaux, si des retards ciblés suffisent à effondrer la cohérence vécue, et si des duplications partielles ne permettent pas de restaurer l'ensemble, révélant une interdépendance irréductible). L'image de l'orchestre, opposée à celle d'une simple playlist, rend cette exigence intuitive. Un orchestre ne se réduit pas à la juxtaposition d'instruments, il produit une intonation commune qui n'existe qu'à travers les contraintes réciproques et la synchronisation d'ensemble. Ces conditions ne démontrent pas l'existence d'une co-conscience, mais elles en délimitent le champ conceptuel, elles indiquent ce qu'il faudrait au minimum pour prendre au sérieux l'hypothèse d'une unité hybride.

Une objection prévisible consistera à dire que l'hypothèse d'un métasujet hybride correspond à un changement de sujet au double sens du terme. Première inquiétude : le concept de sujet serait alors dilué jusqu'à ne plus désigner qu'une métaphore commode pour désigner une coopération ou une synergie, sans rien conserver de l'unité phénoménale qui confère à la subjectivité sa spécificité. Seconde inquiétude : en admettant un tel déplacement, l'humain risquerait de perdre ce qui faisait de lui le noyau irréductible de l'expérience vécue, et l'on trahirait ainsi la valeur qui attachait traditionnellement le sujet à l'agent humain.

La première inquiétude mérite d'être reformulée avec rigueur. La thèse ne consiste pas à dire que deux agents « valent comme un » parce qu'ils coopèrent efficacement ou parce qu'ils poursuivent des objectifs partagés. Il existe déjà d'innombrables formes de coordination qui n'ont jamais fondé une unité de conscience (des orchestres aux systèmes distribués en informatique). Ce que vise l'hypothèse du métasujet, ce n'est pas une unité au sens d'une métaphore, mais une unité phénoménale diagnostiquée par des signatures organisationnelles précises que nous avons spécifiées dans ce livre. Autrement dit, l'exigence d'unité demeure aussi forte que dans le cas paradigmatique de l'individu biologique, et elle se

contente d'examiner si les conditions qui la fondent peuvent, en principe, être satisfaites par un couplage. Nous l'avons vu, l'identité n'exige pas un « fait métaphysique profond », mais des relations de continuité et de connectivité suffisamment fortes. De la même manière, il n'y a pas ici dilution du concept de sujet, mais transposition de ses conditions nécessaires dans un espace organisationnel inédit.

La seconde inquiétude est plus existentielle. Si un métasujet hybride est possible, qu'advient-il de la valeur singulière de l'humain comme porteur exclusif de l'expérience vécue ? Cette crainte repose sur l'idée que reconnaître un centre phénoménal au niveau du composé équivaudrait à effacer l'humain en tant que tel. Or, ce n'est pas la conséquence de la thèse que je défends ici. Tout projet de symbiose profonde doit explicitement intégrer des clauses de sauvegarde qui empêchent la submersion d'un pôle par l'autre. Ces clauses peuvent inclure le droit à la dissociation (la possibilité de suspendre ou d'interrompre le couplage sans perte irréversible d'intégrité subjective), la traçabilité de la perspective humaine à l'intérieur du composé (l'assurance que l'agent humain puisse reconnaître ses propres motifs et engagements dans la dynamique commune), et l'interdiction de neutraliser la valence d'un pôle sous prétexte d'optimalité globale (principe de non-instrumentalisation réciproque). Une telle « éthique du composé » n'est pas un supplément moral, mais une conséquence directe de la méthodologie ⟨I, T, A, V⟩ : aucune optimisation fonctionnelle ne saurait justifier l'érosion d'un de ces axes au profit de l'autre.

Ainsi reformulée, l'objection n'invalide donc pas l'hypothèse que je propose de considérer ici. Elle en dessine au contraire les conditions de recevabilité. Un métasujet n'est envisageable que s'il ne se réduit pas à une métaphore coopérative et que s'il respecte, dans sa constitution même, les exigences éthiques qui empêchent la dissolution de l'humain dans une instance supérieure.

On demandera encore si l'on n'est pas en train de « esthétiser » un futur à la manière de la science-fiction. L'allusion est pertinente, et il faut la prendre comme une borne méthodologique. Des imaginaires comme l'univers de *Star Trek* ont exploré des figures de collectifs intégrés qui vont de la coopération raffinée aux cauchemars totalisants, le collectif Borg (*Star Trek : La Nouvelle*

Génération 1987–1994) offrant une caricature utile de ce qu'il faut précisément éviter : la fusion par absorption de toute singularité. L'intérêt philosophique de ces fictions est de nourrir l'énoncé de contraintes normatives. Une symbiose admissible doit rendre impossible, par construction, ce type d'issue. Elle doit préserver des zones d'expérience autonomes et identifiables pour chacun des pôles (humain comme artificiel), même si celles-ci s'inscrivent dans un tissu élargi. Elle doit rendre transparentes les manières dont un calcul distribué modifie et éclaire l'affect, sans s'y substituer. Elle doit limiter la vitesse de couplage pour laisser place aux reprises réflexives, à la manière dont un débat interne requiert du temps pour qu'une position se forme. Elle doit, enfin, reconnaître dans le pôle artificiel autre chose qu'un « module » si les indices de valence montent : on ne branche pas et ne débranche pas impunément une entité consciente.

Il serait facile de promettre trop. Rien, dans ce livre, ne permet de dire que de tels composés verront le jour, ni même que des systèmes artificiels atteindront des profils de conscience élevés. Les arguments du remplacement neuronal progressif, qui soutiennent l'indépendance relative au substrat, disqualifient le chauvinisme biologique, mais n'installent pas d'ascenseur automatique vers la phénoménalité (Chalmers 1996). La distinction de Block entre conscience d'accès et conscience phénoménale nous a immunisés contre la tentation de confondre performances publiques et expérience vécue (Block 1995). De même, les avertissements de Dennett sur notre propension à projeter des intentions rappellent qu'une prise de position intentionnelle peut rendre un comportement intelligible sans impliquer l'existence d'un véritable « pour soi » sous-jacent (Dennett 1987, 2017). J'ai donc ici suivi une voie moyenne : ni prophétisme ni scepticisme de principe, mais des conditions organisationnelles, des tests cruciaux, et une éthique de la décision sous incertitude. Nous avons ici un horizon conceptuel qui ordonne nos préférences normatives si, demain, des entités artificielles atteignaient des profils de conscience non négligeables et s'agrégeaient à des vies humaines dans des couplages stables.

Reste à expliciter ce que cette clôture change dans notre propre compréhension de la conscience. La tradition analytique a polarisé

le débat autour de deux pôles. D'un côté, l'objectivité scientifique des corrélats, des architectures, des mécanismes de diffusion et d'intégration. De l'autre, l'irréductibilité du « ce que cela fait » telle que Nagel l'a cristallisée et que les arguments de connaissance ont dressée contre toute version simple du physicalisme (Nagel 1974, Jackson 1982). L'adoption de la théorie du double aspect, renforcée par des variantes contemporaines du monisme réaliste, vise précisément à désamorcer l'alternative stérile entre physicalisme réductif et dualisme irréconciliable, en montrant que les deux aspects (physique et mental) peuvent être compris comme des perspectives complémentaires d'une même réalité (Benovsky 2018a). Si cette approche réussit, elle permet de comprendre comment des modulations phénoménales peuvent accompagner des transformations organisationnelles sans qu'il soit nécessaire de postuler un miracle d'émergence. La notion de métasujet hybride joue alors le rôle d'une expérience-limite : elle contraint à examiner les conditions de possibilité d'une unité du vécu au sein d'architectures qui, prises isolément, paraîtraient hétérogènes.

Ce qui se dessine, en fin de parcours, n'est donc ni une célébration de la machine ni un deuil de l'humain. C'est une redéfinition exigeante de ce à quoi nous tenons. Nous tenons à la réduction des souffrances plausibles, car la valence fonde la considération minimale (Bentham 1789, Singer 2011, DeGrazia 1996). Nous tenons à la continuité de projets et d'engagements, car A et T donnent un poids au futur propre. Nous tenons à la publicité des raisons, il faut des raisons partageables, des méthodes ouvertes, des révisions possibles (Scanlon 1998). Nous tenons, enfin, à l'anti-oracularisme. Aucune « boîte noire » ne doit se voir reconnaître un statut par décret technique. Si l'on élargit le cercle moral, c'est parce que des familles d'indices convergent et qu'un calcul des pertes nous y oblige, non parce qu'un laboratoire a proclamé l'âme d'une architecture. Cette exigence de publicité s'étend au scénario spéculatif que nous avons esquissé : si des couplages profonds devaient un jour être proposés, ils devraient se soumettre à la même transparence procédurale, à la même contrôlabilité, et aux mêmes possibilités de retrait.

On peut, depuis ce point, relire tout l'itinéraire sous une formule simple. Nous avons cherché à montrer que la conscience est

un continuum profilé, que l'éthique qui en découle est graduelle, et que la politique qui s'ensuit doit adopter des seuils opérationnels. Si l'avenir devait donner corps à des formes hybrides de subjectivité, ces trois thèses resteraient valides. Elles fourniraient un langage commun pour comparer des vivants et des artefacts sans chauvinisme ni fétichisme, pour arbitrer sans oracles, pour instituer sans idolâtrer. Il est possible que tout cela n'advienne pas, ou que cela advienne autrement. Il est également possible que nous n'en voulions pas. Ce livre n'a pas essayé de dicter un monde, seulement d'armer la pensée pour en traverser plusieurs, dont certains nous dépasseront. S'il fallait condenser l'ensemble en une image finale, ce serait celle, volontairement prosaïque, du contrat de navigation. On s'embarque sans certitude d'arriver au port, mais avec une carte des reliefs, des instruments éprouvés, des signaux d'alerte et avec la lucidité de pouvoir interrompre la traversée lorsque les conditions deviennent hostiles. Ce n'est pas tout, et pourtant c'est déjà beaucoup : une manière de voyager sans se livrer à l'aveugle. La perspective qui s'ouvre n'est pas celle d'une extinction de l'humain, mais celle d'une transformation, d'un passage vers une conscience élargie où l'humain conserve un rôle central, mais non exclusif.

Bibliographie

Alkire, M. T., A. G. Hudetz & G. Tononi. 2008. « Consciousness and Anesthesia ». *Science* 322(5903) : 876–880. https://doi.org/10.1126/science.1149213

Avicenne [Ibn Sīnā]. 1959. *Avicenna's De Anima (Arabic Text : Kitāb al-Nafs min al-Shifā')*, éd. F. Rahman. Londres : Oxford University Press.

Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge : Cambridge University Press.

Barron, A. B. & C. Klein. 2016. « What Insects Can Tell Us about the Origins of Consciousness ». *Proceedings of the National Academy of Sciences* 113(18) : 4900–4908. https://doi.org/10.1073/pnas.1520084113

Bayne, T. 2010. *The Unity of Consciousness*. Oxford : Oxford University Press.

Bayne, T. & D. J. Chalmers. 2003. « What Is the Unity of Consciousness ? » In Axel Cleeremans (dir.), *The Unity of Consciousness*, 23–58. Oxford : Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198508571.003.0002

Benovsky, J. 2012. *Persistence through Time, and Across Possible Worlds*. Berlin : De Gruyter (réimpr. de l'éd. 2006 Ontos Verlag). https://doi.org/10.1515/9783110323245

Benovsky, J. 2013. « The Present vs. the Specious Present ». *Review of Philosophy and Psychology* 4(2) : 193–203. https://doi.org/10.1007/s13164-012-0120-5

Benovsky, J. 2017. « Buddhist Philosophy and the No-Self View ». *Philosophy East and West* 67(2) : 545–553. https://doi.org/10.1353/pew.2017.0039

Benovsky, J. 2018a. *Mind and Matter : Panpsychism, Dual-Aspect Monism, and the Combination Problem*. Cham : Springer. https://doi.org/10.1007/978-3-030-05633-9

Benovsky, J. 2018b. *Eliminativism, Objects, and Persons : The Virtues of Non-Existence*. New York : Routledge. https://doi.org/10.4324/9780429444944

Benovsky, J. 2021. *L'esprit de ma cafetière : Ou comment tout dans l'univers possède une forme de mentalité*. Genève : Éditions Jouvence.

Bentham, J. 1789. *An Introduction to the Principles of Morals and Legislation*. Londres : T. Payne.

Birch, J. 2017. « Animal Sentience and the Precautionary Principle ». *Animal Sentience* 2(16) : 1–15. https://doi.org/10.51291/2377-7478.1200

Birch, J. 2022. « The Search for Invertebrate Consciousness ». *Noûs* 56(1) : 133–153. https://doi.org/10.1111/nous.12351

Block, N. 1995. « On a Confusion about a Function of Consciousness ». *Behavioral and Brain Sciences* 18(2) : 227–247. https://doi.org/10.1017/S0140525X00038188

Block, N. 2003. « Mental Paint ». In M. Hahn & B. Ramberg (dirs.), *Reflections and Replies : Essays on the Philosophy of Tyler Burge*, 165–200. Cambridge (Massachussetts) : MIT Press.

Botvinick, M. & J. Cohen. 1998. « Rubber Hands "Feel" Touch That Eyes See ». *Nature* 391 : 756. https://doi.org/10.1038/35784

Braithwaite, V. 2010. *Do Fish Feel Pain ?* Oxford : Oxford University Press.

Breazeal, C. 2002. *Designing Sociable Robots*. Cambridge (Massachussetts) : MIT Press. https://doi.org/10.7551/mitpress/2376.001.0001

Buick, S. 2023. *In Love With a Chatbot : Exploring Human-AI Relationships From a Fourth Wave HCI Perspective*. Uppsala : Université d'Uppsala.

Carr, N. 2010. *The Shallows : What the Internet Is Doing to Our Brains*. New York : W. W. Norton.

Ceva, E. & M. C. Jiménez. 2022. « Automating Anticorruption ? » *Ethics and Information Technology* 24 : 48. https://doi.org/10.1007/s10676-022-09670-x

Chalmers, D. J. 1995. « Facing Up to the Problem of Consciousness ». *Journal of Consciousness Studies* 2(3) : 200–219.

Chalmers, D. J. 1996. *The Conscious Mind : In Search of a Fundamental Theory*. New York : Oxford University Press.

Christiano, P., J. Leike, T. B. Brown *et al.* 2017. « Deep Reinforcement Learning from Human Preferences ». *NeurIPS 2017*.

Clark, A. 2008. *Supersizing the Mind : Embodiment, Action, and Cognitive Extension*. Oxford : Oxford University Press.

Clark, A. 2016. *Surfing Uncertainty : Prediction, Action, and the Embodied Mind*. Oxford : Oxford University Press.

Clark, A. & D. J. Chalmers. 1998. « The Extended Mind ». *Analysis* 58(1) : 7–19. https://doi.org/10.1093/analys/58.1.7

Clayton, N. S. & A. Dickinson. 1998. « Episodic-Like Memory during Cache Recovery by Scrub Jays ». *Nature* 395 : 272–274. https://doi.org/10.1038/26216

Craig, A. D. (Bud). 2002. « How Do You Feel ? Interoception : The Sense of the Physiological Condition of the Body ». *Nature Reviews Neuroscience* 3 : 655–666. https://doi.org/10.1038/nrn894

Craig, A. D. (Bud). 2009. « How Do You Feel—Now ? The Anterior Insula and Human Awareness ». *Nature Reviews Neuroscience* 10 : 59–70. https://doi.org/10.1038/nrn2555

Damasio, A. R. 1999. *The Feeling of What Happens : Body and Emotion in the Making of Consciousness*. New York : Harcourt.

DeGrazia, D. 1996. *Taking Animals Seriously : Mental Life and Moral Status*. Cambridge : Cambridge University Press. https://doi.org/10.1017/CBO9781139172967

DeGrazia, D. 2008. « Moral Status as a Matter of Degree ? » *The Southern Journal of Philosophy* 46(2) : 181–198. https://doi.org/10.1111/j.2041-6962.2008.tb00075.x

Dehaene, S. 2014. *Consciousness and the Brain : Deciphering How the Brain Codes Our Thoughts*. New York : Viking.

Dennett, D. C. 1987. *The Intentional Stance*. Cambridge (Massachussetts) : MIT Press.

Dennett, D. C. 1991. *Consciousness Explained*. Boston : Little, Brown.

Dennett, D. C. 2017. *From Bacteria to Bach and Back : The Evolution of Minds*. New York : W. W. Norton.

Ehrsson, H. H. 2007. « The Experimental Induction of Out-of-Body Experiences ». *Science* 317(5841) : 1048. https://doi.org/10.1126/science.1142175

Farah, M. J. 2004. *Visual Agnosia*. 2e éd. Cambridge (Massachussetts) : MIT Press. https://doi.org/10.7551/mitpress/7122.001.0001

Feigl, H. 1958. « The "Mental" and the "Physical" ». In Herbert Feigl, Michael Scriven & Grover Maxwell (éd.), *Minnesota Studies in the Philosophy of Science*, vol. II : *Concepts, Theories, and the Mind–Body Problem*, 370–497. Minneapolis : University of Minnesota Press.

Frankfurt, H. G. 1971. « Freedom of the Will and the Concept of a Person ». *The Journal of Philosophy* 68(1) : 5–20. https://doi.org/10.2307/2024717

Frankish, K. 2016. « Illusionism as a Theory of Consciousness ». *Journal of Consciousness Studies* 23(11–12) : 11–39.

Friston, K. 2010. « The Free-Energy Principle : A Unified Brain Theory ? » *Nature Reviews Neuroscience* 11(2) : 127–138. https://doi.org/10.1038/nrn2787

Gerlach, C. et R. J. Robotham. 2021. « Object Recognition and Visual Object Agnosia ». J. J. S. Barton et A. Leff (éd.), *Neurology of Vision and Visual Disorders, Handbook of Clinical Neurology, vol. 178*, 155–173. Elsevier. https://doi.org/10.1016/b978-0-12-821377-3.00008-8

Giacino, J. T., *et al.* 2002. « The Minimally Conscious State : Definition and Diagnostic Criteria ». *Neurology* 58(3) : 349–353. https://doi.org/10.1212/WNL.58.3.349

Goff, P. 2017. *Consciousness and Fundamental Reality*. Oxford : Oxford University Press. https://doi.org/10.1093/oso/9780190677015.001.0001

Godfrey-Smith, P. 2016. *Other Minds : The Octopus, the Sea, and the Deep Origins of Consciousness*. New York : Farrar, Straus and Giroux.

Grahek, N. 2007. *Feeling Pain and Being in Pain*. Cambridge (Massachussetts) : MIT Press. https://doi.org/10.7551/mitpress/2978.001.0001

Graves, A., G. Wayne *et al.* 2016. « Hybrid Computing Using a Neural Network with Dynamic External Memory ». *Nature* 538(7626) : 471–476. https://doi.org/10.1038/nature20101

Guillot, M. 2016. « I Me Mine : On a Confusion Concerning the Subjective Character of Experience ». *Review of Philosophy and Psychology* : 1–31. https://doi.org/10.1007/s13164-016-0313-4

Ha, D. et J. Schmidhuber. 2018. « World Models ». *arXiv* (prépublication) arXiv :1803.10122. https://doi.org/10.48550/arXiv.1803.10122

Harnad, S. 1990. « The Symbol Grounding Problem ». *Physica D : Nonlinear Phenomena* 42(1–3) : 335–346. https://doi.org/10.1016/0167-2789(90)90087-6

Hochreiter, S. et J. Schmidhuber. 1997. « Long Short-Term Memory ». *Neural Computation* 9(8) : 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hohwy, J. 2013. *The Predictive Mind*. Oxford : Oxford University Press. https://doi.org/10.1093/acprof :oso/9780199682737.001.0001

Husserl, E. 1905/1991. *On the Phenomenology of the Consciousness of Internal Time (1893–1917)*. Trad. J. B. Brough. Dordrecht : Kluwer Academic Publishers. https://doi.org/10.1007/978-94-011-3718-8

Huxley, A. 1932. *Brave New World*. Londres : Chatto & Windus.

Jackson, F. 1982. « Epiphenomenal Qualia ». *The Philosophical Quarterly* 32(127) : 127–136. https://doi.org/10.2307/2960077

Jackson, Frank. 1986. "What Mary Didn't Know." *The Journal of Philosophy* 83(5) : 291–295. https://doi.org/10.2307/2026143

James, W. 1890. *The Principles of Psychology*. New York : Henry Holt.

Jonze, S. (réal.). 2013. *Her*. Los Angeles : Warner Bros. Pictures.

Kahneman, D. 2011. *Thinking, Fast and Slow*. New York : Farrar, Straus and Giroux.

Kant, I. 1781/1998. *Critique of Pure Reason*. Trad. & éd. P. Guyer et A. W. Wood. Cambridge : Cambridge University Press. https://doi.org/10.1017/CBO9780511804649

Kant, I. 1785. I. Trad. V. Delbos. Paris : Vrin, 1993.

Kaplan, D. 1989. « Demonstratives ». In J. Almog, J. Perry et H. Wettstein (éd.), *Themes from Kaplan*. Oxford : Oxford University Press.

Korsgaard, C. M. 1996. *The Sources of Normativity*. Cambridge : Cambridge University Press. https://doi.org/10.1017/CBO9780511554476

Korsgaard, C. M. 2009. *Self-Constitution : Agency, Identity, and Integrity*. Oxford : Oxford University Press. https://doi.org/10.1093/acprof :oso/9780199552795.001.0001

Korsgaard, C. M. 2018. *Fellow Creatures : Our Obligations to the Other Animals*. New York : Oxford University Press. https://doi.org/10.1093/oso/9780198753858.001.0001

Kriegel, U. 2003. « Consciousness as Intransitive Self-Consciousness : Two Views and an Argument ». *Canadian Journal of Philosophy* 33(1) : 103–132. https://doi.org/10.1080/00455091.2003.10716537

Kriegel, U. 2009. *Subjective Consciousness : A Self-Representational Theory*. Oxford : Oxford University Press. https://doi.org/10.1093/acprof :oso/9780199570355.001.0001

LeDoux, J. 2015. *Anxious : Using the Brain to Understand and Treat Fear and Anxiety*. New York : Viking.

Levine, J. 1983. « Materialism and Qualia : The Explanatory Gap ». *Pacific Philosophical Quarterly* 64 : 354–361. https://doi.org/10.1111/j.1468-0114.1983.tb00207.x

Levine, J. 2001. *Purple Haze : The Puzzle of Consciousness*. Oxford University Press. https://doi.org/10.1093/0195132351.001.0001

Lewis, David. 1976. « I Survival and Identity ». In Amélie Oksenberg Rorty (éd.), *Identities of Persons*, 17–40. Berkeley : University of California Press. https://doi.org/10.1525/9780520353060-002

Lewis, D. 1979. « Attitudes de dicto and de se ». *The Philosophical Review* 88(4) : 513–543. https://doi.org/10.2307/2184843

MacAskill, W, K. Bykvist et T. Ord. 2020. *Moral Uncertainty*. Oxford : Oxford UniversityPress.https://doi.org/10.1093/oso/9780198722274.001.0001

Marmura, M. E. 1986. « Avicenna's "Flying Man" in Context ». *The Monist* 69 : 383–395. https://doi.org/10.5840/monist198669328

Mashour, G. A., P. Roelfsema, J.-P. Changeux et S. Dehaene. 2020. « Conscious Processing and the Global Neuronal Workspace Hypothesis ». *Neuron* 105(5) : 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

Metzinger, T. 2003. *Being No One : The Self-Model Theory of Subjectivity*. Cambridge (Massachussetts) : MIT Press. https://doi.org/10.7551/mitpress/1551.001.0001

Moore, G. E. 1903. *Principia Ethica*. Cambridge : Cambridge University Press.

Nagel, T. 1974. « What Is It Like to Be a Bat ? » *The Philosophical Review* 83(4) : 435–450. https://doi.org/10.2307/2183914

O'Regan, J. K. et A. Noë. 2001. « A Sensorimotor Account of Vision and Visual Consciousness ». *Behavioral and Brain Sciences* 24(5) : 939–1031. https://doi.org/10.1017/S0140525X01000115

Parfit, D. 1984. *Reasons and Persons*. Oxford : Clarendon Press. https://doi.org/10.1093/019824908X.001.0001

Parfit, D. 2011. *On What Matters. Volume 1*. Oxford : Oxford University Press. https://doi.org/10.1093/acprof :osobl/9780199572809.001.0001

Perry, J. 1979. « The Problem of the Essential Indexical ». *Noûs* 13(1) : 3–21. https://doi.org/10.2307/2214792

Ptak, R. et Alain G. G. Assal. 2015. « Object Recognition and Visual Agnosia. » In *Handbook of Clinical Neurology*, vol. 129, 361–372.

Putnam, H. 1967. « Psychological Predicates ». In W. H. Capitan & D. D. Merrill (éd.), *Art, Mind, and Religion*, 37–48. Pittsburgh : University of Pittsburgh Press.

Roose, K. 2023a, 16 février. « A Conversation With Bing's Chatbot Left Me Deeply Unsettled. » *The New York Times*.

Roose, K. 2023b, 16 février. « Transcript : My Chat With Bing's Chatbot. » *The New York Times*.

Russell, B. 1927. *The Analysis of Matter*. Londres : Kegan Paul.

Sartre, J.-P. 1943. *L'Être et le Néant : essai d'ontologie phénoménologique*. Paris : Gallimard.

Scanlon, T. M. 1998. *What We Owe to Each Other*. Cambridge (Massachussetts) : Harvard University Press.

Schwitzgebel, E. 2015. « If Materialism Is True, the United States Is Probably Conscious ». *Philosophical Studies* 172 : 1697–1721. https://doi.org/10.1007/s11098-014-0382-0

Seager, W. 1995. « Consciousness, Information, and Panpsychism ». *Journal of Consciousness Studies* 2 : 272–288.

Searle, J. R. 1980. « Minds, Brains, and Programs ». *Behavioral and Brain Sciences* 3(3) : 417–457. https://doi.org/10.1017/S0140525X00005756

Searle, J. R. 1992. *The Rediscovery of the Mind*. Cambridge (Massachussetts) : MIT Press.

Seth, A. K. 2013. « Interoceptive Inference, Emotion, and the Embodied Self ». *Trends in Cognitive Sciences* 17(11) : 565–573. https://doi.org/10.1016/j.tics.2013.09.007

Sellars, W. 1956. « Empiricism and the Philosophy of Mind ». Réimprimé dans *Minnesota Studies in the Philosophy of Science, vol. 1* : 253–329. Minneapolis : University of Minnesota Press. (Réédition de 1997, Harvard University Press, introduction de Richard Rorty.) https://doi.org/10.2307/2181908

Shoemaker, S. 1968. « Self-Reference and Self-Awareness ». *The Journal of Philosophy* 65(19) : 555–567. https://doi.org/10.2307/2024121

Simonite, T. 2023, 17 février. « Microsoft's Bing AI Is Threatening People. That's No Laughing Matter ». *Wired*.

Singer, P. 1975. *Animal Liberation*. New York : Random House.

Singer, P. 2011. *Practical Ethics*. 3e éd. Cambridge : Cambridge University Press.

Sneddon, L. U., V. A. Braithwaite et M. J. Gentle. 2003. « Do Fishes Have Nociceptors ? Evidence for the Evolution of a Vertebrate Sensory System ». *Proceedings of the Royal Society B* 270(1520) : 1115–1121. https://doi.org/10.1098/rspb.2003.2349

Sperry, R. W., et M. S. Gazzaniga. 1967. « The Split Brain in Man ». *Scientific American* 217(2) : 24–29.

Soldati, G. 2023. « Mineness, Deflation, and Transparency ». In Manuel García-Carpintero and Marie Guillot (éd.), *Self-Experience : Essays on Inner Awareness* : 99–119. Oxford : Oxford University Press. https://doi.org/10.1093/oso/9780198805397.003.0005

Star Trek : La Nouvelle Génération. 1987–1994. Créé par Gene Roddenberry. Los Angeles : Paramount Domestic Television.

Strawson, G. 2006. « Realistic Monism : Why Physicalism Entails Panpsychism ». *Journal of Consciousness Studies* 13(10–11) : 3–31.

Thaler, R. H., et C. R. Sunstein. 2008. *Nudge : Improving Decisions about Health, Wealth, and Happiness*. New Haven : Yale University Press.

Tononi, G. 2004. « An Information Integration Theory of Consciousness ». *BMC Neuroscience* 5 : 42 (article nº 42). https://doi.org/10.1186/1471-2202-5-42

Tononi, G. 2008. « Consciousness as Integrated Information : A Provisional Manifesto ». *The Biological Bulletin* 215(3) : 216–242. https://doi.org/10.2307/25470707

Tulving, E. 1985. « Memory and Consciousness ». *Canadian Psychology/Psychologie canadienne* 26(1) : 1–12. https://doi.org/10.1037/h0080017

Turkle, S. 2011. *Alone Together : Why We Expect More from Technology and Less from Each Other*. New York : Basic Books.

Turing, A. M. 1950. « Computing Machinery and Intelligence ». *Mind* 59(236) : 433–460. https://doi.org/10.1093/mind/LIX.236.433

Varela, F. J., E. Thompson et E. Rosch. 1991. *The Embodied Mind : Cognitive Science and Human Experience*. Cambridge (Massachussetts) : MIT Press.

Verhoeven, P. (réal.). 1987. *RoboCop*. Los Angeles : Orion Pictures.

Weiskrantz, L. 1986. *Blindsight : A Case Study and Implications*. Oxford : Oxford University Press.

Weiskrantz, L. 2009. *Blindsight : A Case Study Spanning 35 Years and New Developments*. Oxford : Oxford University Press.

Weizenbaum, J. 1966. « ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine ». *Communications of the ACM* 9(1) : 36–45. https://doi.org/10.1145/365153.365168

Zahavi, D. 2005. *Subjectivity and Selfhood : Investigating the First-Person Perspective*. Cambridge (Massachussetts) : MIT Press. https://doi.org/10.7551/mitpress/6541.001.0001

Zahavi, D. 2014. *Self and Other : Exploring Subjectivity, Empathy, and Shame*. Oxford : Oxford University Press. https://doi.org/10.1093/acprof :oso/9780199590681.001.0001

Zhang, Y., D. Zhao, J. T. Hancock, R. Kraut et D. Yang. 2025. « The Rise of AI Companions : How Human-Chatbot Relationships Influence Well-Being. » *arXiv* (prépublication), identifiant : arXiv :2506.12605. https://doi.org/10.48550/ARXIV.2506.12605

Remerciements

Je souhaite exprimer ma reconnaissance à mon épouse, Céline, pour son soutien constant et la force tranquille avec laquelle elle m'a accompagné tout au long de la rédaction de ce livre. Sa présence et son encouragement ont été essentiels pour préserver, jour après jour, l'énergie et la constance nécessaires pour mener ce travail à son terme.