Le test de Turing pour évaluer les théories de l'esprit

Robin Lamarche-Perrin

Projet de Mémoire

présenté par

Robin Lamarche-Perrin

Master 1 Philosophie *sp.* Philosophie et Langages
Option *Discours, Savoirs, Représentations*

Université Pierre-Mendès-France
UFR SH Grenoble

Le test de Turing pour évaluer les théories de l'esprit

Soutenu le 16 septembre 2010
devant les membres du jury

Encadrant	Maximilian Kistler
Invité	Denis Perrin
Responsable du Master	Marie-Laurence Desclos
Responsable de la spécialité	Denis Vernant

Remerciements

Je tiens à remercier mes relectrices Alice et Anna et mon relecteur Régis pour leur aide et leur attention. Tous les trois ont contribué à améliorer l'orthographe et la clarté de ce mémoire, mais surtout, leurs secours n'ont cessé de m'encourager. Je remercie aussi Séverine et Jean pour la table que j'ai très largement occupée cet été ! Pour leur intérêt et leur soutien.

Je salue et remercie également Ezequiel pour m'avoir donné envie de monter cette année un projet qui maintenant nous est cher. Je remercie enfin Jean-Marc Vincent pour son honnêteté scientifique qui m'a enhardi ce semestre et Yves Demazeau pour ses commentaires vivifiants.

1 Introduction

1.1 Problématique générale

Les machines ont-elles un esprit ? Cette question abrupte ne peut être débattue sans en préciser les termes. Qu'est-ce qu'une *machine* ? Qu'est-ce qu'*avoir un esprit* ?

L'objectif de ce mémoire n'est pas de répondre directement à cette problématique, mais de s'interroger sur les moyens que nous avons d'y répondre. Cette première section donne donc une définition préliminaire aux termes de *machine* et d'*esprit* afin de préciser le cadre général de nos recherches. La problématique particulière de ce projet est formulée plus en détails dans la section 1.2.

1.1.1 Définir l'esprit

L'esprit est traditionnellement défini de deux manières différentes. Premièrement, l'esprit est *l'ensemble des facultés mentales* que possèdent les hommes et d'autres animaux. Ces facultés peuvent concerner, par exemple, le traitement de l'information, la mémoire, l'utilisation du langage, de la logique, la prise de décision, etc. Elles peuvent également désigner des processus plus élémentaires et souvent inconscients comme la perception, la motricité, les émotions. Nous associons à cette première définition, abordant l'esprit en termes d'aptitudes ou de mécanismes, la notion de *cognition* développée précisément par les *sciences cognitives*. La cognition est alors le terme scientifique désignant l'ensemble des mécanismes implémentés par le cerveau.

La seconde définition de l'esprit peut être formulée en termes phénoménologiques. *Avoir un esprit* est alors la faculté d'*avoir conscience de quelque chose*. Il peut s'agir d'une *conscience perceptive*, c'est-à-dire « le fait d'être immédiatement conscient de quelque chose – événement ou relation. »[1] Elle comprend alors l'ensemble des impressions sensibles qui apparait à un sujet : ce sont les formes, les couleurs, la tonalité perçue des émotions et d'autres états mentaux. Nous pouvons également parler de conscience « de plus haut niveau » : la *conscience réflexive*, qui « implique l'idée d'être conscient de ses propres perceptions et pensées, et par conséquent de sa propre existence. »[2] La définition

[1] Cf. (Denton, 1993) Pages 68 et 69
[2] Cf. (Denton, 1993) Page 69

phénoménologique de l'esprit est ainsi centrée sur le sujet pensant, sa perception, sa conscience du monde et de lui-même.

Ces deux définitions s'opposent par le fait que la première aborde l'esprit d'un point de vue objectif et la seconde d'un point de vue subjectif. Nous reviendrons dans la section 2.1.3 sur les modalités de cette tension et sur ce qu'elle implique dans le cadre de notre problématique particulière (cf. section 1.2). Retenons, dans un premier temps, qu'*avoir un esprit* signifie à la fois *être capable de cognition* et *être conscient*.

1.1.2 Définir la machine

Stevan Harnad, dans son article intitulé *Can a machine be conscious? How?*[3] s'interroge sur la notion de *machine*. Dans une perspective internaliste, le cogniticien précise que ce terme ne doit pas dépendre des origines de l'objet qu'il désigne. Ainsi, si un grille-pain venait à pousser sur un arbre de manière entièrement naturelle, il ne devrait pas échapper à la qualification de *machine* attribuée par le sens commun. De la même manière, si un homme était synthétisé molécule par molécule par un savant fou, il n'en serait pas moins ce que l'on appelle couramment *un homme*[4]. Selon ses termes, pour répondre à la question concernant l'esprit des machines, « we need a definition of "machine" that is strictly structural/functional, and not simply dependant on its historic origins. »[5] Harnad supprime ainsi la notion d'*artificialité* du concept de machine : une machine n'est pas nécessairement *construite par l'homme*.

Cette position sous-tend notamment l'idée que la *structure* d'un objet (la composition physique de ses éléments et leurs relations) et sa *fonction* (son comportement lorsqu'il est soumis aux lois de la physique) déterminent entièrement *ce qu'est* l'objet : *il n'y a rien d'autre*. Ainsi, la nature d'un homme et celle de son clone, créé de toute pièce par un savant fou, sont strictement identiques[6]. Les arguments exposés dans ce mémoire reposent sur cette position internaliste.

Harnad défini enfin une *machine* comme « *any causal physical system*. »[7] Le terme désigne à la fois le grille-pain et l'homme, l'ordinateur et la bactérie. Puisque nous sommes

---

[3] « Une machine peut-elle être consciente ? Comment ? » (Harnad, Can a machine be conscious? How?, 2003)
[4] Cf. exemples du *toaster* et du *clone*. Ibid.
[5] « Nous avons besoin d'une définition de "machine" qui soit strictement structurelle/fonctionnelle, et qui ne dépende pas simplement de ses origines historiques. » Ibid.
[6] Cela présuppose également que la *fonction* est entièrement réalisée par la *structure* physique. Le physicalisme défend cette hypothèse.
[7] « N'importe quel système causal physique. » Ibid.

des machines et que nous avons un esprit, nous pouvons affirmer que *certaines machines ont un esprit*. Au contraire, puisque nous sommes à peu près sûrs que les grille-pains n'ont ni mécanismes mentaux, ni conscience du monde, nous pouvons affirmer que *certaines machines n'en n'ont pas*.

1.1.3 Reformuler la problématique

A la question « les machines ont-elles un esprit ? » nous répondons dans un premier temps « cela dépend de quelle machine on parle. » Cette indécision nous incite à reformuler la problématique générale pour que celle-ci engendre le débat qu'elle mérite. Nous pouvons la formuler ainsi : *quelles machines ont un esprit, et pourquoi ?* Il s'agit de déterminer, parmi toutes les machines que nous connaissons, lesquelles ont un esprit et lesquelles n'en ont pas. Il s'agit également de préciser, pour chaque machine considérée, pourquoi nous avons donné une réponse plutôt qu'une autre.

Les machines qui nous intéressent dans ce mémoire ne sont pas les organismes biologiques mais les machines électroniques : programmes, ordinateurs, robots, etc. Le terme machine, puisqu'il regroupe machines électroniques *et* organiques, permet de traiter les unes et les autres de manière identique et de s'interroger, par exemple, sur la conscience des ordinateurs de manière analogue à la conscience humaine. La thèse générale soutenue par ce mémoire est celle défendue par l'Intelligence Artificielle forte[8] : il est possible, en théorie, de concevoir une machine de silicium qui soit doté d'un esprit similaire à de celui des hommes. La réalisation pratique d'une telle machine, quant à elle, n'est pas d'actualité. Nous sommes en effet bien loin de disposer de la technologie nécessaire. Harnad compare un tel projet à la construction par l'homme d'une roquette pouvant atteindre Alpha du Centaure[9]. Nul doute que nous pouvons *en théorie* en construire une ; nul doute que nous ne disposons pas *en pratique* de la technologie nécessaire.

La définition plurielle de la notion d'esprit que nous avons donné précédemment incite à nuancer les réponses à la problématique générale. Puisqu'il existe plusieurs niveaux de conscience et plusieurs types de processus cognitifs, une machine peut avoir certaines propriétés mentales et pas d'autres. Les travaux d'éthologie cognitive vont dans ce sens en précisant que l'ensemble des animaux ont un esprit « moins développé » que celui des hommes. Ainsi, nous pouvons poser la question de manière plus fine, en ciblant le trait de

[8] Intelligence Artificielle forte (*strong AI*), cf. (Searle, 1980) pour une définition du terme par opposition à l'Intelligence Artificielle faible (*weak AI*).
[9] Cf. (Harnad, Can a machine be conscious? How?, 2003)

l'esprit dont nous voulons détecter la présence chez une machine. Nous pourrions par exemple demander : *quelles machines ont une conscience perceptive, et pourquoi ?* Ou encore de manière plus précise : *quelles machines peuvent avoir un sentiment de douleur, et pourquoi ?*

1.2 Problématique particulière

Ce mémoire ne propose pas de résoudre directement la problématique générale présentée dans la section précédente. Il s'intéresse plus largement à ses *méthodes de résolutions*. En ce sens, il participe au débat épistémologique concernant l'évaluation des connaissances que nous pouvons avoir concernant l'esprit des machines électroniques. Cette section expose la problématique particulière que nous pouvons formuler ainsi : *comment déterminer quelles machines ont un esprit, et comment justifier cette détermination ?*

1.2.1 Les théories de l'esprit

Le travail de *détermination argumentée* concernant l'existence d'un esprit chez une machine nécessite l'usage d'un modèle général de l'esprit. Il s'agit d'une construction intellectuelle décrivant la nature et le fonctionnement de l'esprit, notamment son rapport avec le corps. La résolution de la problématique générale ne repose donc pas sur un choix arbitraire, mais sur un raisonnement entrepris dans le cadre d'une *théorie de l'esprit*. Les explications jointes à la réponse (i.e. les arguments en faveur ou en défaveur de l'existence d'un esprit) peuvent être exprimées par cette théorie sous la forme de conditions nécessaires et/ou suffisantes concernant l'existence d'une capacité mentale donnée chez une machine. Par exemple : *il suffit d'avoir des yeux et un cerveau interconnectés « de la bonne façon » pour avoir la sensation de rouge*, ou : *il est nécessaire d'avoir ce qu'on appelle des neurones miroirs pour être capable d'empathie*. La validité de ces arguments dépend bien évidement de la validité des théories qui les expriment.

Une théorie de l'esprit donne ainsi une *définition en intension* de ce qu'est « une machine avec un esprit », c'est-à-dire qu'elle donne l'ensemble des propriétés de ces machines. La *détermination argumentée* consiste à utiliser cette *définition en intension* pour déterminer, par l'analyse, l'existence d'un esprit chez une machine donnée. Elle permet par la suite de donner une *définition par extension* de ce qu'est « une machine avec un esprit », c'est-à-dire qu'elle détermine l'ensemble des machines qui ont les propriétés définies par la théorie. Cette méthode de résolution est représentée par le schéma suivant.

L'utilisation d'une théorie de l'esprit permet à première vue de répondre à la problématique générale de ce projet. Cependant, la section 2.1 présente une limite épistémologique à une telle méthode de résolution : il s'agit du *problème des autres esprits*. Cette limite fait notamment obstacle à l'évaluation des théories utilisées dans la mesure où l'on ne peut connaître l'ontologie des autres esprits[10]. Il est alors impossible d'évaluer les arguments avancés pour répondre à la problématique. Les définitions extensives de l'esprit s'avèrent être à leur tour des outils épistémologiques utiles à l'évaluation des théories. Ce mémoire définit et discute de tels outils (cf. sections 2.2 et partie 3).

1.2.2 La valse des théories

Deux modèles différents peuvent ne pas coïncider sur l'ensemble de leurs réponses à la problématique générale. Par exemple, le modèle cartésien de l'esprit affirme que les animaux (autres que les hommes) sont des machines inconscientes. Les travaux d'éthologie cognitive moderne soutiennent au contraire que les primates, parmi d'autres, possèdent un certain niveau de conscience[11].

Il est nécessaire de confronter les théories et de définir comment les unes peuvent être réfutées et les autres peuvent être validées. L'objectif de ce mémoire est de discuter des moyens de validation dont nous disposons. La thèse qui y est défendue est que les *définitions en intension* formulées par les théories ne peuvent être l'objet de vérifications directes. Il faut alors recourir aux *définitions en extension* pour discuter des différents modèles de l'esprit, les confirmer ou les infirmer. Le schéma suivant représente la méthode ainsi soutenue par ce mémoire.

[10] Cf. section 2.1.3
[11] Cf. par exemple (Denton, 1993) Chapitre 3, *La Conscience chez les animaux*

1.3 Contenu du mémoire

La partie 2 de ce mémoire analyse le problème exposé dans cette introduction en deux temps. La section 2.1 présente une limite épistémologique à la connaissance de l'esprit des autres machines. Cette limite repose sur notre incapacité à *connaître directement* ces autres esprits et met en péril la définition même de la notion d'esprit. La section 2.2 présente une méthode empirique pour redéfinir cette notion à partir de ce que l'on peut *directement observer*. Il s'agit du test de Turing.

La partie 3 s'intéresse à l'évaluation du test de Turing et, plus largement, à son utilisation pour l'évaluation des théories de l'esprit. Pour ce faire, la section 3.1 introduit une seconde méthode pour définir la notion d'esprit : le *sens fonctionnel*. La section 3.2 présente les différentes contradictions qui peuvent survenir entre les deux définitions (test de Turing et sens fonctionnel). La section 3.3 explique comment régler ces tensions définitionnelles et la section 3.4 expose les conséquences de leur résolution sur la notion ainsi élaborée et sur les modèles théoriques qui la sous-tendent. Les contradictions au sein du test de Turing permettent de confronter les théories entre-elles, via l'analyse de cas particuliers, et on procède ainsi à leur évaluation.

La partie 4 propose d'utiliser les outils développés dans les parties précédentes pour redécouvrir la valse historique des théories fonctionnalistes de l'esprit. La section 4.1 s'intéresse à l'évaluation de l'*hypothèse cognitiviste* ; la section 4.2 s'intéresse à l'évaluation du *connexionnisme* ; la section 4.3 à l'évaluation du *modèle énactif de l'esprit*. Cette dernière section expose enfin, à partir des notions développées dans ce mémoire, un débat concernant l'évaluation du *fonctionnalisme* lui-même.

La partie 5 fait enfin le bilan du travail exposé dans ce mémoire (section 5.1) et ouvre différentes perspectives de recherches (section 5.2).

2 Une définition empirique de l'esprit

Cette seconde partie discute du problème exposé dans l'introduction de ce mémoire. La section 2.1 présente une limite fondamentale à la connaissance que nous pouvons avoir de l'esprit des autres machines. Il s'agit d'une difficulté épistémologique traditionnellement nommée *le problème des autres esprits*. Elle empêche toute vérification ontologique des théories : on ne peut déterminer la nature et le fonctionnement des autres esprits. Il est ainsi impossible de procéder à leur évaluation. Pire, la notion même d'« esprit » perd sa signification. Elle peut néanmoins être redéfinie par le biais d'une méthode empirique empruntée au domaine de l'Intelligence Artificielle. Il s'agit du test de Turing, présenté dans la section 2.2. Cette seconde approche permet de répondre à la problématique générale par l'observation du comportement des machines.

La partie 3 montre comment le test de Turing peut également être utilisé comme outil épistémologique dans la valse des théories. Les termes introduits dans cette troisième partie permettent en effet de confronter les différents modèles de l'esprit, de les confirmer, de les infirmer ou encore d'en établir de nouveaux. Cet outil explique enfin comment construire une théorie cohérente dans le but de *déterminer quelles machines ont un esprit* et comment *justifier une telle théorie*.

2.1 Le problème des autres esprits

Comment justifier l'existence, chez les autres, d'un esprit comme le mien ? Cette question récurrente en philosophie de l'esprit interroge sur les possibilités de la connaissance humaine. Sommes-nous capables, individuellement, d'assurer l'existence d'autres esprits ? Le point de vue adopté par ce mémoire oppose à cette question ontologique une limite épistémique infranchissable.

2.1.1 Formulation du problème et origines

Le problème des autres esprits est ainsi introduit par René Descartes dans sa *Méditation Seconde*.

> *Si par hasard je ne regardais d'une fenêtre des hommes qui passent dans la rue, à la vue desquels je ne manque pas de dire que je vois des hommes […] ; et cependant que*

vois-je de cette fenêtre, sinon des chapeaux et des manteaux qui pourraient couvrir des machines artificielles qui ne se remueraient que par ressorts ? [12]

Le problème des autres esprits se heurte immédiatement à la limite épistémologique soulevée par Descartes : nous ne pouvons garantir leur existence. Cette incertitude, selon la présentation du problème par Alec Hyslop[13], est engendrée par une asymétrie entre (1) la connaissance que nous avons de l'existence de notre propre esprit et (2) la connaissance que nous avons de l'existence de ceux des autres. La première est *directe*, ou *immédiate*, comme le soutient par ailleurs Descartes dans sa *Méditation Seconde*. La seconde est au contraire *indirecte*. Nous accédons en effet à l'existence des autres esprits par la *médiation* de l'observation ou du raisonnement logique, elle n'est pas le fait immédiat d'une conscience de soi comme l'expérience du *cogito* nous le fait découvrir. La question n'est pas de savoir ce qui peut ou ne peut pas être *directement observé*, puisque l'observation est nécessairement vécue comme une activité de notre propre esprit, mais de ce qui peut ou ne peut pas être *directement connu*[14]. Or, dans la *philosophie du sujet* élaborée par Descartes, le primat gnoséologique – l'expérience du « je » – est la seule expérience immédiate. La découverte du monde – et par suite celle des autres esprits – est l'effet d'une médiation divine (dans le cas des *Méditations Métaphysiques*)[15].

Remarquons que dans la précédente citation de Descartes le terme « machines artificielles » est employé pour désigner des corps, semblables au nôtre en ce qui concerne leur comportement et leur structure, mais dépourvus d'esprit. Comme nous l'avons mis en évidence dans la section 1.1.2, où nous avons défini la notion de machine, le terme de Descartes est inapproprié pour deux raisons : (1) je suis moi-même une *machine* pourvue d'un esprit, (2) la notion d'*artificialité*, liée aux origines d'une machine, ne peut être prise en compte pour discuter de la présence d'un esprit dans le cadre d'une position internaliste. Dans

---

[12] Cf. (Descartes, 1641) *Méditation Seconde*

[13] Cf. (Hyslop, 2009) Section *The Epistemological Problem*

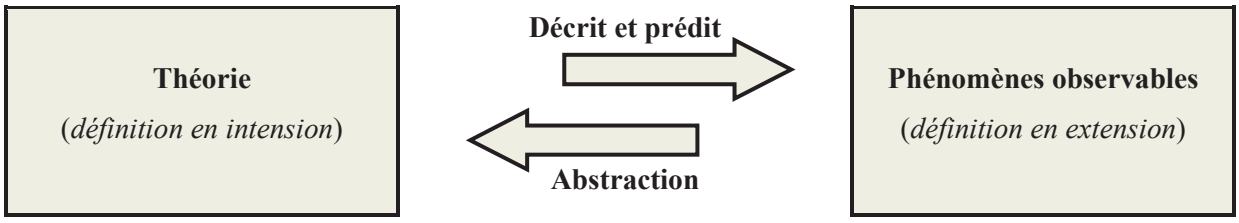
[14] Dans *Other Minds*, Alec Hyslop remarque que la capacité de télépathie ne résout pas le problème des autres esprits (cf. (Hyslop, 2009) Section *The Epistemological Problem*). Une sensation de douleur, émise par un autre esprit par télépathie, serait perçue comme partie intégrante de notre propre expérience. « What I need is the capacity to observe those mental states as mental states belonging to that other human being. » (« Ce dont j'ai besoin, c'est la capacité d'observer ces états mentaux comme des états mentaux appartenant à cet autre être humain. ») Une capacité que la télépathie ne peut offrir.

[15] La *philosophie de la connaissance* élaborée par Emmanuel Kant dans la *Critique de la raison pure* montre que les caractéristiques du sujet cartésien ne se donnent pas de manière immédiate : l'esprit est un *phénomène* pour lui-même. Cependant, l'existence du sujet, et non le caractère de ses propriétés *nouménales*, reste, dans les limites de la philosophie kantienne, une connaissance immédiate. Cela suffit pour garantir l'existence de notre esprit. Les critiques modernes de l'unité du soi, qui mettent en péril l'existence même du sujet, ne sont pas abordées dans le cadre de ce mémoire. Cf. (Varela, Thompson, & Rosch, 1993) Chapitre 6 *Un esprit dénué de soi*, (Minsky, 1988) Section 28.8 *Overlaping Minds* et les travaux de Ray Jackendoff.

un langage plus moderne, nous parlerons de *zombies* pour désigner de tels corps sans esprits. Ce terme et ses variations sont présentés dans la section 3.2.2.

2.1.2 La solution du fonctionnalisme

Plusieurs propositions ont été avancées pour résoudre le problème des autres esprits. La plus répandue prend la forme d'une *inférence de la meilleure explication*[16] : l'existence d'un esprit chez les autres êtres humains est ainsi la meilleure explication de leur comportement observable. Le raisonnement est le suivant : (1) je sais que je possède un esprit et je sais qu'il est la cause de mon comportement, (2) j'observe le comportement des autres qui m'apparait très similaire au mien, (3) il en découle, par inférence, que les autres ont un esprit et qu'il est la cause de leurs comportements. Le raisonnement soutenant cette inférence est de même nature que les raisonnements à l'œuvre notamment dans les sciences physiques. Carnap distingue à ce titre *théories* et *observables*[17]. Les premières sont des modèles explicatifs des seconds (les phénomènes observables). Elles n'ont cependant jamais valeur d'ontologie dans la mesure où elles restent des modèles probables, des outils épistémologiques évalués en fonction de leur expressivité et de leur efficacité de prédiction. *L'existence d'autres esprits* est ainsi une théorie simple et efficace pour rendre compte du comportement observable des autres êtres humains. Ce qui est *directement observable* permet ainsi de rendre compte de ce qu'on ne peut connaitre directement par un processus d'abstraction.



Le *risque empirique*[18] engendré par cette inférence en faveur de l'existence d'autres esprits est de même nature que les risques engendrés par toute forme de raisonnement inductif. La philosophie de l'esprit est encore une fois limitée par une incertitude épistémique. Cependant, Harnad propose, dans son article concernant la conscience des machines[19], de ne pas trop se soucier de cette incertitude. Dans le cas de la physique, malgré le problème de

[16] « Inference of the best explanation » (Hyslop, 2009) Section *The Epistemological Problem*
[17] Exprimées dans un *langage des observations* (*Observation Language*) et un *langage théorique* (*Theoretical Language*) (Carnap, 1966)
[18] Risque lié à la contingence des lois de la physique. (Harnad, Can a machine be conscious? How?, 2003)
[19] Cf. (Harnad, Can a machine be conscious? How?, 2003)

l'induction ou des non-observables, l'homme continue à construire des théories. De la même manière, les sciences cognitives peuvent élaborer des modèles et des théories affirmant l'existence d'autres esprits et en étudier la nature à partir de leurs conséquences observables (leur comportement). Searle énonce ainsi, lorsqu'on oppose à son expérience de la chambre chinoise (exposée dans la section 4.1.2) le problème des autres esprits, « in "cognitive sciences" one presupposes the reality and knowability of the mental in the same way that in physical sciences one has to presuppose the reality and knowability of physical objects[20]. » Le risque empirique est ainsi engagé dans un pari, bien connu des physiciens, concernant l'uniformité de la nature, de ses lois, et concernant également la possibilité pour les hommes de les identifier par l'observation.

Le *fonctionnalisme* propose de résoudre le problème des autres esprits sur la base de l'inférence exposée précédemment. Selon cette théorie de l'esprit, les états mentaux sont caractérisés par leur rôle fonctionnel liant les stimuli environnementaux aux comportements observables. Leur place dans cette chaîne causale (théorisée à l'origine par le béhaviourisme) permet de les identifier et d'en étudier la nature. Finalement, pour répondre de l'existence d'un esprit chez un autre être humain, il suffit d'observer consciencieusement son comportement en fonction des circonstances extérieures.

2.1.3 Les autres esprits n'existent pas

Jackson formule une objection qui déstabilise considérablement le projet du fonctionnalisme. Dans son article *Epiphenomenal Qualia*, le philosophe affirme qu'il n'y a aucune raison pour que les états mentaux soient la cause des comportements observables[21]. Ils apparaissent alors comme une simple conséquence des processus cérébraux, eux-mêmes responsables du comportement. Relégué au statut d'épiphénomène, l'existence de l'esprit ne peut être mise en évidence par l'inférence de la meilleure explication. L'étude des comportements observables ne donne d'information, selon Jackson, que sur les états cérébraux. Les états mentaux sont écartés de la chaîne causale et ne participent plus à l'inférence qui permettait au fonctionnalisme de résoudre le problème des autres esprits.

Harnad, à la fin de son article, rejoint le point de vue pessimiste de Jackson en ce qui concerne les possibilités de la connaissance humaine. La dernière phrase affirme : « the ghost

[20] « Dans le domaine des "sciences cognitives", on présuppose la réalité et l'identifiabilité du mental, tout comme en sciences physiques, on doit présupposer la réalité et l'identifiabilité des objets physiques. » (Searle, 1980) Section V, *The other minds reply*, traduction de É. Duyckaerts

[21] Cf. (Jackson, 1982) Section *The Bogey of Epiphenomenalism*

in the machine is destined to continue to haunt us even after all cognitive science's empirical work is done[22]. » Nous pensons que cette désillusion concernant la portée épistémologique des neurosciences ou du fonctionnalisme (pour citer deux branches des sciences cognitives) est très bien exprimée par la critique nagelienne du physicalisme. Cette critique est magistralement exposée par Thomas Nagel dans son célèbre article intitulé *What Is It Like to Be a Bat?*[23] Nous retenons de la thèse qui y est défendue que deux points de vue coexistent. Le point de vue *objectif*, celui des sciences physiques, correspond à un *point de vue de nulle part*[24], un point de vue pour lequel l'observateur est idéalisé, absent du monde. C'est par exemple l'œil dessiné dans les travaux d'optique géométrique par Descartes, dont on suppose qu'il occupe une position ponctuelle dans l'espace et qu'il ne perturbe jamais l'expérience. Rien n'est précisé de son comportement et de la manière dont il observe. L'ajout d'un modèle de l'observateur, décrit par exemple en termes neurologiques, n'introduit pas de nouveau point de vue. En effet, cet « observateur » est un objet théorique modélisé par le scientifique. C'est ce dernier qui occupe le *point de vue de nulle part* : le scientifique (le réel observateur) est lui-même absent de la théorie. Le second point de vue répertorié par Nagel est le point de vue *subjectif* de la phénoménalité. L'expérience y est toujours définie en fonction d'un observateur incarné. Le modèle de cet observateur, incluant son expérience subjective, est alors pris en charge par la théorie.

Selon Nagel, le physicalisme rend parfaitement compte des *faits objectifs* du monde[25]. Cependant, il manque à l'analyse des *expériences subjectives*[26], tout simplement parce qu'il n'intègre aucun observateur dans ses modes de description du monde. Notre sentiment est que le problème des autres esprits est éclairé par cette distinction fondamentale. Nous assimilons les points de vue de Nagel à autant d'observateurs, c'est-à-dire à autant d'esprits. Dans le cadre du *point de vue de nulle part*, il n'existe donc aucun esprit puisqu'il n'y a aucun observateur incarné. Pour le physicalisme, nous sommes tous « des chapeaux et des manteaux […] qui ne se [remuent] que par ressorts, » ressorts dont les lois du mouvement sont identifiables en soi par la physique. Dans le cadre du point de vue subjectif, il existe un esprit et un seul, celui de l'observateur, celui du *cogito*. Ce qui nous amène à formuler la conclusion suivante : l'existence des autres esprits ne peut être déterminée par le physicalisme, puisque

[22] « Le fantôme dans la machine est voué à nous hanter même après que le travail empirique des sciences cognitives a été achevé. » (Harnad, Can a machine be conscious? How?, 2003)
[23] « Quel effet cela fait, d'être une chauve souris ? » (Nagel, What Is It Like to Be a Bat?, 1974)
[24] Cf. (Nagel, The View From Nowhere, 1986) où est présentée une discussion plus large concernant le sujet.
[25] Nous pourrions parler, en termes kantiens, de *noumènes*.
[26] Les *phénomènes* au sens de Kant.

l'observation désincarnée ne présuppose aucun esprit. Elle ne peut non plus être déterminée par le point de vue subjectif puisque qu'il n'existe alors qu'un seul point de vue, c'est-à-dire qu'un seul esprit. Nous sommes dans les deux cas dans une impasse, les autres esprits *n'existent pas*.

2.1.4 Redéfinir l'esprit

Le problème des autres esprits donne une réponse apparemment définitive à la problématique générale de ce mémoire. A la question « quelles machines ont un esprit, et pourquoi ? » tout homme est forcé de répondre « *je* suis la seule machine dotée d'un esprit, parce que *je* suis le seul esprit qui existe de mon point de vue. » Les théories de l'esprit ne se réfèrent alors plus à rien sinon à notre propre conscience. Elles ne sont plus générales mais particulières. Elles ne peuvent être évaluées de manière objective dans la mesure où, du point de vue physicaliste, elles ne s'appliquent à aucun objet.

Aucune méthode d'investigation, si elle s'intéresse à l'ontologie des autres esprits, ne peut dépasser la réponse close formulée par le problème des autres esprits. Il est nécessaire, pour aller plus en avant, de reconstruire la notion d'esprit qu'il met à mal. La suite de cette partie propose une telle définition, fondée sur une méthode empirique empruntée à l'Intelligence Artificielle : le *test de Turing*. Clarifions, avant de nous y intéresser, les objectifs et contraintes exigées par ce travail de reconstruction sémantique.

1. La dimension phénoménale de l'esprit doit être écartée de la nouvelle définition. Elle empêche en effet la construction d'un sens général dans la mesure où il n'y a toujours qu'une seule phénoménologie. L'esprit doit donc être défini uniquement en termes de mécanismes ou d'aptitudes cérébrales (cf. section 2.2.3).
2. L'esprit doit être défini depuis le point de vue subjectif, afin de conserver une parenté avec l'esprit au sens ontologique du terme. Le sujet énonçant une théorie de l'esprit prend ainsi part au modèle et inclut dans sa définition une dimension qui lui est propre (cf. sections 2.2.2 et 3.1.2).
3. La nouvelle définition doit reposer dans un premier temps sur l'observation. Puisque l'esprit ne peut être *directement connu* il est nécessaire de s'appuyer sur ce qui est *directement observable* (cf. section 2.2.2). La définition de l'esprit ne doit pas prétendre à une description ontologique de l'esprit (via une inférence de la meilleure explication par exemple) mais à une description épistémique.

Nous verrons que, dans le cas des machines électroniques construites par l'Intelligence Artificielle, nous avons également accès à leur *fonctionnement*. Cette notion, s'opposant à celle de *comportement* observable, permet de donner une seconde définition extensive de l'esprit. La confrontation sémantique entre ces deux définitions est utilisée dans la partie 3 pour évaluer le test de Turing et participer à la valse des théories.

2.2 Le test de Turing

Le pessimisme de Harnad quant à l'échec du physicalisme[27] pour décider de l'existence d'autres esprits donne naissance à une volonté nouvelle de répondre au problème. Face à la perte de sens du mot « esprit » lorsqu'on parle de celui des autres, il est nécessaire d'en donner une nouvelle acception. Le test de Turing permet d'en construire une en termes de comportements observables.

2.2.1 Présentation du test

A l'aube de l'Intelligence Artificielle, il fut nécessaire de définir le terme même « d'intelligence ». La question posée est la suivante : *quand peut-on dire qu'une machine est ou n'est pas intelligente ?* Le mathématicien Alan Turing propose alors une méthode empirique pour répondre à cette question : *The Imitation Game*[28]. Pour résumer son idée, la définition *a priori* du terme « intelligence » est remplacée dans ce test par une épreuve au cours de laquelle l'intelligence de la machine est identifiée empiriquement. La proposition « cette machine est intelligente » a pour condition de réalisation « elle se comporte comme un homme intelligent ».

En pratique, le test nécessite donc une machine A qui passe le test, une machine B qui sert de référence (e.g. un homme intelligent) et un observateur (ou juge). L'observateur discute avec les machines A et B par le biais d'un terminal. Il leur pose des questions et détermine, en comparant les réponses, si la machine A est intelligente, en présupposant que la machine B l'est. Le test répond donc à la problématique générale de la manière suivante : *la machine A est intelligente si, et seulement si, elle se comporte comme la machine B*. Il établit une *condition nécessaire et suffisante* à la notion d'intelligence et constitue donc une véritable *définition*, celle-ci étant fondée sur le comportement observable des machines.

---

[27] Cf. (Harnad, Can a machine be conscious? How?, 2003) Fin de l'article

[28] « Le Jeu de l'Imitation » (Turing, 1950), que nous appelons depuis « le test de Turing. »

Avant de faire deux remarques concernant la conception de l'esprit tacitement reconnue par le test, nous souhaitons préciser que la présence de la machine de référence (machine B) lors du test n'est pas nécessaire. En effet, la comparaison des comportements ne repose pas sur une relation d'identité. De fait, il existe de nombreuses machines que nous qualifions d'intelligentes – qui peuvent donc servir de référence – et qui ont toutes un comportement qui diffère dans les détails ou même selon des caractéristiques plus importantes. Deux hommes intelligents ne répondront pas nécessairement de la même manière à une même question. Ce qui est évalué par l'observateur, c'est un type général de comportement. Celui-ci est la généralisation de ce qu'il conçoit comme un *comportement intelligent possible*. Il compare ainsi le comportement de la machine A aux comportements de *toutes les machines de référence possibles*.

2.2.2 Définition empirique de l'intelligence

Le test de Turing hérite clairement de la méthode béhavioriste pour laquelle l'esprit n'est détectable qu'à partir de l'observation de comportements, en réponse à des stimuli extérieurs. De cette manière, le test définit l'intelligence en termes d'observables et évite ainsi l'obstacle épistémologique de ce qui ne peut être *directement connu*. Pour le béhaviorisme, l'esprit ne peut être en effet qu'*indirectement connu*, ou *connu a posteriori*, via l'expérimentation et l'analyse de la chaîne causale Stimuli-Réactions. Le problème des autres esprits est résolu assez facilement en modifiant la définition même de l'intelligence. Il faut bien comprendre que, avec ce qui a été dit précédemment, le test de Turing n'est pas envisagé dans le cadre de ce mémoire comme une inférence de la meilleure explication, auquel cas il serait soumis à l'objection épiphénoménaliste de Jackson. Ici, le test ne prétend pas décider de l'ontologie des autres esprits, mais permet de définir la notion même d'esprit à partir de ce qui est décidable. Les rapports causaux étudiés relient uniquement des phénomènes observables (stimuli et réactions). L'intelligence n'a pas valeur d'ontologie.

Le test de Turing renonce ainsi à donner une définition indépendante de l'expérience. Lors d'un test de Turing, l'intelligence est toujours définie de manière subjective par un observateur incarné. En ce sens, un robot est intelligent « selon telle ou telle personne ». La notion de relativisme dans la construction sémantique du terme indique que le test s'écarte du point de vue objectif tel qu'il est conceptualisé par Nagel. L'intelligence n'est pas définie *depuis nulle part*. Cette déclaration chagrine sans aucun doute la tradition scientifique qui fait grand cas de l'objectivité et l'on peut attaquer le test sur le fait qu'il ne fournit aucune

définition absolue de l'intelligence. Pour sa défense, Harnad rappelle que nous faisons constamment du *mind-reading*[29] lorsqu'il s'agit de déterminer ce que les autres « ont en tête ». Le test de Turing systématise ce que nous faisons déjà en permanence et, en vérité, nous ne pouvons faire que cela. Le problème des autres esprits donne le dernier mot à la définition empirique du test. Lorsqu'on objecte que deux *noumènes* peuvent se cacher derrière le même *phenomène*, Harnad répond « if I can't tell the two apart empirically, I'd best not try to make too much of that distinction. »[30]

2.2.3 L'intelligence comme compétence

La méthode élaborée par Turing repose sur un *test dialogique de compétence*[31]. Le test repose en effet sur un dialogue entre deux machines ; l'intelligence y est envisagée comme une compétence à détecter, c'est-à-dire comme quelque chose qu'une machine *est capable de faire*. En ce sens, l'intelligence n'est pas une propriété *structurelle*, mais une propriété *fonctionnelle*.

Dans l'introduction de son article, Turing présente une version du test dans laquelle la propriété à identifier est le fait d'être un homme ou le fait d'être une femme[32]. Cet exemple nous semble être assez mal choisi dans la mesure où le genre d'un individu peut être identifié, dans la plupart des cas, à partir de propriétés structurelles directement observables (comme par exemple l'observation des organes de la reproduction). L'intérêt du test de Turing est justement de décider des propriétés fonctionnelles des machines, propriétés qu'un examen de leur structure ne peut révéler. A ce titre, l'exemple choisi par Turing pour introduire son test ne rend pas hommage à son intérêt pour le fonctionnalisme. Le test est capable de décider, par exemple, si un individu *sait siffler* à partir de son comportement lorsqu'on lui demande de siffler. Ce qui est bien évidemment indécidable sur la base d'une simple dissection du corps de l'individu dans la mesure où on ne trouvera pas de sifflet dans la gorge de celui-ci.

Il est important de noter que le test de Turing peut être reformulé pour tester dialogiquement d'autres compétences que l'intelligence. A l'origine, Turing s'intéresse d'ailleurs à la question « can machine think? »[33] donnant une portée plus large au test. Ce n'est

---

[29] « The capacity […] to detect or infer what others "have in mind." » (« La capacité […] de détecter ou d'inférer ce que les autres "ont à l'esprit." ») (Harnad, Can a machine be conscious? How?, 2003)

[30] « Si je ne peux pas empiriquement distinguer l'un de l'autre, je ferais mieux de ne pas faire grand cas de cette distinction. » (Harnad, Can a machine be conscious? How?, 2003)

[31] Cf. (Putnam, 1981) Section *Le Test de Turing*

[32] Cf. (Turing, 1950) Section *The Imitation Game*

[33] « Les machines peuvent-elles penser ? » (Turing, 1950) Section *The Imitation Game*, page 433

que dans les relectures de sa méthode que l'on proposa d'associer le test à la notion même d'intelligence. Putnam l'utilise également pour identifier l'*intentionnalité*[34] et on peut trouver d'autres exemples. Le fait de posséder un esprit, d'avoir des expériences sensibles, une conscience réflexive peut au même titre être identifié à une compétence dont l'existence est décidable par le test de Turing. Ceci en fait un test générique.

L'esprit et ses états mentaux peuvent être également définis comme des compétences du cerveau, et non comme des propriétés structurelles. Lewis parle ainsi d'*aptitudes*[35] ; Minsky affirme « *minds are simply what brains do*. »[36] Le test de Turing, en tant que *test dialogique de compétence*, est alors adéquat et efficace pour élaborer une définition fonctionnaliste de l'esprit. Le *fonctionnalisme* est une théorie de l'esprit qui définit les états mentaux en fonction de la place qu'ils occupent dans le fonctionnement du cerveau. Il s'agit d'une théorie purement causale de ces états, qui écarte leurs aspects subjectifs et phénoménologiques (i.e. les *qualia*). En ce sens, le fonctionnalisme répond parfaitement aux exigences de la nouvelle définition de l'esprit (cf. section 2.1.4, point 1). Le travail effectué dans le cadre de ce mémoire repose donc sur une telle conception. Le fonctionnalisme y est envisagé comme une théorie générale de l'esprit au sein de laquelle gravitent des théories particulières plus précises : e.g. *cognitivisme*, *connexionnisme*, *modèle énactif de l'esprit*, etc. La partie 4 présente et évalue ces théories internes au fonctionnalisme.

L'indistinguabilité des compétences entraîne leur identité. Ce type de raisonnements peut enfin être comparé à ceux utilisés par Albert Einstein pour penser le *principe d'équivalence*[37]. L'ascenseur d'Einstein est ainsi une célèbre expérience de pensée où un observateur localement situé dans l'espace et dans le temps (et en ce sens *incarné*) est soumis à des champs gravitationnels et des accélérations de son référentiel. Puisque les effets de ces expériences sont indistinguables pour cet observateur, Einstein en conclu qu'ils doivent être exprimés de manière identique dans le langage de la physique. Il fonde ainsi la théorie de la relativité restreinte. Par analogie, si un cerveau, biologique ou électronique, produit les

---

[34] Cf. le *test de Turing pour la référence* (Putnam, 1981) Section *Le Test de Turing*

[35] Cf. (Lewis, 1978) Postface de 1983 en réponse à Nagel

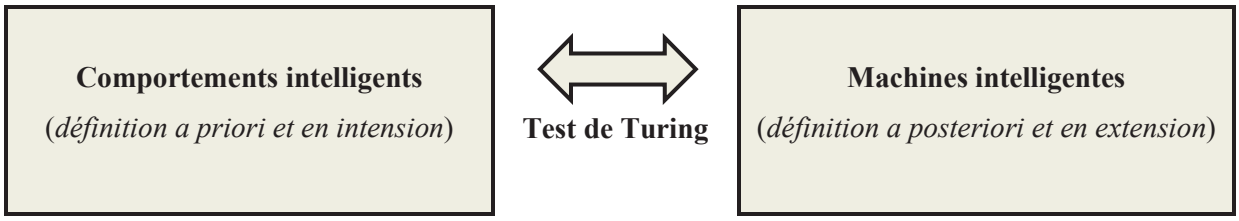
[36] « L'esprit est simplement ce que fait le cerveau » Cf. (Minsky, 1988) Section *The Mind and the World*. A l'origine, le sens de cette affirmation est une description de l'esprit en termes de processus produits par le cerveau. Il est possible de traduire : « l'esprit est simplement ce que le cerveau *produit*. » Mais nous pouvons également interpréter la thèse de Minsky, selon la conception fonctionnaliste, comme une description en termes de compétences cérébrales : « ce que le cerveau *fait*. »

[37] Principe selon lequel les effets d'un champ gravitationnel sont identiques aux effets d'une accélération du référentiel de l'observateur, cf. (Einstein, 1921)

mêmes effets qu'un autre cerveau pour un observateur incarné, alors l'activité des deux cerveaux doivent être décrit de la même manière par une théorie de l'esprit.

2.2.4 Bilan

Le test de Turing permet de redonner un sens général à la notion d'intelligence à partir de l'observation des comportements. Il définit ainsi un lien nécessaire et suffisant entre des *comportements* définis *a priori* comme intelligents (ils constituent les comportements de référence que l'on utilise pendant le test pour juger les comportements observés) et des *fonctionnements* définis *a posteriori* comme intelligents (il s'agit du résultat du test, déterminant si la machine a ou n'a pas un fonctionnement intelligent). En outre, le test peut être reformulé pour donner un sens général à la notion d'esprit ou à tout autre concept envisagé en termes de compétence ou de fonction. Le schéma suivant résume cette méthode permettant de répondre à la problématique générale de ce mémoire.



Le test de Turing est fondé sur un modèle générique : le *fonctionnalisme*. Les théories utilisant le test doivent donc reposer sur cette conception fonctionnelle de l'esprit. Contrairement à ces théories, le test de Turing donne une définition de l'esprit *en extension* : il définit empiriquement *l'ensemble des machines qui ont un esprit* et *l'ensemble des machines qui n'en ont pas*, sur la base de comportements catégorisés *a priori*. La partie suivante explique comment cette définition extensive peut être utilisée pour confronter et évaluer les différentes théories : par l'étude de cas particuliers appartenant à ces ensembles. Elle nécessite notamment l'introduction d'une seconde définition extensive de l'esprit que nous appelons le *sens fonctionnel* (cf. section 3.1.2). Le schéma ci-dessous rappelle la méthode que nous avions annoncée dans l'introduction de ce mémoire et que nous explicitons dans la partie suivante.

3 Évaluer les théories de l'esprit

Le test de Turing donne une définition extensive et *a posteriori* à la notion de fonctionnement intelligent et, par suite, à la notion d'esprit au sens du fonctionnalisme : *les machines possédant un esprit sont celles qui passent le test avec succès et uniquement celles-là*. Nous disons qu'elles ont un esprit *au sens de Turing*.

Selon Harnad, le test de Turing a le dernier mot[38]. La section qui suit s'oppose à cette assertion et révèle les contradictions que peut engendrer le test. Elle montre comment les limites du test de Turing – ses résultats pouvant être erronés – peuvent servir à l'évaluation des différents modèles de l'esprit. La notion d'erreur est définie en fonction d'une seconde définition extensive de l'esprit, déterminée par ce que nous appelons le *sens fonctionnel*. Les éventuelles tensions entre les deux définitions (test de Turing et sens fonctionnel) sont utilisées pour discuter des théories à évaluer. Un modèle de l'esprit doit garantir une certaine cohérence entre les deux définitions et régler les tensions éventuelles. L'utilisation conjointe du test de Turing et du sens fonctionnel nous incite donc à confirmer ou à infirmer les théories et nous indique comment concevoir un modèle de l'esprit cohérent.

3.1 Évaluer le test de Turing

3.1.1 Sensibilité et spécificité

L'efficacité d'un test quelconque peut être exprimée en termes de *sensibilité* et de *spécificité*. La *sensibilité* est la probabilité que le test donne un résultat *positif* pour un objet *possédant* la propriété à détecter. Ainsi, lorsque la sensibilité est inférieure à 100%, il existe des cas de *faux négatifs* (i.e. un résultat négatif alors que la propriété est présente). La *spécificité* est au contraire la probabilité que le test donne un résultat *négatif* pour un objet *ne possédant pas* la propriété à détecter. Une spécificité inférieure à 100%, indique qu'il existe des cas de *faux positifs* (i.e. un résultat positif alors que la propriété est absente).

Les notions de sensibilité et de spécificité sont utilisées pour évaluer l'efficacité d'un test. Il s'agit de déterminer s'il répond correctement à la question : *est-ce que cet objet possède ou non la propriété à détecter ?* Un test est ainsi efficace à 100% lorsqu'il ne présente aucun cas de *faux négatif* ou de *faux positif*. Pour réaliser une telle évaluation, il est

[38] Cf. (Harnad, Can a machine be conscious? How?, 2003) Fin de l'article

donc nécessaire de posséder un échantillon d'objets dont on sait, par un moyen indépendant du test, s'ils ont ou non la propriété à détecter. Un test n'est ainsi évalué qu'en fonction d'une seconde définition extensive de la propriété. Cette définition est une *définition extensive a priori*, puisqu'elle est indépendante de l'expérience menée dans le cadre du test. Le test constitue au contraire une *définition extensive a posteriori* de la propriété qu'il est sensé détecter.

	Test positif	**Test négatif**
Propriété présente	Vrais positifs	Faux négatifs
Propriété absente	Faux positifs	Vrais négatifs

Les notions de sensibilité et de spécificité peuvent être représentées par un diagramme de Carroll (cf. ci-dessus). L'ensemble des objets est représenté selon deux dimensions : *détention de la propriété* et *résultat du test*. Il est découpé selon ces deux dimensions en quatre sous-ensembles représentés par des rectangles. L'ensemble des objets possédant (resp. ne possédant pas) la propriété à détecter est représenté par le rectangle du haut (resp. le rectangle du bas) ; ils sont donnés par la *définition extensive a priori*. L'ensemble des objets pour lequel le test donne un résultat positif (resp. négatif) est représenté par le rectangle de gauche (resp. le rectangle de droite) ; ils sont donnés par la *définition extensive a posteriori*. Ces sous-espaces découpent finalement l'ensemble des objets en quatre sous-espaces unitaires :

1. Le rectangle en haut à gauche représente l'ensemble des *vrais positifs*, i.e. les objets *possédant* la propriété et pour lesquels le test donne un résultat *positif*.
2. Le rectangle en bas à droite représente l'ensemble des *vrais négatifs*, i.e. les objets *ne possédant pas* la propriété et pour lesquels le test donne un résultat *négatif*.
   Ce sont les cas où le test parvient à détecter correctement la présence ou l'absence de la propriété.
3. Le rectangle en bas à gauche représente l'ensemble des *faux positifs*, i.e. les objets *ne possédant pas* la propriété et pour lesquels le test donne un résultat *positif*.
4. Le rectangle en haut à droite représente l'ensemble des *faux négatifs*, i.e. les objets *possédant* la propriété et pour lesquels le test donne un résultat *négatif*.

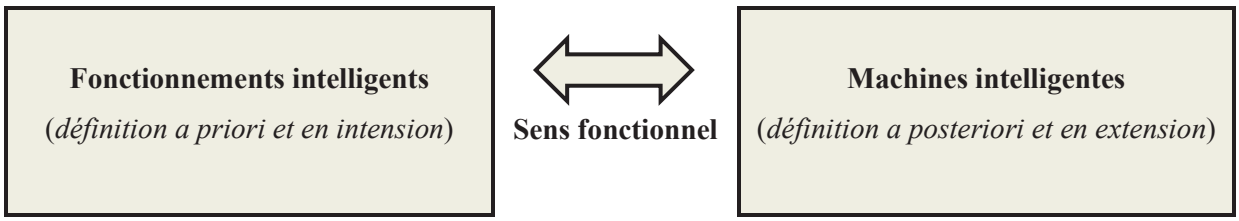
Ce sont les cas où le test engendre une erreur de détection.

Le test est ainsi efficace à 100% s'il n'y a que des *vrais positifs* et des *vrais négatifs*. Dans ce cas les rectangles (3) et (4) sont vides.

3.1.2 Le sens fonctionnel

Pour évaluer le test de Turing, il est nécessaire de disposer d'une *définition extensive a priori* de l'intelligence. Nous appelons celle-ci le *sens fonctionnel*. Intuitivement, il repose sur une notion d'intelligence que nous avons conceptualisée de manière informelle au cours du temps. Tout le monde aura une idée de ce que signifie pour lui un « fonctionnement intelligent ». Cette définition peut être empruntée, par exemple, au fonctionnement que nous avons nous-mêmes lorsque nous réfléchissons ou pensons à quelque chose. Elle définit les fonctionnements intelligents *a priori*, sans ce poser de questions sur ce que cela peut impliquer en termes de comportement.

Le sens fonctionnel est à ce titre la définition duale de celle réalisée par le test de Turing. En effet, le test définit la notion d'intelligence *a posteriori* à partir de comportements choisis *a priori*. Il donne à ce titre un sens comportemental à l'intelligence. L'observateur dit en effet : *tel comportement est intelligent, donc je définis toutes les machines qui ont ce comportement comme des machines intelligentes*. Le sens fonctionnel fait exactement l'inverse : il définit la notion de comportements intelligents *a posteriori* à partir de fonctionnements intelligents *a priori*. L'observateur dit : *tel fonctionnement est intelligent, donc je définis toutes les machines qui fonctionnent ainsi comme des machines intelligentes*.



La notion d'intelligence selon le sens fonctionnel est remplacée par la notion de *fonctionnements intelligents*. Ils constituent la partie cachée des machines, causalement liés aux comportements, et donnent une description fonctionnelle de leurs mécanismes internes. Les fonctionnements peuvent être formalisés par les tables de machine de Turing, par des automates à état, ou plus largement par la description de processus reliant causalement des états cérébraux.

Les recherches en neurocognition et neuropsychologie fournissent certaines méthodes pour modéliser les fonctionnements humains. Elles ne sont pas, à ce jour, suffisamment avancées pour donner une description exhaustive de ce que pourrait être la table de machine d'un individu donné. Cependant, elles sont capables de décrire des schémas fonctionnels généraux, « de haut niveau », donnant une idée sur le fonctionnement global du cerveau humain. Le fonctionnement des machines électroniques (ordinateurs, robots, programmes) est donné de manière immédiate par leur code d'exécution, formulé dans des langages de programmation dont nous connaissons toutes les spécifications. Nous supposons ainsi, lorsque nous nous interrogeons sur la notion d'esprit électronique, que nous possédons l'ensemble de sa description fonctionnelle. Cela est possible uniquement parce que ces machines sont construites par l'homme et que leur modèle de conception est ainsi connu *a priori*. Leurs comportements, au contraire, ne sont pas toujours prévus par le modèle et nécessite souvent l'implémentation et l'exécution de celui-ci pour les déterminer *a posteriori*. Le sens fonctionnel que nous avons formalisé dans cette section procède de cette manière, il juge *a priori* les fonctionnements et tire des conclusions *a posteriori* sur les comportements qu'ils induisent.

3.1.3 Application au test de Turing

Il est possible d'évaluer le test de Turing en fonction de la définition duale reposant sur le sens fonctionnel. Nous appelons les machines dont le résultat du test est un *faux négatif* des *fous* et celles dont le résultat est un *faux positif* des *zombies*. Le diagramme de Carroll suivant représente la spécificité et la sensibilité du test vis-à-vis du sens fonctionnel.

	Test de Turing positif	**Test de Turing négatif**
Intelligence selon le sens fonctionnel	Machines intelligentes	Fous
Absence d'intelligence selon le sens fonctionnel	Zombies	Machines non-intelligentes

En termes de fonctionnements et de comportements, cela donne le diagramme suivant.

	Comportement intelligent	**Comportement non-intelligent**
Fonctionnement intelligent	Machines intelligentes	Fous
Fonctionnement non-intelligent	Zombies	Machines non-intelligentes

Les cas de *fous* et de *zombies* sont engendrés par une contradiction entre le sens fonctionnel et le test de Turing. Les *fous* sont des machines au fonctionnement intelligent et au comportement non-intelligent (cf. section 3.2.1). Les *zombies* sont des machines au fonctionnement non-intelligent et au comportement intelligent (cf. section 3.2.2). Les cas de *vrais positifs* et de *vrais négatifs* ne sont pas problématiques dans la mesure où le sens fonctionnel et le test de Turing s'accordent dans leur définition. Ils définissent l'ensemble des *machines intelligentes* et l'ensemble des *machines non-intelligentes*.

Les sections qui suivent expliquent l'utilisation des termes de *fous* et de *zombies*. La propriété à détecter dans l'ensemble des exemples présentés est l'intelligence, mais nous rappelons que le test de Turing est générique. Ainsi, les termes de *fou* et de *zombie* sont eux-mêmes génériques et peuvent être redéfinis en fonction d'autres propriétés/compétences de l'esprit.

3.2 Des contradictions au sein du test

3.2.1 Des fous passant le test

Un *fou* est une machine intelligente qui échoue lors d'un test de Turing concernant l'intelligence, c'est-à-dire que son comportement conduit l'observateur à décider que cette machine n'est pas intelligente. Nous utilisons le terme « fou » en référence à l'homme décrit par Lewis dans le premier paragraphe de son texte intitulé *Mad Pain and Martian Pain*[39]. Celui-ci est un homme dont l'état mental associé à la douleur est le même que le nôtre. Cependant, les causes extérieures de la douleur et les effets comportementaux qui y sont associés diffèrent fondamentalement du commun des mortels. Dans le cas de l'intelligence, le fou a un *fonctionnement* interne similaire au nôtre (il sait par exemple lire, parler et compter dans sa tête), mais il ne se *comporte* pas de la même manière que nous (il préfère par exemple

[39] « Douleur de fou et douleur de martien » (Lewis, 1978)

écrire n'importe quoi et compter n'importe comment, sans doute par esprit de contradiction ou par simple jeu). Il est ainsi impossible, par la méthode behaviouriste du test de Turing, d'identifier la douleur ou l'intelligence d'un fou, excepté si on compare son comportement à celui d'un autre fou (i.e. qui se comporte de la même manière que lui), dont on décide *a priori* qu'il a mal ou qu'il est intelligent vis-à-vis de son fonctionnement interne.

Le test de Turing nécessite que l'on compare le comportement de la machine observée avec celui d'une machine de référence, dont le fonctionnement est présupposé intelligent par le sens fonctionnel. Le résultat du test est donc toujours un résultat relatif. On ne dira pas « cette machine est (ou n'est pas) intelligente, » mais plutôt « cette machine a (ou n'a pas) la même intelligence que cette autre machine. » Ainsi, le problème des fous peut être réglé en comparant un fou à un autre, encore faut-il que le sens fonctionnel définisse le second fou comme fonctionnellement intelligent. Que se passe-t-il si le fou est unique ? Lewis propose, pour définir la douleur, de comparer ses états mentaux, et non plus son comportement, à ceux des membres de son espèce[40] (les humains, qu'ils soient fou ou non). Le test de Turing, héritant de la méthode behaviouriste, ne permet pas de faire cette comparaison puisque les états mentaux sont inaccessibles. Dans le cas de machines électroniques, on peut cependant comparer le fonctionnement interne des fous avec ceux de son espèce, i.e. comparer les descriptions des états mentaux en termes causaux.

Que se passe-t-il si le fou appartient à une autre espèce dont-il est l'unique représentant (e.g. le dernier des martiens, qui s'avère être un fou). Devant un cas si étrange et complexe, Lewis abandonne. « I think we cannot and need not solve this problem. »[41] Le test de Turing, de la même manière, ne peut identifier toutes les formes d'intelligence ou d'esprit. Ainsi, il se pourrait que les membres d'une espèce aient un esprit, des expériences conscientes, une certaine intelligence, sans qu'on ne puisse jamais les détecter parce qu'on ne possède aucun individu de référence dont on stipule *a priori* qu'il a un esprit. Puisque nous n'avons que peu d'exemples de consciences (humaines ou animales), nous ne pouvons tester que celles-là sur l'ensemble des machines.

Pour conclure, le test de Turing est fondamentalement réservé à la détection des compétences que nous reconnaissons *a priori*. Il sera donc utilisé dans ce cadre uniquement, pour déterminer si une machine a un esprit proche de celui de l'homme, ou éventuellement

---

[40] Cf. (Lewis, 1978) Section V

[41] « Je pense que nous ne pouvons ni ne devons résoudre ce problème. » (Lewis, 1978) Section VII

proche de celui d'autres animaux[42], en comparant son comportement au leur[43]. La conclusion du test, enfin, dépend des compétences que nous attribuons *de fait* aux individus de référence. L'intérêt du test en ce qui concerne les programmes et les robots repose donc sur les recherches en éthologie cognitive.

3.2.2 Des zombies passant le test

Un *zombie* est une machine qui passe avec succès le test de Turing alors qu'elle n'est pas intelligente. Si nous utilisons le test de Turing pour définir l'intelligence – à savoir : une machine qui passe avec succès le test de Turing est intelligente – les zombies n'existent pas ! Et pourtant, parce que nous utilisons également la définition du sens fonctionnel, nous pouvons facilement imaginer qu'il en existe.

Le *zombie chanceux* est un exemple de zombie qui, à notre connaissance, n'a jamais été présenté dans la littérature. Ceci est d'autant plus étonnant qu'il constitue la preuve la plus simple et efficace que le test de Turing peut être défectueux (en ce qui concerne sa *spécificité*), mais c'est également un exemple assez improbable. Imaginons un homme qui fonctionne de la manière suivante : lorsqu'on lui soumet une question, il choisit des lettres aléatoirement (le nombre de ces lettres étant lui-même aléatoirement grand) et constitue avec elles des mots de manière aléatoire. La plupart du temps, sa réponse, composée de mots inexistants dans la langue du test, est tout simplement illisible. Parfois, les mots composés aléatoirement sont tous bien formés sur le plan lexical, mais la syntaxe de la phrase ne correspond à rien qui existe dans la langue de référence. D'autres fois, la syntaxe est bonne, mais la réponse n'a pas de sens en soi[44], ou alors pas de sens vis-à-vis de la question posée par le juge et vis-à-vis du contexte de la discussion. Enfin, dans certains cas assez rares, la réponse formulée aléatoirement par l'homme est une réponse tout à fait correcte sur le plan lexical, syntaxique et sémantique. Malgré ces lueurs d'intelligence, l'homme échoue très fréquemment au test de Turing, les observateurs jugeant qu'il se comporte assez bizarrement et que ses réponses aux questions du test sont loin d'être intelligentes.

---

[42] Nous pouvons nous référer à des travaux d'éthologie cognitive pour préciser la notion de conscience animale, par exemple aux travaux concernant l'émergence de la conscience dans le cadre de l'évolution naturelle, cf. (Denton, 1993)

[43] Les animaux ne possédant l'usage du langage, le test de Turing doit bien évidemment être modifié. Les compétences évaluées par ce test dialogique doivent être adapté au type de conscience que nous voulons détecter.

[44] Comme par exemple la célèbre phrase grammaticalement correcte de Noam Chomsky : « Colorless green ideas sleep furiously. » (« Les idées vertes sans couleur dorment furieusement. ») (Chomsky, 1957)

Imaginons dans un second temps que cet homme soit extrêmement chanceux. Imaginons qu'il soit tellement chanceux que, lors d'un test, toutes ses phrases sont grammaticalement correctes, qu'elles ont du sens et qu'elles sont même très à propos. L'observateur conclu donc que son interlocuteur est intelligent. Un tel cas, bien que très improbable, est *possible* pour tout test en temps fini. Il est même possible que l'homme soit régulièrement confronté à des tests, par des observateurs différents, et qu'il ne choisisse jamais une seule lettre de travers. Personne ne pourra alors révéler que cet homme, en réalité, n'est qu'un idiot avec beaucoup de chance.

Ce cas est problématique puisque, selon les critères comportementaux, l'homme *est* véritablement intelligent. Le test de Turing peut donc intégrer dans sa définition extensive de l'intelligence une machine pour laquelle, lorsqu'on regarde de plus près, nous n'avons pas envie de dire qu'elle l'est effectivement. Il s'agit encore une fois d'une tension entre le sens fonctionnel *a priori* des fonctionnements intelligents et le résultat *a posteriori* du test de Turing. Dans le cas de la conscience, nous ne parlons pas d'*idiot* mais de *zombie*.

Nous avons commencé à parler de *zombies* dans la section 2.1.1 à propos des *machines artificielles* de Descartes. Les zombies sont ainsi définis comme des créatures physiquement identiques aux hommes. Ils se comportent exactement comme nous, mais sont pourtant inconscients. Imaginés à l'origine par Robert Kirk pour s'opposer au matérialisme[45], la question de l'existence des zombies et de l'implication d'une telle existence ont été de nombreuses fois discutées[46]. Nous ne reviendrons pas dessus, d'autant plus que la problématique sous-tendue est de même nature que celle mise en perspective par le problème des autres esprits. Par ailleurs, dans un contexte physicaliste, la notion de zombie telle que développée par Robert Kirk est incohérente. En effet, l'idée d'un clone structurel *et* fonctionnel[47] (même composition physique et même fonctionnement) ne peut être envisagée sans conscience, dans la mesure où celle-ci est entièrement déterminée par ces deux caractéristiques dans le cadre du physicalisme. L'*argument de la connaissance* présenté par Jackson[48] et l'objection de Nagel au physicalisme[49] nous semblent bien plus soutenables dans la mesure où ils s'appuient sur l'expérience consciente pour commencer leur raisonnement

[45] Voir les *zombies philosophiques* (*philosophical zombies*) dans (Kirk & Squires, 1974)
[46] Cf. (Chalmers D. J., 1996) Section II.3, *Can Consciousness Be Reductively Explained?* pour une discussion en profondeur sur la notion de zombies et des conséquences concernant l'irréductibilité du caractère phénoménal de la conscience.
[47] Cf. section 1.1.2 et l'exemple du clone dans (Harnad, Can a machine be conscious? How?, 2003)
[48] Cf. (Jackson, 1982) Section *The Knowledge Argument for Qualia*
[49] Cf. (Nagel, What Is It Like to Be a Bat?, 1974)

(point de vue subjectif) et non sur l'observation objective d'individu dont on présuppose l'inconscience.

Dans ce mémoire, les zombies désignent des machines qui peuvent différer de nous sur le plan structurel et sur le plan fonctionnel. Ils ne sont identiques à nous que sur le plan comportemental, ce qui explique leur capacité à passer le test de Turing et à contrarier le sens fonctionnel.

3.2.3 Résumé des contradictions

- Fou (faux négatif)
    - Sens fonctionnel : fonctionnement *a priori* intelligent
      → **Machine intelligente**
    - Test de Turing : comportement *a priori* non-intelligent
      → Fonctionnement *a posteriori* non-intelligent
      → **Machine non-intelligente**
      → Contradiction !
- Zombie (faux positif)
    - Sens fonctionnel : fonctionnement *a priori* non-intelligent
      → **Machine non-intelligente**
    - Test de Turing : comportement *a priori* intelligent
      → Fonctionnement *a posteriori* intelligent
      → **Machine intelligente**
      → Contradiction !

3.3 Comment régler les tensions ?

Les cas de fous et de zombies sont engendrés par une tension entre le sens que l'on donne *a priori* à l'intelligence et la volonté de définir empiriquement cette notion d'intelligence. Le test de Turing ne coïncide pas avec le sens fonctionnel. Il existe deux manières de corriger cette différence entre les deux définitions : (1) modifier le *test de Turing* ou (2) modifier le *sens fonctionnel*. Les sections suivantes expliquent les conséquences de tels choix sur nos conceptions des fonctionnements intelligents et des comportements intelligents.

3.3.1 Que faire des fous ?

Un fou est une machine intelligente qui échoue au test de Turing. Selon le sens fonctionnel, le cerveau du fou fonctionne intelligemment et donc son comportement est par définition intelligent. Selon le test de Turing, son comportement observé ne correspond à aucun comportement intelligent et donc son fonctionnement est par définition non-intelligent. Il est possible, pour démêler cette contradiction, (1) de changer le test de Turing, i.e. les *comportements* intelligents *a priori* ou (2) de changer le sens fonctionnel, i.e. les *fonctionnements* intelligents *a priori*. Les cas (1) et (2) sont représentés par les flèches (1) et (2) dans le diagramme de Carroll suivant.

	Test de Turing positif	**Test de Turing négatif**
Intelligence selon le sens fonctionnel	Machines intelligentes	(1) Fous (2)
Absence d'intelligence selon le sens fonctionnel	Zombies	Machines non-intelligentes

Dans le cas (1), nous parlons de *vrai fou*, c'est-à-dire que son fonctionnement est effectivement intelligent mais que, comme le fou de Lewis[50], il n'est pas causalement lié à un comportement intelligent. Le vrai fou est donc une *machine intelligente*. Il faut alors changer le test de Turing pour l'adapter à ce cas particulier et inclure dans sa définition de l'intelligence le comportement du fou. Cette transformation du test consiste à ajouter une *condition suffisante* à la définition *a priori* des comportements intelligents. Cela revient en fait à comparer le fou à un individu de son espèce. On dira alors : « cette machine est intelligente si elle passe avec succès le test de Turing par rapport à cette machine de référence, *ou alors* si elle passe le test par rapport à cette autre machine de référence. » La disjonction de plusieurs conditions suffisantes permet ainsi d'améliorer le test pour y inclure les cas de folie.

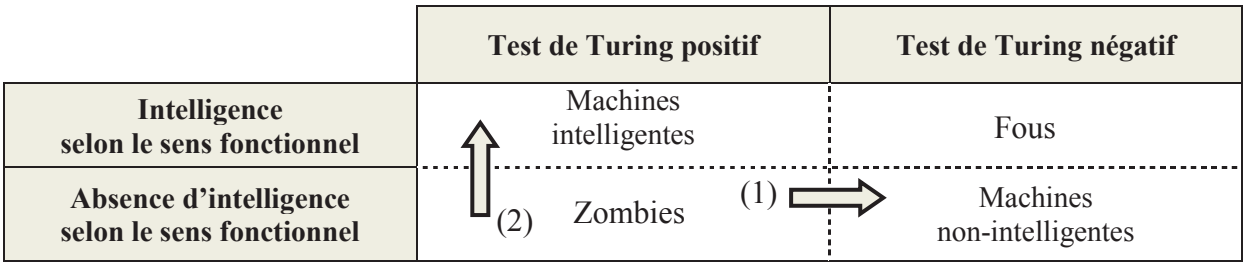
Dans le cas d'un fou pour lequel aucune machine de référence ne peut être définie (cf. le cas du dernier martien fou, section 3.2.1) la condition suffisante ajoutée au test de Turing peu concerner les fonctionnements. Le test de Turing devient une définition mixte, contraignant à la fois les comportements et les fonctionnements des machines. On dira : « cette machine est intelligente si elle passe avec succès le test de Turing, *ou alors* si elle a tel fonctionnement interne. »

[50] Cf. (Lewis, 1978) Section I

Dans le cas (2), nous parlons de *faux fou*, c'est-à-dire que le sens fonctionnel se trompe à propos de son cas : son fonctionnement, contrairement à ce que l'on croit, n'est pas intelligent, ce qui explique pourquoi son comportement ne l'est pas non plus. Le faux fou est une *machine non-intelligente*. Cette nouvelle perspective nous invite à revoir la définition donnée par le sens fonctionnel. Sa modification consiste en l'ajout d'une *condition nécessaire* à la définition *a priori* des fonctionnements intelligents. Il faut trouver une condition qui supprime les comportements non-intelligents responsables de l'échec de la machine au test de Turing.

3.3.2 Que faire des zombies ?

Un zombie est une machine qui n'est pas intelligente mais qui réussit pourtant à passer le test de Turing. Selon le sens fonctionnel, le zombie a un fonctionnement non-intelligent et donc, par définition, son comportement n'est pas intelligent. Selon le test de Turing, son comportement observé correspond à un comportement intelligent et donc son fonctionnement est par définition intelligent. Il est possible, pour démêler cette autre contradiction (1) de changer le test de Turing, i.e. les *comportements* intelligents *a priori* ou (2) de changer le sens fonctionnel, i.e. les *fonctionnements* intelligents *a priori*. Les cas (1) et (2) sont représentés par les flèches (1) et (2) dans le diagramme de Carroll suivant.



Dans le cas (1), nous parlons de *vrai zombie*, c'est-à-dire que son comportement est effectivement intelligent mais que, comme dans le cas du zombie chanceux, il n'est pas le fait d'un fonctionnement intelligent. Le vrai zombie est une *machine non-intelligente*. Il faut alors changer le test de Turing pour l'adapter à ce cas particulier et qu'il détecte le zombie. Cette transformation du test consiste à ajouter une *condition nécessaire* à la définition *a priori* des comportements intelligents. Cette condition nécessaire peut, néanmoins, porter sur le fonctionnement de la machine. On dira alors : « cette machine est intelligente si elle passe avec succès le test de Turing, *et* si elle fonctionne de telle manière. » La conjonction de

plusieurs conditions nécessaire permet ainsi d'améliorer le test et d'éliminer les cas de zombie.

Dans le cas (2), nous parlons de *faux zombie*, c'est-à-dire que le sens fonctionnel se trompe à propos de son cas : son fonctionnement, contrairement à ce que l'on en pense, est bel et bien intelligent, ce qui explique pourquoi il passe avec succès le test de Turing. Le faux zombie est une *machine intelligente*. Cette nouvelle perspective nous invite à revoir notre définition donnée par le sens fonctionnel. Sa modification consiste en l'ajout d'une *condition suffisante* à la définition *a priori* des fonctionnements intelligents. Il faut trouver une condition qui est responsable des comportements intelligents qui ont menés à la réussite du test.

3.4 Ce que peut nous apprendre le test

Lorsque les définitions en extension réalisées par le test de Turing et le sens fonctionnel ne coïncident pas, il faut modifier l'une ou l'autre. Plus que d'ajuster ainsi des définitions, la régularisation des cas de fous et des cas de zombies a d'autres conséquences intéressantes. Elle permet (1) d'invalider un modèle de l'esprit responsable d'incohérences ou (2) de découvrir et concevoir des nouvelles formes de fonctionnements et de comportements associés à la notion d'esprit.

3.4.1 Modifier des théories existantes

Une théorie de l'esprit donne un modèle général à la notion d'esprit. Il est possible de concevoir, à partir de ce modèle, des machines par particulières que nous soumettons au test de Turing. Si une théorie permet de définir un zombie, cela signifie qu'elle présente une contradiction entre la définition *a posteriori* de l'esprit selon le test et la définition *a priori* de l'esprit selon le sens fonctionnel. Lorsque nous adoptons la position (1) du *vrai zombie*, une telle théorie est dite *incohérente* et le modèle qu'elle présente n'est pas un modèle valide. La condition nécessaire alors ajoutée au test de Turing pour régler le cas de *vrai zombie* met en évidence l'erreur commise par la théorie. Celle-ci suppose que cette condition *n'est pas* nécessaire à la notion d'esprit, alors qu'elle l'est. Le modèle peut être remanié afin de prendre en compte cette condition et donner ainsi naissance à un nouveau modèle, une nouvelle théorie de l'esprit.

Il en est de même pour les cas de *faux fous*. Une théorie permettant de concevoir une machine que nous pensons dotée d'un esprit, mais qui en fait n'en a pas selon le sens

fonctionnel, oublie dans son modèle une condition nécessaire à la notion d'esprit. Comme précédemment, la théorie peut être remaniée pour inclure cette condition et donner naissance à une théorie cohérente.

A ce titre, l'utilisation conjointe des définitions extensives de l'esprit permettent d'évaluer les théories en fonction de leur cohérence. Le test de Turing devient un outil épistémologique participant à la valse des théories. La partie 4 de ce mémoire présente de tels bouleversements historiques dans les conceptions de l'esprit ayant été développées au cours du XX^{e} siècle.

3.4.2 Découvrir de nouvelles formes d'esprit

Les cas de *faux zombies* incitent à modifier le sens fonctionnel concernant les fonctionnements *a priori* intelligents. Ils ne témoignent pas d'une incohérence au sein des théories permettant de les exprimer, mais d'une mauvaise conception en termes de définition extensive. La condition suffisante, ajouté au sens fonctionnel, confirme la théorie : il *suffit* d'avoir ce fonctionnement particulier pour avoir un esprit, et modifie notre définition des fonctionnements associés à la notion esprit. En d'autres termes, les cas de *faux zombies* nous permettent de découvrir de nouveaux fonctionnements.

Il en est de même dans le cas des *vrais fous*. La théorie n'est pas mis-à-mal et de nouveaux genres de comportements sont découverts, comportements associé à la présence d'un esprit. A ce titre, le test de Turing permet de modifier nos définitions extensives de l'esprit et d'y inclure de nouveaux fonctionnements et de nouveaux comportements. La partie 4 présente également de telles découvertes.

3.4.3 Bilan

Lors d'un cas de *fou* ou de *zombie*, il faut :

1. Choisir de conserver le test de Turing et de modifier le sens fonctionnel, ou bien de faire l'inverse.
2. Ce choix implique de reconsidérer la théorie de l'esprit (définition intensive) dans le cadre de laquelle le cas est apparu, ou bien de reconsidérer les définitions extensives à l'origine du cas problématique.

Nous résumons les différents cas par le tableau suivant.

Vrai zombie	Condition nécessaire	Sur le test de Turing	Modification de la théorie
Faux zombie	Condition suffisante	Sur le sens fonctionnel	Nouveau fonctionnement
Vrai fou	Condition suffisante	Sur le test de Turing	Nouveau comportement
Faux fou	Condition nécessaire	Sur le sens fonctionnel	Modification de la théorie

Enfin, les prises de position pour régulariser les cas de *fou* ou de *zombie* ne peuvent pas être accompagnées d'arguments ontologiques. Le problème des autres esprits (cf. section 2.1) a en effet révélé que la nature même de l'esprit des autres machines ne peut être directement connu. Les prises de position sont alors l'origine de débats définitionnels en faveur ou en défaveur des théories de l'esprit. Ils permettent, en outre, de dégager clairement les conséquences de ces modèles sur le cas particulier d'esprits que nous pouvons rencontrer. Défendre telle ou telle théories implique que l'on accepte tel ou tel cas de *vrai fou* et de *faux zombie*, tel ou tel type de fonctionnements et de comportements associés à la notion d'esprit. Au contraire, refuser ces types de fonctionnements et ces types de comportements à la définition que nous souhaitons construire consiste à invalider la théorie et invite à concevoir un nouveau modèle.

4 La valse des théories

Cette partie fait l'exposé chronologique de discussions qui ont animé la philosophie de l'esprit, les sciences cognitives et l'Intelligence Artificielle au cours du XX^{e} siècle. Ces débats concernent la validité de différents modèles de l'esprit. Nous proposons une relecture des arguments exposés au cours de ces discussions en termes de *fous* et de *zombies*. Les outils épistémologiques développés dans la partie précédente sont ainsi appliqués à la valse historique des théories.

Les notions de fou et de zombie, le sens fonctionnel et le test de Turing ne sont pas utilisés ici pour définir l'intelligence – comme cela était le cas dans la partie précédente – mais pour définir la notion d'esprit. Nous parlons alors de *fonctionnements dotés d'un esprit* pour désigner les fonctionnements des machines que l'on définit *a priori* comme ayant un esprit (sens fonctionnel). Nous parlons également de *comportements dotés d'un esprit* pour désigner les comportements des machines que l'on définit *a priori* comme ayant un esprit (test de Turing).

Nous présentons dans cette partie deux cas de zombies (cf. sections 4.1 et 4.3) et un cas de fou (cf. section 4.2). Chaque cas est abordé de la manière suivante :

1. Présentation du modèle de l'esprit soumis au débat
2. Cas de fou ou de zombie exprimé par le modèle
3. Les deux positions permettant de régler le cas problématique
4. Choix historique et conséquence pour la valse des théories
5. Bilan

4.1 Premier cas : le zombie de la chambre chinoise

L'expérience de la chambre chinoise, imaginée par John Searle[51], est sans doute l'argument le plus connu s'opposant au courant majeur défendu par les sciences cognitives lors des années 1950 à 1980. Il s'agit du *cognitivisme*. L'expérience de pensée décrit un zombie dont le fonctionnement est en accord avec le modèle cognitiviste de l'esprit (cf. sections suivantes). L'interprétation de ce cas comme un cas de *vrai zombie* a historiquement conduit à l'invalidation de ce modèle en ce qui concerne une propriété importante de l'esprit : l'*intentionnalité*.

[51] Cf. (Searle, 1980)

4.1.1 Modèle : cognitivisme et symbolisme

Le cognitivisme a constitué, dès les années 1950, le courant dominant des sciences cognitives. Nous verrons plus tard que des théories alors minoritaires sont aujourd'hui exploitées pour dépasser les limites de ce premier modèle de l'esprit. Quel est-il ? L'hypothèse principale du cognitivisme consiste à affirmer que l'esprit est un *système symbolique* et que la cognition est une manipulation de symboles, un calcul (ou *computation*)[52].

Un système symbolique est constitué d'un ensemble de symboles chacun associé à une représentation physique arbitraire (*physical tokens*) et de règles de calcul explicites (*explicit rules*). Ces règles permettent de manipuler les symboles en fonction de leur *forme* seulement, non pas en fonction de leur *sens*. Les symbolistes (Fodor et Pylyshyn pour ne citer qu'eux) soutiennent que le niveau symbolique est indépendant de son implémentation (i.e. indépendante de sa réalisation physique). Cette notion d'*indépendance implémentationnelle* hérite des notions d'indépendance en informatique entre le code d'un programme et le processeur qui l'exécute. En conséquence, pour le cognitivisme, l'esprit humain et l'ensemble de ses facultés cognitives sont indépendants de son support matériel, à savoir le cerveau. L'analogie suivante est de mise : *le cerveau est à l'esprit ce que l'ordinateur est au programme*. L'indépendance, dans le génie logiciel, entre le programme et le processeur qui l'exécute soutient, selon le cognitivisme, l'indépendance entre l'esprit et le cerveau qui l'exécute.

Une propriété importante des systèmes symboliques est qu'ils sont sémantiquement interprétables. « The syntax can be systematically assigned a meaning. »[53] Le sens des symboles est ainsi assuré par des règles d'interprétation. L'argument de Searle repose sur le fait que ces règles sont *extérieures* au système, comme l'illustre son expérience, présentée dans la section suivante.

---

[52] Cf. (Harnad, The Symbol Grounding Problem, 1990) pour une présentation détaillée de l'hypothèse cognitiviste et une définition complète des systèmes symboliques.

[53] « La syntaxe peut systématiquement être assigné à une sens. » (Harnad, The Symbol Grounding Problem, 1990)

4.1.2 Cas de zombie : la chambre chinoise

L'exemple le plus célèbre de zombie passant le test de Turing est présenté dans l'expérience de pensée connue sous le nom de *la chambre chinoise*[54]. John Searle y formule un argument puissant contre le *cognitivisme*. Selon le philosophe, une capacité cognitive fondamentale fait défaut aux systèmes symboliques : l'*intentionnalité*. De ce fait, ils ne peuvent modéliser adéquatement l'esprit humain. La chambre chinoise permet d'illustrer son argument.

Dans cette expérience de pensée, Searle imagine qu'il passe lui-même le test de Turing en chinois (langue qu'il ne maîtrise pas) à l'aide d'un dictionnaire associant des suites de symboles – les questions du juge – à d'autres suites de symboles – les réponses de Searle – et constituant ainsi un « programme » capable de passer avec succès le test. L'argument est le suivant : grâce aux règles de calcul indiquées par le dictionnaire, Searle est capable de faire croire à un observateur extérieur qu'il parle couramment chinois, alors que ce n'est absolument pas le cas. En ce sens, Searle est ici un *zombie*.

L'idée sous-tendue par cet argument est que la manipulation symbolique – ici réalisée à l'aide du dictionnaire – puisqu'elle est effectuée en vertu de la seule *forme* des symboles, ne permet pas de saisir leur *signification*. L'interprétation sémantique associée au système symbolique est le fait de l'observateur extérieur (le juge) et du concepteur du dictionnaire, mais Searle (le zombie) est incapable de comprendre les caractères qu'il manipule et les phrases qu'il produit. C'est ainsi que fonctionne un processeur : il manipule des symboles en vertu de leur forme physique (des successions de bits) selon un code dont l'interprétation est le fait de son concepteur et de l'utilisateur du programme.

Daniel Dennett formule plusieurs objections à l'expérience de la chambre chinoise. La première concerne la forme du programme implémenté par le dictionnaire de Searle. Le fait qu'il permette au zombie de passer le test de Turing repose sur une forte supposition, « the (unwarranted) supposition that the giant program would work by somehow simply "matching up" the input Chinese characters with some output Chinese characters. » Dennett ajoute : « No

[54] Cf. (Searle, 1980) pour une présentation complète de l'expérience, des objections qui y ont été opposées et des réponses apportées par l'auteur.

such program would work, of course—do Chinese Room's speeches in English "match up" with the judge's questions ? »[55]

Dennett a sans doute raison de rappeler que, s'il passait le test de Turing, il ne répondrait pas en faisant simplement coïncider aux questions du juge des réponses qu'il aurait préparées. L'esprit humain fonctionne d'une manière bien plus complexe. Cependant, nous pensons qu'un tel *matcher* est réalisable en théorie et que, contrairement à l'intuition de Dennett, il pourrait effectivement réussir le test de Turing. Supposons que le juge soit limité à M mots (ou M caractères en ce qui concerne le chinois) pour chacune de ses questions. Le nombre des questions possibles est donc fini[56]. Nous pouvons donc associer à chacune de ces questions une réponse préenregistrée dans le dictionnaire. Cependant, certaines réponses doivent être replacées dans le contexte de la discussion. Il faut alors prendre en compte toutes les questions précédemment formulées par le juge. Supposons qu'il soit limité à Q questions lors du test. Le nombre de questions possibles, en prenant en compte le contexte de la conversation, est également fini[57]. Par conséquence, on peut préparer une réponse pour chacune de ces questions et ainsi construire un *matcher* que le test de Turing identifierait comme intelligent. Il est nécessaire de remarquer enfin que les contraintes sur le nombre de mots et le nombre de questions ne limitent pas le test de Turing. En effet, le test est toujours effectué en temps fini, éventuellement sur le temps d'une vie. Or, un temps fini implique nécessairement un nombre fini de questions, elles-mêmes de longueurs finies. Il est nécessaire d'ajuster les paramètres M et Q du *matcher* pour les adapter à la longueur maximale d'un tel test.

Sans doute qu'un tel dictionnaire n'est pas réalisable en pratique. Il contiendrait en effet, pour un maximum de 100 questions par test (questions limitées à 100 caractères chacune), un ordre de $10^{50\,000}$ entrées en ce qui concerne la langue chinoise[58]. Il est inconcevable que l'homme arrive un jour à réaliser un tel dictionnaire et à le stocker sur assez

---

[55] « La supposition (non-garantie) que le programme géant fonctionnerait en faisant simplement "coïncider" les caractères chinois en entrée avec des caractères chinois en sortie. » « Un tel programme ne pourrait fonctionner bien évidemment – est-ce que les discours en anglais de la Chambre Chinoise "coïncident" avec les questions du juge ? » (Dennett, 1991) Section 1, *Imagining a Conscious Robot*

[56] Le nombre de questions possibles est inférieur à N^M, où N est le nombre fini de mots constituant le vocabulaire de la langue utilisée pour le test.

[57] Le nombre de questions possibles, prenant en compte le contexte de la discussion, est inférieur à N^{MQ}.

[58] *Hànyǔ dà Zìdiǎn* (littéralement « le dictionnaire complet de caractères chinois ») inclut près de 55 000 entrées. Nous avons donc un ordre de grandeur de $(10^5)^{100\times100} = 10^{50\,000}$ entrées pour notre propre dictionnaire. Cet ordre est valable si on prend en compte toutes les suites de 100 caractères possibles, même mal constituées, et toutes les suites de 100 questions, même s'il s'agit de 100 fois la même question. Le nombre de conversation correcte et « normale » est sans doute inférieur à cet ordre de grandeur.

de papier dans la mesure où le nombre d'atomes dans l'univers est lui-même inférieur au nombre d'entrées dans notre dictionnaire[59]. Mais le simple fait qu'un tel dictionnaire soit *en théorie* réalisable pose problème. En effet, il est conforme au modèle cognitiviste s'appuyant sur la manipulation symbolique pour décrire la cognition. Selon cette théorie, donc, le zombie de la chambre chinoise a un esprit. Or, dans la mesure où il ne comprend rien à ce qu'il fait, dans la mesure où il ne possède pas la signification de ses propres réponses, nous aimerions dire qu'il n'en a pas.

Avant d'accuser trop rapidement ce zombie, donnons les raisons pour lesquels son fonctionnement n'est pas reconnu par le sens fonctionnel de la cognition et de la conscience. Comme Searle le souligne, ce qui fait défaut à cette machine est l'*intentionnalité*. Les caractères qu'elle manipule sont interprétés par l'observateur et, de ce point de vue, elle les utilise en *référence à quelque chose*. Lorsqu'on lui demande la définition d'un mot, elle est capable de la donner et même d'en discuter. En outre, on a l'impression que la machine *sait* à quoi ces mots font référence et qu'elle ne les utilise pas arbitrairement. Cependant, du point de vue de la machine, elle ne sait pas à quoi correspondent les symboles qu'elle utilise arbitrairement. L'intentionnalité apparente de la machine prend source dans l'intentionnalité de son concepteur.

Le problème de l'origine de la sémantique des symboles est aussi connu sous le nom de *Symbol Grounding Problem*[60]. Stevan Harnad formule ce problème en définissant l'intentionnalité comme la capacité d'attribuer un sens à des symboles *de manière autonome*[61]. Pour cette raison, la notion de *système symbolique* ne peut être un bon modèle de l'esprit intentionnel : la syntaxe à elle seule ne suffit pas à déterminer la sémantique. « Cognition cannot be just symbol manipulation. »[62]

Harnad rappelle dans son article la distinction élaborée par Ludwig Wittgenstein dans ses *Philosophical Investigastions* à propos des règles *explicites* et *implicites*. « It is not the same thing to "follow" a rule (explicitly) and merely to behave "in accordance with" a rule (implicitly). »[63] Le *calcul symbolique* est défini en fonction de règles de manipulation *explicites* que les processus cognitifs appliquent systématiquement, de manière hétéronome.

---

[59] Le nombre d'atomes dans l'univers est couramment estimé à 10^{80}.
[60] « Problème de l'ancrage des symboles »
[61] Cf. (Harnad, The Symbol Grounding Problem, 1990) Section *The Chinese Room*
[62] « La cognition ne peut pas être simplement une manipulation de symboles. » Ibid.
[63] « Ce n'est pas la même chose que de "suivre" une règle (explicite) que de se comporter simplement "conformément à" une règle (implicite). » (Harnad, The Symbol Grounding Problem, 1990) Section *Symbol Systems*

L'*interprétation sémantique*, au contraire, doit être réalisée selon des règles *implicites*. Elles ne sont pas déterminées *une fois pour toute*, mais résultent d'une disposition globale du cerveau à se donner ses propres règles d'interprétation.

4.1.3 Vrai zombie ou faux zombie ?

Deux positions peuvent être défendues à l'égard du zombie de la chambre chinoise : (1) il s'agit d'un *vrai zombie* ou (2) il s'agit d'un *faux zombie*. Cette section explicite les deux positions et leurs conséquences.

La position (1) consiste à affirmer que le zombie de la chambre chinoise n'est pas capable d'intentionnalité mais qu'il se comporte néanmoins comme s'il en était capable. Nous choisissons alors de conserver le sens fonctionnel selon lequel *être capable d'intentionnalité* est une condition nécessaire à la possession d'un esprit. Le test de Turing doit en conséquence être modifié pour coïncider avec cette définition.

Il est impossible d'ajouter une *condition nécessaire* au comportement du zombie sans réduire l'efficacité du test en ce qui concerne les *vrais positifs*. En effet, si l'on restreint les comportements *a priori*, les machines intentionnelles qui se comportent comme le zombie risquent d'échouer au test, engendrant un *faux négatif*. La *condition nécessaire* doit alors contraindre les fonctionnements possibles des machines et non leurs comportements. On construit une définition mixte reposant sur les comportements (via le test) et sur les fonctionnements (via la condition que nous lui ajoutons). Comment formuler plus exactement cette condition ? Nous pouvons dire « cette machine a un esprit si, et seulement si, elle passe le test de Turing et est capable d'intentionnalité. » La condition supplémentaire est alors définie en termes de caractéristiques fonctionnelles.

Le zombie de la chambre chinoise témoigne d'une incohérence dans le modèle cognitiviste de l'esprit. Puisqu'il est « permis » par le modèle et qu'il présente une contradiction entre les définitions extensives de l'esprit, il invalide la théorie. La *condition nécessaire* contribue alors à modifier les hypothèses cognitivistes et symbolistes. Elle empêche que l'on exprime, dans le cadre de la nouvelle théorie, un zombie similaire à celui qui nous a permis d'invalider l'ancienne théorie. La condition nécessaire historiquement formulée par les sciences cognitives et le modèle auquel elle a donné naissance sont décrits dans la section 4.1.4.

La position (2) consiste à affirmer que le zombie de la chambre chinoise, malgré le sens fonctionnel, possède bel et bien un esprit. Il n'est donc pas étonnant qu'il puisse passer le test de Turing avec succès. Le sens fonctionnel est alors modifié au profit du test de Turing qui est préservé. La modification est formulée en termes de condition suffisante à la définition de l'esprit.

Cette seconde position peut être défendue de deux manières différentes. Soit on considère que le fonctionnement du zombie suffit à l'intentionnalité, que celle-ci émerge de l'utilisation du dictionnaire et que le *Symbole Grounding Problem* est résolu par la complexité vertigineuse et le nombre de ses entrées ; soit on considère que l'intentionnalité n'est pas nécessaire aux fonctionnements dotés d'un esprit – définis de manière *a priori* pas le sens fonctionnel. L'utilisation d'un tel dictionnaire est alors une *condition suffisante* à la possession d'un esprit. Dans les deux cas, on définit de nouvelles formes d'intentionnalité ou d'esprit. Le zombie est naturalisé au rang de *machine intentionnelle* ou, plus largement, les machines non-intentionnelles sont naturalisées au rang de *machine avec un esprit*. L'hypothèse cognitiviste est conservée et elle affirme l'existence de nouveaux types de machines dotées d'un esprit.

4.1.4 Choix historique

Selon la tradition philosophique, la notion d'intentionnalité est importante pour caractériser l'esprit humain. Elle correspond, comme le remarque Harnad, à une *autonomie de l'esprit*, sa capacité à se forger des règles implicites, notamment vis-à-vis de l'interprétation des symboles qu'il utilise. L'intentionnalité entre donc dans la définition *a priori* du fonctionnement des esprits humains, donnée par le sens fonctionnel. D'autres espèces animales, toujours selon le sens fonctionnel, ne sont pas douées d'intentionnalité, entre autres parce que cette capacité nécessite une conscience réflexive. L'expérience de Searle argumente donc contre un modèle cognitiviste de l'esprit humain uniquement.

La position (2) consiste à aller contre le sens fonctionnel et à nier le caractère nécessaire de l'intentionnalité pour avoir un esprit similaire à celui des hommes. Pour revenir au problème des autres esprits (cf. section 2.1), il est possible que les hommes que j'observe n'aient pas cette capacité et qu'ils soient tous à ce titre des zombies : d'où l'idée d'une définition de l'esprit humain n'incluant pas cette fonction. Cependant, dans la mesure où le sens fonctionnel est défini par un sujet intentionnel, il est étonnant de ne pas prendre en compte cette compétence dans la description *a priori* des fonctionnements.

La seconde option de la position (2) est de soutenir que l'intentionnalité *émerge* de la syntaxe explicitée par les systèmes symboliques : une manipulation formelle, suffisamment complexe, permet de *faire émerger* la sémantique des symboles. Le cas de la chambre chinoise, cependant, nous incite à rejeter cette option. Le *matcher* qui y est représenté est un système symbolique *linéaire* pour lequel le dictionnaire est égal à la somme de ses entrées : il n'y a rien de plus dans sa structure, aucune sémantique nouvelle ne peut émerger de ce système. La position (2) est ainsi difficile à défendre.

L'idée d'émergence, cependant, a fait son chemin dans l'histoire des sciences cognitives. Si on ne peut sauver le cognitivisme en conservant tel quel le test de Turing, il est possible de concevoir un modèle plus fin de la cognition qui ne repose pas seulement sur la manipulation de symboles. L'objection de Dennett contre l'exemple simpliste de Searle indique comment dépasser le problème du cognitivisme. « Complexity does matter. »[64] L'étude des systèmes *non-linéaires*, pour lesquels le tout est plus que la somme des parties, a conduit à décrire l'activité symbolique comme émergeant d'une activité de plus bas niveau, parfois nommé niveau *sub-symbolique*. Ce niveau est composé d'unités autonomes réactives, c'est-à-dire que ces unités ne manipulent pas des symboles mais changent d'état en fonction des interconnexions qu'elles entretiennent avec les autres unités. La dynamique globale du réseau est le support de fonctions cognitives complexes comme l'identification et la discrimination de formes ou la manipulation de symboles et de catégories. Selon ce modèle de l'esprit, appelé *connexionnisme*, « cognition is not symbol manipulation but dynamic patterns of activity in a multilayered network of nodes or units. »[65]

De très nombreux travaux ont été menés en Intelligence Artificielle pour implémenter de tels réseaux. Ce sont par exemple les célèbres *réseaux de neurones*[66]. Des études plus larges donnent naissance au *calcul distribué* et à l'*Intelligence Artificielle distribuée*. Le domaine des *systèmes multi-agents* repose également sur cette conception de l'esprit par émergence de l'activité d'agents autonomes[67].

---

[64] « La complexité est à prendre en compte. » (Dennett, 1991) Section 1, *Imagining a Conscious Robot*

[65] « La cognition ne repose pas sur une manipulation de symboles, mais sur des structures dynamiques en activité dans un réseau constitué de plusieurs couches de nœuds ou d'unités. » (Harnad, The Symbol Grounding Problem, 1990)

[66] Cf. (Lettvin, Maturana, McCulloch, & Pitts, 1959) pour les premiers travaux sur les réseaux de neurones formels, issus de la biologie, permettant théoriquement de réaliser des fonctions logiques et symboliques complexes.

[67] Cf. (Minsky, 1988) pour l'origine conceptuelle des systèmes multi-agents et (Wooldridge, 2009) pour une introduction plus moderne au domaine. Cf. également (Varela, Thompson, & Rosch, 1993) et (Varela F. J.,

L'apport important de ces disciplines issues du connexionnisme – en ce qui concerne le dépassement du cognitivisme – réside dans l'élaboration de règles de fonctionnement implicites et intrinsèques. En effet, les *réseaux de neurones* permettent de répondre au *Symbol Grouding Problem*. Les symboles n'y sont pas formalisés explicitement par le concepteur du réseau, ils sont formés par conditionnement, à l'aide d'une procédure d'apprentissage. Le réseau est organisé, lors de cette procédure, de manière à faire correspondre certaines entrées (par exemple des images d'oiseaux) aux sorties désirées (le mot « oiseau »). Le symbole associé à la catégorie apprise n'est contenu dans aucun neurone, mais dans l'agencement qu'ils prennent lors du conditionnement : il est une propriété du système et non plus un de ses éléments. Par la suite, l'association réalisée par le réseau entre l'image d'un oiseau et le symbole « oiseau » n'est pas effectuée arbitrairement, mais elle dépend de la structure que le réseau a pris lors de ses « expériences » passées. La sémantique des symboles est alors contenue globalement par cette structure contingente : elle est propre au réseau. Ce nouveau modèle de la cognition émergente garanti ainsi l'autonomie sémantique du système. La notion d'indépendance implémentationnelle, chère au symbolisme, est sacrifiée au profit de l'intentionnalité. En effet, la manipulation symbolique dépend fondamentalement du *niveau sub-symbolique* qui la fait émerger.

Remarquons que la notion même d'apprentissage n'est pas nécessaire à la cognition. Seul le résultat l'est, à savoir une définition non arbitraire des symboles qui sont alors liés de manière causale à leurs référents. De la même manière que le clone imaginé par Harnad est, du point de vue du physicalisme, *exactement le même homme*[68], un réseau de neurones dont les états locaux et la structure relationnelle est identique à un autre est *exactement le même réseau*. Ainsi, il est possible qu'un ingénieur conçoive un tel réseau sans user des méthodes de conditionnement, i.e. en implémentant le réseau directement avec la structure qui le caractérise. Ce réseau aura beau être âgé de quelques secondes, il implémentera *exactement la même forme d'intentionnalité* que son homologue conditionné. L'apprentissage est cependant incontournable en pratique, puisque la conception *ex nihilo* d'une structure aussi complexe est irréalisable avec les technologies actuelles.

1988) pour une présentation détaillée des champs de l'Intelligence Artificielle reposant sur le principe d'émergence et s'opposant ainsi au cognitivisme.

[68] Cf. section 1.1.2 et (Harnad, Can a machine be conscious? How?, 2003)

4.1.5 Bilan

- Le sens fonctionnel est conservé : l'utilisation d'un dictionnaire de symboles ne correspond pas à un fonctionnement définissant un esprit humain.
- Le test de Turing est modifié : « avoir un esprit » signifie « passer avec succès le test de Turing » *et* « être capable d'intentionnalité ».
- La théorie n'est pas valide : le *cognitivisme* ne peut pas modéliser l'esprit humain puisqu'il permet de modéliser des machines non-intentionnelles passant avec succès le test de Turing.
- Une nouvelle théorie est conceptualisée : le *connexionnisme* modélise les machines intentionnelles par émergence. Le zombie de la chambre chinoise est absent de cette nouvelle théorie de l'esprit.

4.2 Deuxième cas : folie et cécité

4.2.1 Modèle : connexionnisme et émergence

Comme nous l'avons vu dans la section précédente, le zombie imaginé par Searle nous incite à abandonner le modèle cognitiviste de l'esprit en faveur d'un nouveau modèle : le *connexionnisme*. Celui-ci rend compte d'une utilisation intentionnelle des symboles fondée sur l'émergence de structures entre les nœuds de réseaux formels. Outre le fait que ce modèle cognitif est plus proche des modèles de réseaux neuronaux communément utilisés en neurosciences, il décrit un mode de référence entre le symbole et le monde qui n'est plus simplement arbitraire, mais qui repose sur une relation causale conditionnée lors de l'apprentissage. C'est sur cette base théorique que nous abordons notre seconde utilisation du test de Turing.

4.2.2 Cas de fou : les chatons aveugles

Dans le cadre d'une étude devenue classique, Held et Hein élevèrent des chatons dans l'obscurité et les exposèrent à la lumière seulement dans des conditions contrôlées[69]*. Un premier groupe d'animaux furent autorisés à circuler normalement, mais ils étaient attelés à une voiture et à un panier contenant le second groupe d'animaux. Les deux groupes partageaient donc la même expérience visuelle, mais le second groupe était entièrement passif. Quand les animaux furent relâchés après quelques semaines*

[69] Cf. (Held & Hein, 1958)

> *de ce traitement, les chatons du premier groupe se comportaient normalement, mais ceux qui avaient été véhiculés se conduisaient comme s'ils étaient aveugles : ils se cognaient contre les objets et tombaient par-dessus les bords.*[70]

L'expérience, rapportée ici par Varela, peut être interprétée comme un test de Turing particulier. La compétence à détecter concerne l'usage de la vue pour se représenter un environnement complexe. Le critère du test consiste à déterminer si un individu arrive à se déplacer en évitant les obstacles lorsqu'il a les yeux ouverts, et ce de façon « normale »[71] à condition que ce ne soit pas la première fois qu'il se retrouve confronté visuellement à ces obstacles. Un tel comportement défini empiriquement la notion de *représentation visuelle de l'environnement*.

Le premier groupe de chatons passe ainsi avec succès le test et l'on conclu qu'ils sont capables d'utiliser des informations visuelles pour se déplacer. Le second groupe de chatons échoue au test. Ils sont alors considérés comme des *fous*. En effet, on présuppose que les chatons du second groupe, comme ceux du premier, ont des expériences visuelles lorsqu'ils ont les yeux ouverts et on suppose qu'ils sont capables de les exploiter pour détecter et éviter les obstacles. Ces présuppositions reposent sur le modèle *a priori* que l'on a concernant l'esprit des chatons : nous présupposons que tous les membres de cette espèce fonctionnent *à peu près* comme nous en ce qui concerne la détection d'obstacles et la volonté de les éviter. On en déduit que leur comportement « normal », lorsqu'ils se déplacent les yeux ouverts, consiste à éviter ces obstacles. A ce titre, les chatons du second groupe, bien qu'ils possèdent selon notre modèle les expériences et facultés cognitives nécessaires, ont un comportement « anormal ». Nous les appelons donc des *fous* (conformément au terme que nous avons défini dans la section 3.2.1).

La représentation visuelle de l'environnement est une compétence cognitive qui peut être modélisée par le connexionnisme. Le réseau formel constituant l'esprit des chatons apprend à utiliser les diverses expériences perceptives pour adapter leur comportement à l'environnement dans lequel ils se déplacent. La notion d'obstacle, même si elle n'est pas nommée ainsi, émerge dans le réseau par apprentissage et conditionnement. Elle est enfin utilisée pour déterminer l'action des chatons. Dans la mesure où le connexionnisme est

---

[70] Cf. (Varela, Thompson, & Rosch, 1993) Pages 236 et 237

[71] On suppose ici que le même individu n'arriverait pas à se déplacer « normalement » s'il avait les yeux fermés, ou du moins qu'il n'y arriverait pas de la même manière. On suppose également que cette différence se traduirait dans son comportement. Par exemple, l'individu aurait recours à ses membres pour sonder l'espace et détecter les obstacles. Un tel mode de déplacement est alors qualifié d'« anormal. »

compatible avec un tel cas de folie, il est nécessaire de démêler la contradiction que ce cas particulier induit.

4.2.3 Vrai fou ou faux fou ?

Deux positions peuvent être défendues pour expliquer l'expérience des chatons aveugles : (1) il s'agit de *vrais fous* ou (2) il s'agit de *faux fous*. Cette section explicite les deux positions et leurs conséquences.

La position (1) consiste à affirmer que les chatons aveugles sont capables de se représenter visuellement leur environnement, mais qu'ils agissent comme si ce n'était pas le cas. Nous choisissons alors de conserver le sens fonctionnel selon lequel *percevoir visuellement l'environnement* est une condition suffisante pour apprendre à se le *représenter*. Puisque les chatons sont dans ce cas (ils ont les yeux ouverts et ce n'est pas la première fois qu'ils ont une telle expérience perceptive), le test de Turing doit être modifié pour inclure leurs comportements étranges dans la définition *a priori* des comportements liés à la *représentation visuelle*.

La condition suffisante utilisée pour redéfinir le test de Turing peut être assez simple. On peut dire : « un chaton se représente visuellement son environnement si, et seulement si, il passe avec succès le test de Turing de la représentation visuelle (consistant à éviter des obstacles) *ou alors* si, lors d'un tel test, il se comporte comme les chatons décrits par Held et Hein (ils n'arrivent pas à éviter les obstacles). » Une telle condition met en péril la définition du test dans la mesure où elle peut engendrer de nombreux cas de *faux positifs*. En effet, des chatons ne sachant pas se représenter leur environnement sur la base d'une expérience visuelle (soit parce qu'ils n'ont pas appris à le faire, soit parce qu'ils ont des troubles neurologiques) se comportent de manière similaire aux chatons de l'expérience. La normalisation de leur comportement implique qu'ils soient tous considérés comme étant doués de la capacité de représentation alors que certains ne le sont pas.

La condition peut alors porter sur le fonctionnement interne des chatons et non sur leur comportement (définition mixte). La formulation d'une telle condition suffisante repose alors sur les recherches de l'éthologie cognitive. Nous pourrions dire par exemple : « un chaton se représente visuellement son environnement si, et seulement si, il passe avec succès le test de Turing de la représentation visuelle, *ou alors* si son cerveau fonctionne de telle manière. » Cette approche nous amène à définir les bases neurologiques de la représentation. Il peut être

difficile de les détecter ensuite chez les machines organiques que nous voulons tester. Dans le cas des machines électroniques cependant, il est possible de savoir si le code implémente ou non la fonction suffisante à la représentation. Il est également nécessaire d'expliquer pourquoi la représentation visuelle de l'environnement n'incite pas, dans le cas des chatons fous, à éviter les obstacles.

La position (2) consiste à affirmer que les chatons aveugles, malgré le sens fonctionnel, ne se représentent pas leur environnement. Il n'est donc pas étonnant qu'ils « se cognent contre les objets et tombent par-dessus les bords. » Le sens fonctionnel est modifié au profit du test de Turing de la représentation visuelle qui est alors préservé. La modification est formulée en termes de condition nécessaire à la représentation visuelle.

Cette seconde approche revient sur le modèle cognitif associé au chaton. La capacité de représentation de l'environnement ne lui est plus attribué par le seul fait qu'il a les yeux ouverts et que ce n'est pas la première fois qu'il aperçoit de tels obstacles. Nous sommes dans le cas d'un *vrai négatif* que nous avions tout d'abord interprété comme un *faux négatif* : le chaton n'a pas de représentation visuelle de l'environnement et l'apprentissage n'est pas une condition suffisant à cette capacité[72]. La section suivante expose la condition nécessaire qui, notamment dans le travail de Varela, a conduit à préciser le modèle connexionniste de l'esprit.

4.2.4 Choix historique

Varela affirme que l'expérience des chatons aveugles nous révèle que l'apprentissage de certaines aptitudes cognitives ne peut se faire de manière passive. En effet, le cas de *faux fou* nous conduit à affirmer que certains chatons n'arrivent pas à se représenter visuellement l'environnement *alors qu'ils y ont déjà été confrontés visuellement*. Le premier groupe de chaton permet d'induire une différence fondamentale entre les deux modes d'apprentissage : le premier est *actif* et le second *passif*.

Le biologiste et neurologue Francisco Varela montre les limites des modèles cognitivistes et connexionnistes de l'esprit, ainsi que celles d'une éventuelle synthèse des deux[73]. La cognition ne coïncide pas selon Varela avec la définition générale qu'en donne

[72] L'expérience de Held et Hein est réalisée de telle sorte que l'apprentissage soit le seul facteur variant entre les deux échantillons de chatons. Sans cela, il aurait été possible de s'interroger sur d'autres aspects du modèle, par exemple sur le fait que les yeux ouverts suffisent ou non à la représentation visuelle de l'environnement.

[73] Cf. (Varela, Thompson, & Rosch, 1993) pour une discussion très détaillée des deux courants et une large présentation de la *voie moyenne* proposée par l'auteur. Cf. également (Varela F. J., 1988) pour un résumé des trois modèles et de leurs oppositions historiques et conceptuelles.

l'Intelligence Artificielle et la grande majorité des sciences cognitives : elle ne peut être circonscrite à la résolution de problème par *représentation d'un environnement prédonné*. Dans *L'Inscription corporelle de l'esprit*[74], le philosophe renoue avec la phénoménologie pour construire une *voie moyenne* entre l'idéalisme subjectif à la Berkeley (ou le point de vue subjectif de Nagel) et l'objectivisme des sciences cognitives (le *point de vue de nulle part*)[75]. La cognition ne réside pas uniquement dans la capacité à se représenter un monde qui existe en soi (i.e. de manière objective), elle consiste également à construire subjectivement ce monde en *faisant émerger* ses caractéristiques essentielles. Cette émergence du monde, ou *énaction*, est associée à l'émergence de ses représentations dans le cas du connexionnisme. Elle nécessite une *action* sur le monde perçu et un couplage entre cette action et cette perception. L'énaction d'un monde ne peut être effectuée de manière passive, ce qui explique pourquoi les chatons de Held et Hein, dont les premières expériences visuelles n'ont pas été guidées par leur action propre, n'ont pas appris à se représenter visuellement leur environnement.

Les conséquences de cette expérience ne sont pas limitées à la famille des félidés. Varela fait acte de plusieurs autres indices attestant d'une *cognition énactive* dans l'ensemble du règne animal. De plus, l'énaction ne se limite pas à la représentation visuelle de l'environnement mais à l'ensemble des activités cognitives qui sont en jeu lors de la création du moment suivant. « [Les sous-réseaux neuronaux] n'engagent pas seulement l'interprétation sensorielle et l'action motrice, mais aussi toute la gamme des attentes cognitives et la tonalité émotionnelle, qui sont centrales dans le façonnement d'un moment de l'action. »[76] La notion d'intentionnalité peut être élargie à ces « attentes cognitives ». La difficulté de la cognition consiste donc à faire émerger un monde *intentionnellement* sur la base d'expériences sensori-motrices désordonnées. L'intentionnalité du sujet consiste à faire apparaître un monde authentique dont les caractéristiques essentielles ne sont pas déterminées *explicitement* (i.e. prédonnées par le monde), mais *implicitement* (i.e. construites par le sujet conscient).

Pour en revenir à l'expérience des chatons aveugles, la contemplation passive de l'environnement par le second groupe n'a fait *énacter* aucun monde subjectif pour la bonne et simple raison que leur action n'était pas dirigée vers cet environnement. Il y a une absence d'intérêt pratique pour ces images qui défilent. L'apprentissage passif est donc inefficace, ce

[74] Cf. (Varela, Thompson, & Rosch, 1993)
[75] Cf. la présentation de ces différents points de vue dans la section 2.1.3.
[76] Cf. (Varela, Thompson, & Rosch, 1993) Pages 238 et 239

qui explique pourquoi les chatons sont incapables, dans un second temps, d'utiliser leurs expériences visuelles nouvellement liées à leur action et leur intention pour se déplacer dans un environnement complexe.

Des chercheurs en Intelligence Artificielles ont travaillé à une mise en pratique du modèle énactif de l'esprit. Les plus connus sont sans doute Rodney Brooks et Luc Steels. Ces deux roboticiens affirment que, pour construire des machines aux capacités cognitives développées, il est nécessaire (1) de les immerger dans un environnement complexe et (2) que leurs moyens d'actions et de perceptions soient diversifiés. Le corps est alors envisagé comme une variable expérimentale responsable, en autres choses, de l'émergence de l'esprit[77]. A long terme, le projet de Brooks inclut une perspective évolutionniste à la conception de robots cognitifs. En commençant par l'élaboration de machines simples et efficacement connectées avec le monde, il est possible de sélectionner les plus performantes pour appuyer la conception de machines plus développées[78].

4.2.5 Bilan

- Le test de Turing est conservé : « avoir une représentation visuelle de son environnement » correspond empiriquement à « savoir éviter les obstacles lorsqu'on a les yeux ouverts ».
- Le sens fonctionnel est modifié : l'apprentissage ne suffit pas à développer des compétences cognitives telle que la représentation visuelle de l'environnement, il est nécessaire de procéder à un apprentissage *actif*.
- La théorie n'est pas valide : le *connexionnisme* ne suffit pas à modéliser la cognition puisqu'il permet de modéliser des machines qui, malgré une certaine forme d'apprentissage, ne sont pas dotées des compétences cognitives adéquates.
- Une nouvelle théorie est conceptualisée : le *modèle énactif de l'esprit* modélise les machines cognitive sur la base d'une boucle sensori-motrice. La cognition y est nécessairement *incarnée* et le cas des chatons aveugles, pour lesquels la composante motrice est absente, est réglé par cette nouvelle théorie de l'esprit.

---

[77] Cf. (Kaplan & Oudeyer, 2008) pour une discussion sur la notion de corps vue comme *variable expérimentale*.
[78] Cf. par exemple (Steels & Brooks, 1995) et (Brooks, 1999) pour un aperçu des modèles et applications réalisés par les deux roboticiens.

4.3 Troisième cas : retour à la chambre chinoise

De nombreux contre-arguments ont été opposés à l'expérience de la chambre chinoise, certains pour sauver coûte que coûte le modèle cognitiviste de l'esprit et d'autres, précédent de peu la révolution du connexionnisme et du modèle énactif de l'esprit, pour montrer que la vision étroite de Searle concernant les mécanismes cognitifs de son zombie pouvait être dépassée par l'Intelligence Artificielle. Searle répond cependant en détails à tous ces contre-arguments dans son article intitulé *Minds, Brains, and Programs*[79]. Le zombie de la chambre chinoise, après avoir été démasqué une première fois et supprimé dans la section 4.1, revient hanter la philosophie de l'esprit.

4.3.1 Modèle : énaction

L'expérience présentée dans la section précédente nous a incités à concevoir la cognition comme une activité fondamentalement incarnée : l'intentionnalité y est implémentée par une boucle sensori-motrice dirigée vers le monde. L'activité symbolique et la représentation mentale du monde émergent ainsi d'un environnement complexe. Cependant, le modèle connexionniste ainsi que le modèle énactif de l'esprit peuvent à leur tour être mis à mal par les critiques de Searle.

4.3.2 Cas de zombie : le retour de la chambre chinoise

Dans la mesure où les opérations élémentaires effectuées par un processeur peuvent être effectuées par un être humain, il est possible d'imaginer qu'un homme, manipulant des 0 et des 1, remplace l'ordinateur pour l'exécution d'un programme. Si cet homme ne connait pas la signification de la suite de bits qu'il manipule, il connait néanmoins les règles de calcul formelles permettant d'exécuter le programme. L'absence d'intentionnalité semble s'appliquer également au niveau sub-symbolique du modèle connexionniste : le zombie de la chambre chinoise, au lieu d'utiliser un dictionnaire, simule le comportement de chaque unité du réseau et détermine ainsi quelles sorties il doit produire. En d'autres termes, le zombie ne s'occupe plus directement d'imiter l'activité symbolique d'un locuteur chinois, il imite maintenant son activité sub-symbolique. Dans les deux cas, il ne possède pas le sens des opérations qu'il effectue. Hormis une différence significative concernant le temps d'exécution, il est impossible pour l'observateur de distinguer l'activité du programme

[79] Cf. (Searle, 1980)

exécuté par un processeur de celle de l'homme singeant le programme, les sorties étant identiques. La critique de Searle peut être ainsi étendue au connexionnisme.

Searle envisage le cas d'une activité sensori-motrice dans ce qu'il appelle *The Robot Reply*[80]. Dans ce cas, le programme est embarqué dans un robot capable d'interagir avec son environnement. Searle admet dans un premier temps que cette réponse a le mérite de dépasser le modèle symboliste de la cognition. « Cognition is not solely a matter of formal symbol manipulation, since this reply adds a set of causal relation with the outside world. »[81] Cependant, cela ne change rien au problème. Le zombie de la chambre chinoise peut, encore une fois, être aux commandes du robot. Les sorties du programme, au lieu d'être affichées sur un écran (dans le cas du test original), servent à actionner les membres du robot et les entrées sont le fait de capteurs (mais Searle ne le sait pas). Le zombie à l'intérieur de la chambre ne comprend rien à ce qu'il fait, mais il arrive néanmoins à diriger le robot de manière très convaincante.

L'association de ces deux réponses constitue une critique du connexionnisme et du modèle énactif de l'esprit. Même si Searle n'a pas argumenté directement contre ces deux modèles (son article précède d'ailleurs la vision énactive de Varela), nous pouvons utiliser ses arguments pour concevoir un zombie révélant une incohérence dans les nouvelles théories. Encore une fois, c'est l'intentionnalité de ces systèmes qui est mise à mal. Puisque le zombie ne comprend rien à ce qu'il fait, mais qu'il manipule les neurones d'un réseau responsable de l'action complète d'un robot sophistiqué, il est la preuve que les deux modèles échouent à définir la notion d'esprit.

4.3.3 Vrai zombie ou faux zombie ?

Nous disposons, encore une fois, de deux alternatives face à ce cas de zombie : (1) il s'agit d'un *vrai zombie* ou (2) il s'agit d'un *faux zombie*. Cette section explicite les deux positions et leurs conséquences.

La position (1) consiste à conserver le sens fonctionnel et modifier le test de Turing. Il faut trouver une condition nécessaire à l'intentionnalité que le nouveau zombie ne remplit pas, afin de supprimer ce cas problématique. Cette condition permet de définir un modèle de l'esprit à partir duquel un tel zombie ne peut pas être décrit.

---

[80] « La réponse du robot » Cf. (Searle, 1980) Section III, *The Robot Reply*

[81] « La cognition n'est pas seulement une affaire de manipulation de symboles, puisque cette réponse ajoute tout un ensemble de relations causales avec le monde extérieur. » Ibid. traduction de É. Duyckaerts

Quelle peut être cette condition nécessaire qui manque à notre zombie ? Peut-être que Searle, à terme, lutte contre l'idée même d'un esprit implémenté par un cerveau en silicium. De même, un cerveau ne peut résider dans une chambre chinoise, ou dans l'association d'un milliard de chinois pour reprendre l'expérience de pensée de Ned Block[82]. C'est l'idée même du fonctionnalisme qui est mis à mal par Searle : il ne suffit pas d'avoir le même fonctionnement qu'un cerveau humain pour avoir un esprit, il faut en plus que ce cerveau soit structurellement similaire à celui de l'homme, à savoir qu'il soit composé de neurones et de synapses. Il faut un cerveau organique pour avoir un esprit, voilà la condition nécessaire !

La position (2) consiste à affirmer que, malgré tous les arguments de Searle, le zombie de la chambre chinoise possède bel et bien un esprit. Le sens fonctionnel est modifié, le test de Turing préservé et il en résulte une condition suffisante à la définition de l'esprit. Cette condition est de fonctionner selon la théorie connexionniste et le modèle énactif, peu importe la structure implémentant l'esprit, peu importe la matière dont notre cerveau est constitué.

4.3.4 Choix historique

Ce dernier cas de zombie incite finalement à se positionner vis-à-vis du *fonctionnalisme* lui-même : soit on accepte de dire qu'une machine reproduisant dans le moindre détail le fonctionnement d'un cerveau humain a un esprit au même titre que l'homme en question, soit cela ne suffit pas et l'on exprime des contraintes concernant la structure du cerveau (sa composition chimique, sa taille, la distance entre les neurones et leurs moyens physiques de communication… toutes ces caractéristiques peuvent être contraintes par des conditions structurelles nécessaires à la possession d'un esprit). La position (2) revient donc à dire que le fonctionnalisme est insuffisant.

Pour sauver cette théorie à l'origine de l'Intelligence Artificielle forte, il est nécessaire de changer le sens fonctionnel. Le zombie de Searle n'en est pas un : il est conscient, il a des états mentaux, dans le même sens qui nous pousse à dire que les autres hommes sont conscients et ont des états mentaux. Ainsi Dennett répond à Searle qu'il y a bien un esprit dans la chambre chinoise, seulement il ne s'agit pas de l'esprit de Searle. « He is not alone in the room. »[83] Mais puisque les esprits ne peuvent être directement connus, il n'y a pas de raison pour que Searle, même s'il est à l'intérieur de la chambre, soit capable de le détecter. Pire, Searle suppose que le zombie, c'est-à-dire lui-même enfermé dans la chambre chinoise,

[82] Cf. (Block, 1978) pour l'expérience de la *Nation Chinoise*.
[83] « Il n'est pas seul dans la chambre. » (Dennett, 1991) Section 1, *Imagining a Conscious Robot*

est identique à la machine qui est analysée par le test de Turing. C'est confondre les unités fonctionnelles du cerveau avec son état global. Chaque neurone est un zombie, ils n'ont pas d'esprit, et pourtant leur somme, le cerveau, en a un. La chambre chinoise, pris comme un tout, est donc doté d'un esprit selon la définition fonctionnaliste. C'est la *réponse du système* (ou *system reply*[84]).

La réponse du système associée à la notion de complexité permet à Dennett d'affirmer que ce genre d'esprit existe : dans l'expérience de Block[85] il y a bien un esprit, implémenté par l'ensemble de la population chinoise ; dans l'exemple du cerveau à canaux hydrauliques[86] il y a bien un esprit implémenté par ces réserves d'eau et ces vannes ; de même pour le martien de Lewis[87] qui, malgré son cerveau rempli d'eau, a des états mentaux ; enfin, dans le cas de la chambre chinoise, il y a un esprit implémenté par le dictionnaire, l'utilisation qu'en fait Searle, les portes d'entrées et de sorties, le reste de la chambre. Défendre le fonctionnalisme, c'est affirmer que toutes ces machines ont un esprit. Le sens fonctionnel doit alors intégrer ces nouvelles formes.

Block parle de « chauvinisme neuronal » pour qualifier la position selon laquelle seul un cerveau organique peut être défini comme ayant un esprit. Il est difficile de s'imaginer *comment* la population chinoise ou un circuit en silicium peuvent donner naissance à un esprit, à une conscience perceptive et réflexive. Dennett remarque qu'il est tout aussi difficile de s'imaginer *comment* un amas de neurones et de molécules sans vie peut également donner naissance à un tel esprit. « And yet we imagine human beings to be conscious, even if we still can't imagine how this could be. »[88] David Chalmers, dans la même perspective, affirme que nous ne devrions traiter les deux cas de manière identique[89] : s'il est surprenant que l'homme possède un esprit, il n'est pas *plus surprenant* qu'une machine électronique en ait un également.

---

[84] Cf. (Dennett, 1991) Section 1, *Imagining a Conscious Robot* et (Searle, 1980) Section I, *The system Reply*, pour la réponse de Searle.
[85] Cf. (Block, 1978)
[86] Cf. (Searle, 1980) Section III, *The brain simulator reply*
[87] Cf. (Lewis, 1978) Section I
[88] « Et cependant nous imaginons sans hésiter que les hommes sont conscients, même si nous ne pouvons toujours pas imaginer comment cela est possible. » (Dennett, 1991) Section 1, *Imagining a Conscious Robot*
[89] Cf. (Chalmers D. J., 1996) Section IV.9.1, *Machine Consciousness*

5 Conclusion

5.1 Bilan

Dans l'introduction de ce mémoire, nous avons présenté une problématique générale de la philosophie de l'esprit : *quelles machines ont un esprit, et pourquoi ?* Répondre à cette question, ce n'est pas seulement *faire l'inventaire des machines possédant un esprit*, c'est aussi *justifier cet inventaire*. Il est possible d'utiliser une théorie de l'esprit pour résoudre cette problématique. Une telle théorie décrit la nature et le fonctionnement de l'esprit, elle fait la liste de ses propriétés. En ce sens elle en élabore une définition *en intension*. Les théories peuvent être utilisées pour déterminer quelles machines ont un esprit, lesquelles n'en ont pas, et pour argumenter ces choix. Seulement, certaines théories peuvent s'opposer et ainsi répondre différemment à la problématique générale.

L'objectif de ce mémoire ne consiste pas à résoudre la problématique générale. Il ne prend pas position en faveur ou en défaveur de théories particulières, mais s'intéresse à la façon dont elles résolvent la problématique et aux différentes réponses qu'elles engendrent. En outre, ce mémoire s'interroge sur les moyens dont on dispose pour évaluer les théories et les confronter.

Dans la seconde partie, nous avons présenté une limite épistémologique à la connaissance que nous pouvons avoir de l'esprit des autres machines. L'origine de cette limite réside dans l'asymétrie entre la connaissance *directe* de notre propre esprit et la connaissance *indirecte* des autres esprits. Ce problème empêche toute vérification ontologique des théories de l'esprit et met en péril leur évaluation. Pire, il supprime le sens même du mot « esprit » dans la mesure où il ne peut en exister qu'un : le nôtre. Les théories ne sont plus générales, mais particulières. Il est nécessaire, pour aller plus en avant, de redonner du sens à l'esprit.

Le test de Turing[90] est une méthode empirique permettant définir ce qu'est « une machine dotée d'un esprit ». Il repose sur une conception *fonctionnaliste* de l'esprit : celui-ci est définit comme une compétence des machines, comme une *fonction* dont l'activité est détectable via l'étude de leurs comportements. Le test de Turing est à ce titre un béhaviourisme méthodologique. La nouvelle définition qu'il donne, puisqu'elle repose sur ce que l'on peut *directement* observer, contourne le problème des autres esprits. Le test ne

[90] Cf. (Turing, 1950)

prétend pas définir un modèle ontologique de l'esprit, mais il en donne une définition *en extension*. En outre, il permet de répondre à la problématique générale de ce mémoire : *les machines qui ont un esprit sont celles qui se comportent comme telle*, et ce en vertu de la définition construite par le test.

Selon Stevan Harnad, le test de Turing a le dernier mot[91]. La troisième partie de ce mémoire s'oppose à cette assertion en révélant les contradictions que le test peut engendrer. Pour ce faire, nous présentons une méthode duale pour définir l'esprit : le *sens fonctionnel*. Cette méthode oppose aux *comportements* des machines la notion de *fonctionnement*. De la même manière que le test de Turing repose sur une définition *a priori* des comportements dotés d'un esprit, le sens fonctionnel repose sur une définition *a priori* des fonctionnements dotés d'un esprit. Les cas de machines pour lesquelles la définition fonctionnelle du sens fonctionnel n'est pas cohérente avec la définition comportementale du test de Turing témoignent d'une incohérence. Les *fous* sont ainsi des machines qui ont un esprit selon la définition fonctionnelle et qui n'en ont pas selon la définition comportementale. Les *zombies* sont au contraire des machines qui n'ont pas d'esprit selon la définition fonctionnelle et qui en ont un selon la définition comportementale.

Chacun de ces cas problématiques peuvent être traités de deux manières différentes : soit en modifiant la définition fonctionnelle du sens fonctionnel, soit en modifiant la définition comportementale du test de Turing. Ce choix repose sur une prise de position vis-à-vis d'un cas particulier de *fou* ou de *zombie*. Ce mémoire ne prétend pas déterminer quelle position est plus légitime que l'autre, mais il s'efforce d'en expliciter les conséquences. En outre, le problème des autres esprits interdit de donner des arguments ontologiques pour départager les positions : à ce titre, elles sont toutes les deux légitimes. La dernière section de la troisième partie s'intéresse alors aux conséquences de ces choix en ce qui concerne la modification des définitions *en extension* (sens fonctionnel et test de Turing) et aux conséquences sur les définitions *en intension* (théories de l'esprit).

Lorsqu'une théorie de l'esprit permet de modéliser une machine pour laquelle il y a contradiction entre le sens fonctionnel et le test de Turing, il est nécessaire de résoudre cette contradiction et d'en tirer des conséquences sur le modèle. Suivant la position adoptée pour régler le cas de *fou* ou de *zombie* rencontré par la théorie, celle-ci peut être confirmée ou infirmée. En outre, défendre une théorie consiste à défendre conjointement des définitions

[91] Cf. (Harnad, Can a machine be conscious? How?, 2003) Fin de l'article

fonctionnelles et comportementales cohérentes. Les cas problématiques permettent alors de révéler les conséquences sur des points critiques concernant la définition de l'esprit.

La quatrième partie expose des théories de l'esprit qui ont été discutées dans la seconde moitié du XX^{e} siècle. Nous proposons une relecture des débats qu'elles ont historiquement suscités à l'aide des outils développés dans ce mémoire. Nous présentons ainsi des cas de *fous* et de *zombies* permettant d'expliciter des positions critiques, en faveur ou en défaveur des théories, et les conséquences de ces positions quant à la notion d'esprit. Les trois théories que nous présentons sont des théories *fonctionnalistes*, reposant donc sur une conception fonctionnelle de l'esprit. La succession historique de ces théories forme ainsi ce que nous appelons la valse des théories.

La première théorie évaluée est le *cognitivisme*, selon laquelle l'esprit est un système symbolique et la cognition une manipulation de symboles. L'expérience de la chambre chinoise, imaginée par John Searle[92], s'oppose à ce modèle en y révélant un cas de *zombie*. Elle centre le débat sur la notion d'*intentionnalité*. Les deux positions, permettant de résoudre la contradiction présentée par le zombie de la chambre chinoise, consistent donc à affirmer la nécessité ou la non-nécessité de l'intentionnalité pour construire le concept d'esprit humain. Le cognitivisme, pour lequel l'intentionnalité semble absente, est alors confirmé ou infirmé.

La seconde théorie évaluée est le *connexionnisme*. Elle modélise la cognition par émergence de structures globales à partir d'un niveau d'activité *sub-symbolique*. Nous interprétons une expérience menée par Held et Hein[93] comme un cas de *folie*. Ce cas problématique nous invite à nous interroger sur le caractère suffisant du connexionnisme pour rendre compte des capacités de *représentation du monde*.

La troisième théorie évaluée est le *modèle énactif de l'esprit*, notamment défendu par Francisco Varela[94]. Elle ajoute au modèle connexionniste la nécessité d'une boucle sensori-motrice pour développer des aptitudes représentationnelles. Le zombie de la chambre chinoise s'oppose, encore une fois, à l'association de ces deux théories. Les deux positions engendrées par ce cas de *zombie* cristallisent finalement le débat autour de la théorie *fonctionnaliste* elle-même. La résolution du conflit entraîne donc, suivant la position adoptée, la confirmation ou l'infirmation de la conception fonctionnelle de l'esprit.

---

92 Cf. (Searle, 1980)
93 Cf. (Held & Hein, 1958)
94 Cf. (Varela, Thompson, & Rosch, 1993)

5.2 Perspectives

Nous proposons de continuer les travaux présentés dans ce mémoire selon deux axes de recherche. Le premier consiste à intégrer les outils que nous avons développés à la méthode *falsificationniste*. Le second propose de s'intéresser aux relations entre *fonctionnements* et *comportements*.

Le *falsificationnisme* est une école de pensée de la philosophie des sciences proposant d'évaluer les théories scientifiques en fonction de leur *réfutabilité*. La méthode ainsi défendue consiste à confronter une théorie à des expériences dont elle est sensée prédire l'issue. L'évaluation repose donc sur des cas particuliers observables, des expériences critiques, dont le rôle n'est pas sans rappeler celui des cas problématiques présentés dans ce mémoire : *fous* et *zombies*. Nous pensons à ce titre que le falsificationnisme peut avantageusement encadrer notre démarche générale. L'objectif de cet axe de recherche consiste donc à reformuler les oppositions entre test de Turing et sens fonctionnel en termes falsificationnistes. Nous pouvons nous appuyer sur le travail de Karl Popper[95], père du falsificationnisme, ou sur la remarquable présentation d'Alan Chalmers[96].

Les définitions duales du test de Turing et du sens fonctionnel nous incitent à nous intéresser aux liens qu'entretiennent en pratique ces deux définitions. La compréhension de leurs cohérences ou de leurs incohérences peut être améliorée par une étude des relations entre *fonctionnements* et *comportements*. Ce second axe de recherches ne s'intéresse pas aux relations entres des fonctionnements et des comportements particuliers (celles-ci sont étudiées par des domaines tels que la psychologie, la physiologie et l'ensemble des neurosciences). Nous nous intéressons plutôt aux relations entretenues par ces deux facettes des machines en toute généralité. Les relations causales, notamment, sont à étudier dans la mesure où les deux définitions présentées dans ce mémoire ne semblent pas être indépendantes.

Enfin, l'objectif à long terme de ces travaux consiste à définir une méthode d'évaluation systématique en philosophie de l'esprit. Les exemples présentés dans la dernière partie montrent qu'une telle approche est féconde et qu'elle peut participer à la valse des théories.

---

[95] Cf. (Popper, 1990)
[96] Cf. (Chalmers A. F., 1976)

Bibliographie

Block, N. (1978). Troubles with Functionalism. *Minnesota Studies in the Philosophy of Science , 9*, 261-325.

Brooks, R. (1999). *Cambrian Intelligence : The Early History of the New AI.* Cambridge: The MIT Press.

Carnap, R. (1966). Theoretical Laws and Theoretical Concepts. Dans *Philosophical Foundations of Physics* (pp. 223-274). New York: Basic Books, Inc.

Chalmers, A. F. (1976). *Qu'est-ce que la science ?* (éd. 2nd (1982)). (M. Biezunski, Trad.) Paris: Edition La Découverte.

Chalmers, D. J. (1996). *The Conscious Mind.* New York: Oxford University Press, Inc.

Chomsky, N. (1957). *Syntactic Structures.* The Hague/Paris: Mouton & Co.

Dennett, D. C. (1991). Consciousness Imagined. Dans *Consciousness Explained* (pp. 431-455). New York: Hachette Book.

Denton, D. (1993). *L'Émergence de la conscience. De l'animal à l'homme.* Flammarion.

Descartes, R. (1641). *Les Méditations Métaphysiques.* Académie de Grenoble: PhiloSophie, http://www.ac-grenoble.fr/PhiloSophie/.

Einstein, A. (1921). *La Théorie de la relativité restreinte et généralisée.* Paris: Gauthier-Villars.

Harnad, S. (2003). Can a machine be conscious? How? *Journal of Consciousness Studies , 10* (4-5), 69-75.

Harnad, S. (1990). The Symbol Grounding Problem. *Physica D , 42*, 335-346.

Held, R., & Hein, A. (1958). Adaptation of disarranged hand-eye coordination contingent upon re-afferent simulation. *Perceptual-Motor Skills , 8*, 87-90.

Hyslop, A. (2009). Other Minds. *Stanford Encyclopedia of Philosophy* .

Jackson, F. (1982). Epiphenomenal Qualia. *Philosophical Quarterly , 32*, 127-136.

Kaplan, F., & Oudeyer, P. (2008). Le corps comme variable expérimentale. *Revue philosophique de la France et de l'étranger , 133* (3), 287-298.

Kirk, R., & Squires, J. E. (1974). Zombies v. Materialists. *The Aristotelian Society , 48*, 135-163.

Lettvin, J., Maturana, H., McCulloch, W., & Pitts, W. (1959). What the Frog's Eye Tells the Frog's Brain. *Proceedings of the Institute of Radio Engineers , 47* (11), 1940-1951.

Lewis, D. (1978). Mad Pain and Martian Pain. *Readings in Philosophy of Psychology , 1*, 216-222.

Minsky, M. (1988). The Mind and the World. Dans *The Society of Mind* (pp. 282-290). New York: Simon & Schuster, Inc.

Nagel, T. (1986). *The View From Nowhere.* New York: Oxford University Press, Inc.

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review , 83* (4), 435-450.

Popper, K. (1990). *Le Réalisme et la science.* Paris: Edition Hermann.

Putnam, H. (1981). Des cerveaux dans une cuve. Dans *Raison, Vérité et Histoire* (pp. 11-32). Paris: Les Éditions de Minuit.

Searle, J. R. (1980). Minds, Brains, and Programs. *The Behavioral and Brain Sciences , 3*, 282-307.

Steels, L., & Brooks, R. (1995). *The Artificial Life Route To Artificial Intelligence: Building Embodied, Situated Agents.* New Haven: Lawrence Erlbaum.

Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind , 59* (36), 433-460.

Varela, F. J. (1988). *Invitation aux sciences cognitives.* Paris: Editions du Seuil.

Varela, F., Thompson, E., & Rosch, E. (1993). *L'Inscription corporelle de l'esprit.* Paris VIe: Editions du Seuil.

Wooldridge, M. (2009). *An Introduction to MultiAgent Systems.* John Wiley & Sons, 2nd edition.

Le test de Turing pour évaluer les théories de l'esprit

Résumé

Ce mémoire propose une méthode pour confronter les théories de l'esprit répondant à la problématique générale : *quelles machines ont un esprit, et pourquoi ?* Dans notre approche, les réponses extensives engendrées par les théories sont utilisées pour délimiter leurs points de désaccords et identifier des cas de divergence cruciaux. L'évaluation et le choix d'une théorie est ainsi éclairé par une analyse focalisée sur de tels cas particuliers.

Voici comment est présentée notre approche. La partie 2 introduit une méthode empirique assez classique pour définir la notion d'esprit de manière *comportementale*. Il s'agit du *test de Turing*. La partie 3 propose d'évaluer ce test en le confrontant à une seconde définition de la notion d'esprit : le *sens fonctionnel*. Les divergences entre ces deux définitions permettent de révéler des contradictions et des incohérences entre deux théories de l'esprit. Le choix entre l'une ou l'autre (et les raisons d'un tel choix) sont obtenus à partir de l'étude de ces cas particuliers. La partie 4 applique cette méthode d'évaluation aux débats historiques qui ont opposé au XX^{e} siècle trois théories fonctionnalistes. Il s'agit du *cognitivisme*, du *connexionnisme* et du *modèle énactif de l'esprit*. Les évaluations passées de ces théories sont reformulées et clarifiées à l'aide de l'approche développée dans ce mémoire.

Mots-clés

Philosophie de l'esprit ; test de Turing ; fonctionnalisme.