Précis de *La Nature de la Volonté,* suivi de la *Disputatio*
A paraître : *Philosophiques,* 2008

Joëlle Proust Institut Jean-Nicod (CNRS, EHESS, ENS)

Le propos de l'ouvrage est à la fois introductif et constructif. Il vise à exposer les deux grandes théories causales en matière d'action, celles de Donald Davidson et de John Searle, de défendre une conception alternative, la théorie volitionniste, et d'en examiner les conséquences sur la liberté de l'agir et sur l'identité personnelle. La défense de la théorie volitionniste passe par l'examen des problèmes que doivent surmonter toutes les tentatives de définir l'action qui font appel à sa structure causale-représentationnelle, en montrant que la conception volitionniste est mieux équipée pour leur apporter une réponse.

Le premier est le problème de Malebranche. Comment expliquer que nous puissions remuer notre bras à volonté, lors même que nous ne savons pas ce qui fait que nous pouvons le remuer? Le problème de Malebranche, s'il était sans solution, donnerait un argument de poids aux philosophes qui, à la suite de Wittgenstein, rejettent la pertinence du registre causal en matière d'action. Malebranche tentait de comprendre le lien entre volonté et exécution comme une relation causale, alors qu'en fait ce lien est d'après eux rationnel: il relève non de l'efficacité causale mais de la justification. La solution du problème de Malebranche passe par plusieurs étapes. On doit la première à Donald Davidson. Elle consiste à montrer que la représentation des raisons d'agir peut être associée à l'efficacité causale dans la mesure où un même événement a des propriétés physiques et mentales. Quand on explique une action par ses raisons, on s'intéresse à l'un des aspects de l'événement mental-cérébral considéré; même si ce n'est pas cet aspect là qui permet de parvenir à des lois causales strictes, c'est le même événement, pris sous un autre aspect, qui intervient à titre de cause ou d'effet.[1] La seconde avancée consiste à réduire l'écart représentationnel entre la représentation mentale de l'action et de son effectuation. On peut en effet montrer que l'agent peut former de ses façons d'agir une représentation plus précise que la simple formulation conceptuelle du résultat ne permet de le faire. C'est là la fonction de ce que John Searle appelle « l'intention en action ». Le contenu d'une intention en action s'articule de la manière suivante: « cette intention en action présente tel mouvement à exécuter, et en vertu de cette présentation, elle cause tel mouvement correspondant ». Grâce au concept d'intention en action, l'effectuation du mouvement devient officiellement partie du contenu intentionnel. La troisième étape consiste à permettre au contenu de l'intention en action d'être non-conceptuel, ce que suggère déjà le recours aux démonstratifs pour caractériser les contenus d'intentions en action. Searle utilise le terme de « présentation », parce que l'intention en action donne directement accès à son contenu dans une expérience caractéristique. Mais ce contenu est d'après lui conceptuel. Or on a objecté depuis que le grain « conceptuel » n'est pas adapté à la représentation de contenus « analogiques » (comme le sont les contenus perceptifs, émotionnels ou agentifs).[2] On verra plus loin que la théorie de la volition permet d'apporter une pierre de plus à la solution du problème de Malebranche, en exposant la structure ontologique de l'action : cette structure est circulaire et non pas linéaire, ce qui permet à la fois d'expliquer pourquoi le sujet peut ignorer les maillons corporels de son action, et de clarifier le statut de la réflexivité dans l'agir.

---

1 On trouvera plus bas, dans la réponse à Daniel Laurier, une discussion de l'épiphénoménalisme des états mentaux à laquelle cette position est associée.

2 C'est pour cela que plusieurs auteurs, suivant les suggestions de Peacocke et d'Evans relatives au contenu non conceptuel de la perception, ont avancé que le contenu des intentions en action devait lui aussi inclure des représentations non-conceptuelles, impliquant entre autres des référentiels égocentriques ou allocentriques, des propriétés de rythme et de fluidité, etc

Le second problème des théories causales de l'agir est celui de la causalité déviante : il peut arriver que le but d'une action soit atteint en vertu de l'impact causal de l'intention de l'agent sur l'environnement, mais d'une manière qui ne permette plus de dire que l'action était intentionnelle. On connaît l'exemple de Chisholm : un neveu tente de tuer à la carabine l'oncle dont il espère l'héritage, le rate, mais déclenche la fuite d'une harde de sangliers, à l'occasion de laquelle l'oncle est tué. Quoique l'oncle soit mort, et qu'il le soit parce que le neveu a pressé sur la gâchette, l'action finale ne satisfait pas les conditions de l'action, parce que les moyens d'agir ne sont pas ceux que l'agent avait en tête. Cet exemple n'est en rien exceptionnel. Il y a une variété de cas de causalité déviante selon qu'elle affecte tel ou tel niveau de la séquence menant des raisons d'agir au résultat final de l'action. C'est du fait de ce problème que Davidson renonce à définir l'action intentionnelle : il parvient à une condition nécessaire, mais non suffisante de l'action intentionnelle. Le concept searlien d'intention en action, dont on a vu qu'il précisait le « comment » de l'action, vise explicitement à répondre à ce problème. Il permet d'écarter les cas de déviance présenté ci-dessus, en indiquant « la condition de satisfaction » de l'intention en action : la victime doit mourir par balle du fait du tir à la carabine. Mais cette stratégie, Searle le reconnaît, ne réussit pas à bloquer entièrement la menace de la déviance causale. En effet, le lien entre l'intention en action et le mouvement corporel peut lui aussi être manipulé par un malin génie. Si chaque fois que j'ai l'intention de lever le bras, le signal est envoyé sans que je le sache à un autre agent qui peut relayer ou non le signal vers le membre, l'action réussie ne sera pas la mienne, quoique mon intention ait pu causer le mouvement. La structure causale déviante s'oppose de nouveau à ce que le lien causal de l'intention en action avec le mouvement corporel puisse *définir* ce qu'est une action.

Searle propose d'exclure ce type de cas par convention, en ajoutant que «la relation de l'intention en action à ses conditions de satisfaction exclut l'intervention d'autres agents ou d'autres états intentionnels » (1985, 138). La solution nominale est-elle le seul recours ? Là aussi, la théorie de la volition permet, on le verra, de faire un pas de plus, en montrant que même les cas où l'action est parasitée par un agent extérieur, il existe un critère non nominal permettant de dire quand l'action est intentionnelle. Quels que soient les mérites de la solution apportée, on ne doit pas méconnaître qu'existe nécessairement une limite asymptotique aux manières de prémunir un organisme contre la causalité déviante : elle demeure nécesssairement possible, du seul fait que nous identifions des mécanismes causaux qui, en tant que tels, peuvent toujours être perturbés. Une théorie qui prémunirait une théorie causale contre *toute* forme de déviance serait ainsi en contradiction avec ses propres présupposés.3 Cette remarque ne dispense pas de chercher à donner une définition qui soit à la fois nécessaire et suffisante de l'action, en montrant que les cas de déviance sont ceux que laisse prévoir la théorie proposée.

Le troisième problème est le problème de Wittgenstein : "Que reste-t-il si l'on soustrait le fait que mon bras se lève du fait que je lève le bras ?". En d'autres termes, qu'est-ce qui distingue mon action de lever le bras du résultat auquel elle aboutit, à savoir que mon bras soit levé? Etant donné qu'une action reste une action quand elle se résume à une tentative inaboutie, la question de Wittgenstein doit recevoir une réponse claire. Les théories de

3 Le problème de la déviance causale ne constitue une difficulté que dans les cas où il s'agit d'exposer la structure causale des représentations du contenu de l'action intentionnelle réussie. Mais il ne porte pas sur la limite de fiabilité de mécanismes qui, on le constate dans de multiples cas, ne sont pas à l'abri d'interférences ou de perturbations pathologiques. L'existence de perturbations dans l'efficience des représentations de l'agir intentionnel est au contraire une prédiction pour toute théorie causale représentationnelle de l'agir et une marque de leur adéquation.

l'action de Davidson et de Searle ne peuvent pas identifier ce "reste" parce qu'en tant que théories des intentions, elles s'exposent au cas fréquent des illusions intentionnelles. Dans les deux cas, les théories avancées conduisent à considérer comme actions des cas où aucune action n'a, en réalité, été accomplie; il suffit que l'agent *croie* qu'il peut agir, qu'il ait des *raisons* de le faire, et qu'il *observe le résultat attendu en conséquence de son* intention (ou de sa raison d'agir) pour qu'il y ait eu action. Mais cette lacune théorique provient du fait que ces théories traitent bien des intentions d'agir, et des croyances nécessaires à l'action, mais ne disent rien du déclenchement ni du contrôle de l'action. Or il est facile de voir, sur un exemple, que les conditions de type épistémique ou conatif ne sont pas suffisantes. Hugh McCann[4] prend l'exemple d'un apprenti yogi qui croit à tort qu'il peut accélérer ses propres battements cardiaques. L'état involontaire d'excitation que crée chez lui l'idée d'y parvenir lui fait battre le coeur plus vite. Soustrayons l'effet. Reste-t-il une action? Davidson doit l'affirmer, puisque le sujet a une raison d'agir: il croit pouvoir le faire, il désire le faire, et il obtient ce résultat parce qu'il a cette croyance et ce désir. Même si l'action échouait, elle resterait une action du fait de sa structure. Searle doit aussi admettre que l'exemple répond aux conditions de l'action : le sujet a l'intention d'accélérer ses battements cardiaques, et cette intention préalable cause l'expérience correspondante (l'intention en action, ici illusoire). A ce problème, le volitionniste s'efforce d'apporter une réponse différente: non, l'apprenti yogi n'a pas exécuté d'action, parce qu'il ne peut pas effectivement contrôler son battement de coeur; ne disposant pas de ce savoir-faire, il ne peut pas lancer l'action correspondante.

Mais voyons maintenant en quoi consiste au juste la théorie de la volition. Toute théorie volitionniste part de l'hypothèse qu'on ne peut agir sans former une volition. Ce qui est essentiel dans l'agir c'est non pas l'arrière-plan rationnel explicite qui conduit le sujet à agir et qui justifie l'action; non pas l'intention que le sujet a pu former avant d'agir, mais l'opération volontaire qui préside à l'exécution de l'action. Comme l'observe Locke :

> "Une chose du moins qui est évidente, à mon avis, c'est que nous trouvons en nous-mêmes la puissance de commencer ou de ne pas commencer, de continuer ou de terminer plusieurs actions de notre esprit, et plusieurs mouvements de notre corps, et cela simplement par une pensée ou un choix de notre esprit, qui détermine et commande, pour ainsi dire, que telle ou telle action particulière soit faite, ou ne soit pas faite". *"* (J. Locke, 1690, II, XX1, 5, 182).

Deux objections sont fréquemment opposées à ce type de définition. Si l'on considère que la volition est elle-même une action mentale présidant au déroulement du mouvement corporel et de ses effets, on rend le concept d'action indéfinissable puisqu'il reparaît dans la définition. Pire encore : cette action mentale semble nécessiter à son tour une volition, et ainsi de suite à l'infini. En fait, on peut dissoudre ces deux objections en examinant de plus près la structure réflexive de la volition. La volition est l'effort de réaliser un changement en vertu de cet effort. Cet effort est constitué par les modifications du sujet qui se met en mesure de produire l'effet visé. Pour bien comprendre pourquoi ces deux dernières formules ne présentent pas de risque de régression à l'infini ni de circularité, il faut analyser plus en détail ce qu'est un épisode de contrôle volontaire et mettre en évidence la structure téléologique de la volition. Appelons « représentation exécutive » le modèle dynamique d'un mouvement (ou d'une activation mentale) et de son bénéfice. L'ensemble de ces représentations exécutives constituent le répertoire dans lequel les volitions sélectionnent leur contenu. Comment cette sélection procède-t-elle ?

La sélection d'un contenu volitif dépend de ce que les théoriciens ont appelé « loi de l'effet » ou « principe de l'action-effet ». Ce principe téléologique pose que l'agent a

4 McCann (1974, 460-1).

tendance à sélectionner les actions qui, dans le passé, ont eu un effet favorable, et à écarter celles qui ont eu un effet défavorable. Le renforcement étant indépendant du contenu particulier de l'action, il s'opère même lorsque le sujet est incapable d'analyser les divers facteurs causalement impliqués dans la réussite de l'action. Pour illustrer cette remarquable propriété : Marteniuk et al. (1987), ont montré, par exemple, que lorsque des sujets doivent saisir un disque soit pour l'insérer dans une fente, soit pour le lancer, la cinématique de leur geste est d'emblée différente. Les conséquences attendues de l'action modulent le début du mouvement. Seule l'analyse téléologique permet d'expliquer ce curieux phénomène de préadaptation du mouvement, qui échappe d'ailleurs à la conscience de l'agent. Ainsi, par parenthèse, le problème de Malebranche reçoit-il enfin une réponse précise. Si le sujet n'a pas à connaître ce qui lui permet d'exécuter son action, c'est en vertu du fait que la volition a une structure circulaire régie par le principe de l'action-effet (nous reviendrons plus bas, dans la réponse à Daniel Laurier, sur les raisons métaphysiques de voir dans le principe d'action-effet le guide heuristique de la définition téléologique de la volition).

Quelque éclairant qu'il soit, le principe de l'action-effet ne peut suffire à rendre compte de la complexité des actions qu'un sujet peut accomplir. S'il explique pourquoi un type d'action est sélectionné et reproduit, il n'explique pas encore comment le mouvement peut être chaque fois *ajusté et modifié* pour répondre aux contraintes éminemment variables auxquels l'agent doit s'adapter. La généralisation du principe téléologique exige que l'on rappelle en quoi consiste le contrôle adaptatif, c'est-à-dire la forme de contrôle qui opère dans des contextes partiellement inconnus. Pour interagir avec un environnement incertain, l'organisme doit pouvoir utiliser les rétroactions antérieures pour sélectionner une représentation exécutive particulière étant donné un but nouveau, la corriger et vérifier l'atteinte du but. Il est important, dans ce cas, de prédire intérieurement les effets qui vont se dérouler durant l'exécution, et comparer avec ces prédictions les observations recueillies. On l'aura compris, ce sont les représentations exécutives qui permettent une telle prédiction intérieure. L'élément capital du contrôle adaptatif est ainsi le *comparateur*, qui permet d'évaluer et de corriger l'action avant même qu'elle ne soit exécutée, sur la base du feedback interne de la volition.

Les deux types de contraintes qui président à la sélection volitive et à l'opération du comparateur sont bien connues par les mathématiciens du contrôle adaptatif (Aubin, à paraître). 1) Les "lois de régulation" associent à chaque nouveau contexte les commandes susceptibles d'atteindre une cible étant donné les contraintes. 2) les "lois de rétroaction" déterminent quelle partie de l'espace de régulation l'organisme peut atteindre à une étape donnée de son apprentissage. Confrontés à ces deux types de contraintes, les agents mémorisent les solutions efficaces et construisent des "modèles internes" des divers contextes exécutifs.[5]

Pour résumer ce que l'on vient de survoler trop rapidement, un épisode volitif consiste à activer un ensemble hiérarchisé de représentations exécutives. La volition de faire X en vue de P (tendre la main pour saisir un verre) dépend alors nécessairement de trois conditions:

(1) Contrôle existant: il existe dans l'espace de régulation accessible à l'agent pour ce contexte au moins une trajectoire de P à X (en vertu des lois de rétroaction) .

(2) Saillance présente: P est momentanément saillant (en vertu des lois de régulation).

(3) Motivation existante: la motivation présente pour obtenir P suffit à ce que l'agent se mette en état de produire X.

Notons que la condition 3 met à jour le soubassement fonctionnel de la définition donnée plus haut, selon laquelle la volition est l'effort de réaliser un changement en vertu de cet

5 On trouvera plus bas, dans la réponse à Pierre Livet, une analyse de la pertinence de la théorie de la viabilité pour comprendre l'action.

effort, les conditions 1 et 2 constituant des conditions de possibilité, respectivement générales et occurrentes, de 3. Observons ensuite que les trois conditions énoncées expriment des dispositions mentales et des propriétés occurrentes *qualifiant les relations avec l'environnement*. Elles n'impliquent pas nécessairement de propriété occurrente *extérieure* (la saillance peut être interne), ni non plus de changement corporel (le changement recherché peut être mental, même s'il covarie avec une propriété cérébrale). L'action corporelle est une extension de l'activité mentale en direction des effecteurs, qui crée une forme spécifique de rétroaction visuelle et proprioceptive absente des actions strictement mentales. Dans ce dernier cas, comme par exemple la tentative de se souvenir d'un mot, des rétroactions renseignent aussi l'agent sur ses chances de réussir: ce sont les sentiments épistémiques, comme le sentiment de savoir, ou le sentiment de familiarité.

Ces trois conditions sont nécessaires à toute forme de volition. Elles ne sont pas tirées de la psychologie expérimentale, mais de l'analyse mathématique des systèmes de contrôle adaptatif, et disposent de ce fait d'une parfaite généralité. Notons en particulier qu'elles ne déterminent pas la nature des mécanismes, c'est-à-dire des dispositifs, susceptibles dans chaque espèce organique de satisfaire les contraintes. Elles s'appliquent à un agent quelconque, animal, robot, dès qu'il vise un résultat et en lance l'exécution. Pour montrer que ces conditions sont suffisantes, il faudra vérifier qu'elles sont à l'abri de la déviance causale, dans les limites indiquées plus haut. Nous y reviendrons plus bas.

Il est maintenant possible de répondre succintement aux deux objections traditionnelles contre le volitionnisme. La définition proposée ne comporte aucune référence à l'action, comme le font certaines théories de la volition qui la *définissent* comme une action mentale, parce que le concept de contrôle est plus général que celui d'action. La condition 3 n'implique pas non plus de circularité: elle n'est dans notre définition qu'une des conditions de l'action, et non l'action elle-même. "la mise en état" n'est pas une action mentale, mais une opération.[6] La définition n'implique pas non plus de régression à l'infini, parce que le contenu d'une volition porte sur l'exécution de X en vue de l'effet P, et non sur la mise en état d'exécuter X.[7] Le cerveau de l'agent exploite les régularités dont il dispose pour placer l'agent dans l'état d'exécuter un changement. Cette mise en activité n'est pas décidé par un vouloir antérieur, il constitue le vouloir. Former la volition d'ouvrir la fenêtre n'implique donc pas, en abyme, le recours à une volition pour former cette volition.

Ces précisions sur la théorie volitionniste étant apportées, voyons comment elle peut répondre aux trois difficultés classiques des théories causales. Le problème de Malebranche, on l'a vu, trouve une réponse dans la structure télélogique de la volition. On sait que la structure associant la reproduction à la sélection par les conséquences dispense d'identifier explicitement les raisons qui ont causé la sélection. La loi de l'effet, étant téléologique, ne passe pas non plus par la mémorisation des causes de la sélection. Elle suppose simplement, dans sa version généralisée, que l'agent accède à une portion de l'espace de régulation (qu'il sache comment atteindre une cible quelconque), qu'il ait appris à discerner des saillances, et soit suffisamment motivé par elles.

---

6 Sur cette distinction, voir Proust (2001)

7 Il est intéressant de voir que cette dernière formule mobilise un théorème portant sur les conditions générales de régulation dans les systèmes mécaniques. A supposer en effet que la volition V1 suffise à atteindre l'objectif P, il ne peut avoir de volition V2 distincte de V1 qui soit sélectionnée pour atteindre l'effet recherché P, quand V2 n'y parviendrait qu'en activant V1. Car une règle de moindre action gouverne les stratégies de régulation, en vertu d'un théorème dérivant la moindre action de l'existence de symétries (Granger, 1979).

Dans la conception volitionniste, le problème de la déviance causale reçoit un traitement original. En effet, même si un démon malin intervient dans la commande (soit pour l'initier, soit pour la fausser), l'agent conserve la capacité de se réapproprier ou de rejeter la commande du fait de la structure dynamique de la séquence commande/comparaison intérieure/révision. De deux choses l'une, donc. Soit la volition de l'agent "reprend la main" en quelque sorte, comme le font les agents qui s'adaptent à un changement imprévu du monde, en révisant ou inhibant la commande "manipulée". Soit la volition reste également manipulée dans le suivi de l'action: l'agent est incapable de réviser et d'inhiber la dérégulation en cours. On doit alors refuser de considérer qu'il s'agit d'une volition, puisque la condition 3 ci-dessus n'est pas satisfaite. Cette fois, le refus n'est pas lié à une convention, mais à la structure réflexive du suivi de la volition.

Enfin le problème de Wittgenstein trouve une réponse, au demeurant très peu wittgensteinienne, dans une distinction fonctionnelle que la définition de la volition met en évidence: l'action est la volition A d'atteindre le résultat R en vertu de cette volition. C'est, en d'autres termes, la mise en état réflexive occurrente de produire un changement conduisant à R, selon une trajectoire contenue dans l'espace de régulation, avec une motivation occurrente d'atteindre R. Si R fait son apparition, comme l'accélération cardiaque dans l'exemple du yogi, alors qu'il n'existe pas de trajectoire possible dans l'espace de régulation occurrent pour atteindre R, ou si R n'est pas motivant, on ne pourra attribuer de volition d'obtenir R à un agent putatif. En revanche, si les trois conditions sont remplies, mais que, pour des raisons étrangères à ces conditions, le résultat R n'apparaît pas, l'action aura effectivement eu lieu, elle sera simplement non réussie.

La définition proposée admet que les volitions peuvent être déclenchées de manière inconsciente. Quel est alors le rôle de la conscience dans l'agir? L'ouvrage examine trois types de théories. Margaret Anscombe l'identifie à la connaissance non-observationnelle de ses intentions, laquelle permet de connaître immédiatement le contenu de ses pensées (conatives ou épistémiques). Le problème est que cette théorie n'explique pas la manière dont s'effectue la prise de conscience, ni par conséquent les illusions ou les erreurs qui peuvent y intervenir. Comment acquérir une connaissance sinon en accédant à une source d'information ? Si la source n'est pas observationnelle, quelle est elle ? La deuxième théorie est la conception inférentielle défendue entre autres par Dan Wegner, selon laquelle le sentiment d'agir dériverait de l'analyse rétrospective par le sujet d'un certain nombre d'indices prélevés pendant l'exécution de l'action. Cette conception a été expérimentalement mise à l'épreuve et rejetée par Pat Haggard et ses collègues. Haggard montre que la conscience de l'action est sensible à des paramètres temporels qui structurent l'exécution et la perception des effets. D'où la troisième théorie : la conscience de l'action est essentiellement perceptive. On perçoit sa volition d'atteindre un certain but dans le retour perceptif des boucles de contrôle associées. La conscience de l'effort, autre dimension de la conscience d'agir, passe aussi par des marqueurs perceptifs, construits quant à eux par des lois de rétro-action proprioceptive.

Le chapitre six s'intéresse à la question classique du libre-arbitre: la volonté peut-elle être libre ? On y pose la question dont dépend la position même du problème: la liberté est-elle ou non compatible avec le déterminisme des causes qui déterminent la volonté ? Les incompatibilistes répondent par la négative, soit pour rejeter la possibilité conceptuelle de la liberté, soit pour rejeter l'impact des causes sur la volonté. Les compatibilistes en revanche soutiennent qu'on peut trouver un sens à la liberté, que le monde soit ou non déterminé. Cette position a le mérite, du point de vue naturaliste, de rejeter le principe des possibilités alternatives: être libre ne veut pas dire que l'on aurait pu vouloir faire une autre action que

celle que l'on a voulu faire. Pour Harry Frankfurt, une volition est libre s'il y a conformité entre elle et une volition de second ordre, la seconde constituant la première comme la volonté de l'agent. Cette théorie séduisante soulève diverses objections, sur la nature des volitions de second ordre, sur le caractère momentané de la liberté du vouloir, et sur la cohérence entre le compatibilisme et la théorie proposée. Frankfurt échappe au risque de la régression à l'infini des niveaux de volonté en trouvant une dimension transversale, qui est la dimension affective, dans laquelle le sujet peut s'unifier. Mais cette solution rend difficile l'adhésion au compatibilisme. La deuxième objection qu'on peut opposer à Frankfurt met en évidence la vulnérabilité de toutes les solutions compatibilistes au subjectivisme: le sujet peut se croire libre sans l'être. Ces observations conduisent à donner du poids à la troisième objection. Il ne semble pas que le compatibilisme soit une position cohérente si on donne au concept de liberté une dimension non subjectiviste. Ce chapitre évoque en conclusion la possibilité de développer une approche strictement comparative de la liberté de vouloir, qui soit cohérente avec l'incompatibilisme. On dira d'un comportement qu'il est plus libre qu'un autre s'il fait appel à un système de régulation plus flexible, sans occulter le fait que toute décision de régulation est elle-même causalement déterminée.

Le chapitre final se tourne vers une question très familière aux lecteurs de Frankfurt : peut-on vouloir être une personne ? On montre d'abord l'impasse dans laquelle se trouvent les théories qui tentent de construire la personne sur la base de la seule conscience instantanée d'agir, de percevoir, ou d'avoir un corps. Puis on examine les théories mémorielles simples, qui identifient la personne à ce que la personne elle-même se souvient d'elle-même. Ces théories tombent généralement sous le coup de la circularité, et de l'argument leibnizien dit de la "duplication" : elles autorisent deux personnes numériquement distinctes à être la même personne si elles partagent leurs souvenirs. Dans la théorie mémorielle révisée, finalement retenue, l'identité personnelle se fonde non sur le seul recouvrement des souvenirs, mais sur l'exercice de la capacité d'auto-affection. La représentation de soi comme personne identique au fil du temps est la représentation motivante qui émerge du résultat de cet exercice de construction par sélection et révision des valeurs qui déterminent les décisions métacognitives, et ne lui préexiste pas. Mais, une fois en place, elle organise et motive de nouvelles formes de contrôle individuel et social, et de ce fait acquiert un rôle causal.

Réponses à Daniel Laurier, Pierre Livet et Stéphane Chauvier

Joëlle Proust Institut Jean-Nicod (CNRS, EHESS, ENS)

L'exercice consistant à répondre à trois critiques aussi avisés n'est pas sans risque, mais il s'est avéré infiniment stimulant et gratifiant. Je les remercie chaleureusement du temps qu'ils ont investi dans leur lecture et leur critique, et remercie ausi les éditeurs de la revue d'avoir programmé une disputatio sur *La Nature de la Volonté*. Puisse cette disputatio clarifier les enjeux méthodologiques et métaphysiques qui constituent l'arrière-plan des théories de l'action. Parmi mes trois critiques, seul Pierre Livet partage avec moi une conception radicale du naturalisme, en cherchant lui aussi à parvenir à une théorie de l'action compatible avec ce qu'on sait, non seulement du cerveau, mais des systèmes dynamiques auto-organisés que le cerveau instancie. Le débat avec Livet porte sur des sujets qui, quoique techniques, me semblent au cœur de la philosophie de l'action; il s'agit de comprendre les contraintes dynamiques des systèmes représentationnels évoluant dans des paysages changeants, et d'y inscrire une théorie du contrôle qui soit suffisamment générale pour

s'appliquer à l'action motrice et à l'action mentale. Avec Daniel Laurier, le terrain cesse d'être partagé ; la rigueur de ses objections suscite l'exigence symétrique d'affuter davantage mes arguments. Je lui dois l'occasion d'expliiciter la métaphysique de l'esprit agissant, restée implicite dans *La nature de la volonté* suite à une choix éditorial. J'ai également beaucoup apprécié de pouvoir m'expliquer sur la sémantique de la volition. Stéphane Chauvier enfin retrace les étapes de ma théorie de la volition en idiome hobbesien (ou watsonien ?), depuis de primitifs et mécaniques 'attrapages de pommes' jusqu'à l'apparition d'un sujet qui les veuille. La gageure était de le convaincre qu'il y a plus d'une façon d'être mécaniste, et que la théorie du contrôle peut remplir le cahier des charges de l'auto-affection et de la conscience d'êre identique à soi-même, sans rétablir l'homuncule ni sombrer dans le fictionalisme.

I – Réponse à Daniel Laurier[8]

Daniel Laurier s'intéresse tout particulièrement à la métaphysique de l'action. Il considère que je paye le prix fort d'avoir candidement opté pour l'image scientifique, au dépens de l'image manifeste du monde. Il relève que ma position est celle d'un « anti-réductionnisme qui a pour effet de ne laisser aucune place à l'efficacité causale des propriétés mentales ou des contenus intentionnels ». Les autres objections se répartissent en deux sections. La première examine *la théorie de la volition*. La deuxième section aborde les questions de *la liberté et de l'identité personnelle*. Ma réponse suivra l'ordre de ces objections, en commençant par expliciter la métaphysique de l'action qui inspire mon ouvrage, même si elle est restée, il est vrai, confinée à quelques passages.

1- Métaphysique de l'action : réductionnisme ou antiréductionnisme ?

Quelques philosophes, dont je fais partie, ont pris au sérieux les conséquences du rejet par Quine de la distinction entre propositions analytiques et synthétiques. On ne peut plus, dès lors, se borner à philosopher en fauteuil. Les concepts sont l'effet collectif d'une recherche indissociable des avancées empiriques. Comme d'autres ouvrages issus de ce rejet, La *Nature de la Volonté* se propose de construire une théorie philosophique de l'action compatible d'une part, avec l'expérience commune de l'agir, soit son « image manifeste », et, d'autre part, avec les apports des sciences cognitives, et en particulier des neurosciences, où des avancées sans précédents ont été faites durant la dernière décennie, soit son « image scientifique ».[9] A ce double système de contraintes, s'ajoutent les exigences proprement philosophiques du réalisme représentationnel. Il n'est pas évident d'avoir à faire de la philosophie en tenant compte de contraintes aussi diverses. Je tente, comme d'autres, depuis vingt ans de les gérer rationnellement, en réfléchissant sur les questions de méthode que cela pose à tous ceux qui ont fait le choix raisonné du naturalisme.[10] Ce préambule est destiné à désamorcer l'impression de « candeur élégante », c'est-à-dire de scientisme obtus, qu'a pu provoquer mon analyse à travers l'interprétation qu'en donne Daniel Laurier.

L'une des caractéristiques du débat métaphysique sur les rapports corps-esprit est de ne prendre en considération rien de ce que l'on peut savoir sur le corps ou sur l'esprit, ni même sans s'expliquer sur le concept de causalité utilisé, au profit d'une caractérisation abstraite de

---

8 Je remercie vivement Max Kistler d'avoir débattu avec moi des aspects métaphysiques de cette disputatio et stimulé ma réflexion. Les erreurs éventuelles restent les miennes.

9 On remarquera que, dans le cas de la philosophie de l'esprit, la tâche est de comprendre comment la première peut être comprise par la seconde, ce qui n'est pas le cas, par exemple, de la philosophie de la physique.

10 Cf. en particulier Proust (2004) et Proust & Pacherie (à paraître).

relations de survenance. La causalité invoquée est-elle le concept du sens commun, lié à l'explication causale, celui qui caractérise les pratiques du raisonnement scientifique, ou bien d'un phénomène objectif de transfert d'énergie entre des événements physiques? S'il s'agit de rendre compte de l'efficacité causale du mental, en adoptant une forme ou une autre de réalisme représentationnel, les formes purement attributives de l'explication ne nous concernent pas. Les rationalisations de l'action sont une récente acquisition de l'évolution de l'esprit. Il est fort probable que beaucoup d'actions ne sont pas structurées de cette manière, tout en satisfaisant des contraintes implicites de rationalité (Proust, 2006_a). Daniel Laurier passe à plusieurs reprises de considérations attributives à des hypothèses ontologiques. Or il n'est pas clair que nos propensions attributives (nos "intuitions" naïves) soient un bon guide pour identifier l'ontologie mentale. On peut légitimement douter que la structure moléculaire du cerveau, ni la neurophysiologie fonctionnelle (voir plus bas) suffisent à nous renseigner sur ce point. En revanche, la manière dont le cerveau se construit dans ses interactions avec le monde, au cours de la phylogenèse et de l'ontogenèse, peut nous orienter vers le découpage causal adapté, et nous éviter de commettre les erreurs les plus grossières sur la constitution de l'esprit. En résumant la métaphysique qui forme l'arrière-plan de cet ouvrage: la fonction causale des neurones est de transférer de l'énergie entre les structures physiques pertinentes (cerveau-corps-environnement) en conformité avec les contenus représentés, conformité qui suggère à son tour une intervention structurante de l'information sur la sélection neuronale.

J'en viens maintenant au raisonnement qui conduit à cette métaphysique. Comment le réalisme représentationnel s'applique-t-il dans le cas de l'action (où, on l'a vu, des représentations exécutives sont sélectionnées en vertu de certaines motivations) ? Le réalisme représentationnel peut inspirer deux types de théories de la causalité mentale ici pertinentes. Pour les unes, l'information dispose d'une causalité occasionnelle (qu'on peut qualifier aussi d'occurrente ou de déclenchante): le contenu mental est censé assumer en tant que tel un rôle causal *dans chaque épisode cognitif*.[11] Pour les autres, l'information possède un rôle causal *structurant* dans l'organisation neuronale - acquis au cours de la phylogenèse et/ou de l'ontogenèse. Une fois la structuration effectuée, l'activation d'un véhicule représentationnel procède par *transfert d'énergie* entre des assemblées neuronales (l'information n'intervient plus à titre de facteur causal *à ce stade*).[12]

Le métaphysicien doit tirer toutes les conséquences *du double niveau causal* dans lequel se construit une fonction représentationnelle donnée. Drestke a été le premier, à ma connaissance, à l'introduire, en distinguant une cause structurante d'une cause déclenchante. Millikan reconnaît aussi ce double niveau quand elle distingue la causalité reproductive, en vertu de laquelle un type d'élément se trouve copié et sélectionné par compétition avec des organismes qui n'en disposent pas, et la causalité dispositionnelle, en vertu de laquelle l'élément individuel produit des effets dans l'organisme où il a été copié.

Je propose de considérer la causalité structurante comme l'effet rétro-actif de besoins informationnels sur le développement du cerveau (à la fois dans la phylogenèse et dans l'ontogenèse). L'information a un rôle causal pour *sélectionner ceux des couplages physiques organisme-milieu qui sont favorables à la flexibilité adaptative*. C'est en effet parce qu'un organisme peut extraire et conserver de l'information sur les associations présentes dans l'environnement qu'il est également capable de prédire les événements favorables et nuisibles et de réagir comme il convient. La causalité structurante inscrit ainsi les contraintes informationnelles dans les propriétés générales d'une architecture mentale. A l'échelle de la phylogénèse, l'information constitue une affordance à exploiter: elle détermine, par exemple,

11 C'est la position de Dretske (1988), à qui l'on doit la distinction entre causalité structurante et déclenchante.
12 Voir Kistler (1999).

les divers dispositifs permettant à un organisme de percevoir des phénomènes d'une fréquence donnée, de mémoriser des événements ou des associations, de raisonner sur eux etc. Au cours de l'ontogénèse, l'information disponible contraint le développement du cerveau immature; que l'on considère comme Quartz et Sejnowski que les neurones se développent par l'interaction avec l'environnement, ou comme Changeux & Dehaene et Edelman, qu'ils se développent par sélection et destruction des neurones inemployés, la rétro-action de l'environnement permet au cerveau d'"internaliser" les caractéristiques spatiales et dynamiques de l'environnement, c'est-à-dire de les représenter à des fins anticipatrices.[13]

La causalité déclenchante est la manifestation occasionnelle des capacités représentationnelles construites au fil de cette double structuration. Afin de comprendre pourquoi l'information ne joue pas un rôle causal *déclenchant*, je propose l'analogie suivante. Un artisan conçoit un marteau en tenant compte des contraintes de la tâche: bois dur, métal dense et résistant, etc. Une fois les matériaux sélectionnés et marteau construit, nul besoin de faire intervenir une caractéristique fonctionnelle de second ordre. De même pour le cerveau. Une fois qu'un contrôle donné est profilé et structuré par les besoins informationnels, nul besoin d'identifier une étape où l'information jouerait un rôle nouveau à chaque utilisation. Le rapport entre les propriétés fonctionnelles et les proxys déclencheurs (une mise en correspondance de configurations neuronales), suffisent à assurer le rôle causal attendu. L'instrument étant calibré, l'"usager" peut exploiter les propriétés informationnelles acquises par le système d'absorber de l'information et de la traiter. Aucun rôle causal ne revientainsi à l'information sinon pour créer de nouveaux apprentissages, c'est-à-dire de nouvelles structures de contrôle. Sans quoi, effectivement on devrait renoncer au principe de l'exclusion explicative de J. Kim, et considérer que les structures neuronales (porteuses d'information), et les significations mentales immatérielles, se combattent en permanence pour influencer les comportements, ou que ces dernières sont purement épiphénoménales.

Une fois qu'un ensemble de neurones a été sélectionné dans la fonction de porter une certaine information, il résonne avec l'environnement de manière sélective. Si une assemblée neuronale donnée tire son efficacité causale de ses couplages informationnels avec l'environnement (eux-mêmes issus de couplages causaux impliquant une sélection et une reproduction, c'est-à-dire des couplages fonctionnels), on peut comprendre que *le même* état porte *de par sa fonction* une information spécialisée (propriété "mentale") et transfère de l'énergie (propriété "physique").

J'en viens à la question légitime de Daniel Laurier : l'ouvrage est-il ou non réductionniste ? Le réductionnisme est la thèse selon laquelle les propriétés mentales sont explicables soit par identification à des propriétés physiques, soit par mise en correspondance nomique avec des lois physiques, soit par la mise en oeuvre de formes plus faibles de réduction, comme l'émergentisme. L'anti-réductionnisme est la thèse selon laquelle les propriétés mentales ne peuvent pas recevoir ce genre d'analyse, en particulier du fait de leur multiréalisabilité.[14] L'hypothèse du réductionnisme par identité semble la plus facilement intelligible, et l'emporter sur les autres par sa simplicité métaphysique. Mais cette simplicité reste illusoire tant qu'on ne met pas à jour les relations nomiques entre les propriétés représentationnelles et les propriétés physiques de leurs véhicules, relations qui s'établissent nécessairement à des échelles temporelles multiples, et selon des mécanismes sélectifs très variés. Il ne paraît guère prometteur, en ce domaine, de se fier à des intuitions. La description

13 Sur ce point, voir "What is a mental function ? ", à paraître.

14 Max Kistler a restitué son attrait à la thèse de l'identité entre propriétés mentales et physiques, en argumentant que la réalisabilité multiple d'une propriété mentale ne constitue pas plus un obstacle à cette identité que la multiréalisabilité du concept de température n'est un obstacle à sa réduction à l'énergie cinétique des molécules. Voir en particulier Kistler (1999_b).

des "états mentaux" par les philosophes est biaisée par la représentation de ces états comme fixes, réplicables, et déterminés. Elle passe sous silence le caractère dynamique de la vie mentale, et la complexité du jeu causal où croissance et apprentissage se contraignent mutuellement.

Supposons, comme je l'ai proposé ci-dessus, que les propriétés physiques des événements mentaux soient les seules à être causalement déclenchantes, et que l'information portée par un indicateur joue un rôle causalement structurant dans la sélection des états physiques dont la fonction sera d'indiquer des états sémantiques. Cette fonction repose-t-elle sur des correspondances nomiques ? Cela peut paraître douteux. En effet, les relations fonctionnelles d'indication sont marquées par diverses caractéristiques architecturales, dont l'encapsulation générative des processus au cours dequels une fonction informationnelle est stabilisée à la fois par l'évolution et par les apprentissages. La multiplicité des paramètres, et la non-réplicabilité des contraintes d'architecture tant dans la phylogenèse que dans l'ontogenèse rendent l'idée de covariation nomique problématique.

En résumé, la position défendue est proche de l'anti-réductionnisme, si l'on entend par là que l'information forme une contrainte causale sui generis, liée à un certain type de bénéfice, la flexibilité cognitive. L'information joue un rôle causal structurant, et façonne l'organisation du cerveau – sa connectivité, ses activations synchrones – en formant un enjeu central de la sélection ou du développement neuronal. Mais la position s'apparente au réductionnisme en considérant que l'activité mentale *occurrente* dépend systématiquement des propriétés énergétiques des neurones mis en rapport avec des stimuli motivants. Symétriquement, la position n'est une forme d'épiphénoménalisme que si l'on s'intéresse à l'activité mentale occurrente. Mais il n'y a aucun épiphénoménalisme du mental si on examine la manière dont cette activité mentale a été simultanément sélectionnée et organisée à long terme. Loin d'être 'ad hoc', cette distinction dérive du fait que le cerveau est un système dynamique auto-organisé. Notons en passant que la question de l'épiphénoménalisme *de la conscience* est une question différente ; mais on peut utilement l'aborder à l'aide de la distinction proposée.

2- ***La théorie de la volition.***

a) La pertinence d'une théorie volitionniste face à une théorie intentionnaliste de l'action

Laurier objecte d'abord que le propos d'une théorie de la volition ne peut pas écarter la pertinence des théories de Davidson et de Searle, lesquelles portent non sur l'action volontaire, mais sur l'action intentionnelle. Si l'on prend l'argument de Laurier au sérieux, on devra dire que toute définition n'a qu'une portée nominale. De ce point de vue, les théories intentionnelles et volitives ne constituent effectivement pas des théories alternatives de l'action : les unes caractérisent les actions i*ntentionnelles*, les autres les actions *volontaires*. Pourtant, les divers auteurs qui ont réfléchi sur l'action cherchent à identifier ce qui définit l'action au sens courant du terme, quitte à introduire pour ce faire des raffinements ignorés du sens commun. Si l'on cherche à définir l'action par la structure des représentations qui la causent, le *definiendum* reste le même, qu'on le nomme, préthéoriquement, « action ordinaire », « comportement orienté vers un but », « action faite avec une intention », « action volontaire » ou « essai ». Ce n'est pas la même chose, évidemment, de faire de l'action un couplage entre raisons d'agir et changement, le produit réflexif d'une intention en action, ou d'une volition immédiatement exécutive. Mais ces différences touchent le sens, et non la référence du concept d'action.

Cela dit, toutes les définitions ne se valent pas. Contrairement à ce qu'affirme Laurier, la définition qui tient pour acquis que l'intention, ou la raison d'agir, disposent automatiquement d'une efficacité exécutive est de moindre généralité que celle qui manifeste la centralité à la

fois causale et sémantique de la sélection d'une représentation exécutive, qu'elle s'accompagne ou non de rationalisation.

b) La distinction entre comportements et actions

Daniel Laurier craint que, si l'on détache le mouvement de la satisfaction d'un désir ou de la réalisation d'une intention, comme le propose la théorie volitive, on n'ait plus les moyens de distinguer « les comportements/activités qui sont des actions, de ceux qui n'en sont pas ». Mais évidemment la théorie volitive donne les moyens de faire cette distinction fondamentale. Un comportement non voulu est un mouvement effectué sans volition, c'est-à-dire sans que se forme l'attitude que le langage commun traduirait comme « l'effort d'atteindre tel résultat en vertu de cet effort ». Cette attitude est analysée par les trois conditions évoquées plus haut (contrôle, saillance, motivation). Pour simplifier, si la représentation du but ne guide pas l'exécution du mouvement, le comportement n'est pas une action.[15]

Daniel Laurier considère que la théorie proposée conduit à la nécessité de distinguer, dans toute action ordinaire, des volitions involontaires et volontaires : « En quoi consiste l'action volontaire de bouger le bras ? Peut-être est-ce simplement l'action de vouloir que mon bras bouge, mais cette action ne peut être volontaire que si je veux l'accomplir, c'est-à-dire si j'accomplis une volition d'ordre supérieur ». Il trouve cette conclusion surprenante, et s'interroge sur la capacité des animaux à effectuer des « volitions volontaires ». La proposition qu'exprime cette conclusion est en effet surprenante, mais elle ne peut pas être dérivée de la théorie. De même que dans les versions intentionnelles de l'action, on ne parle pas d'une action qui ne serait pas intentionnelle sous au moins une description, on ne peut pas, dans une théorie volitionnelle de l'action, considérer qu'une action peut avoir lieu sans avoir été effectuée volontairement, au moins sous une description. L'effort d'atteindre le résultat X peut ne pas être couronné de succès, il peut atteindre en fait un résultat Y incompatible avec X. Mais cet effort de faire X (qualifié dans la triple condition décrite plus haut) est la caractéristique nécessaire et suffisante – dans la présente théorie - de l'action volontaire.

Le malentendu résulte peut-être de la confusion avec les théories « métareprésentationnelles » de la volition humaine, comme celle de Tim Shallice, ou de Dienes et Perner,[16]. De leur point de vue, la volition, au sens cette fois de programmation du mouvement, est involontaire si elle n'exige pas l'intervention d'un mécanisme de sélection consciente, et volontaire si elle implique la conscience explicite de la volition. La conscience de la volition suppose en outre, dans ces théories, que l'agent se représente son attitude volitive. Si l'on adopte ces prémisses, il s'ensuit que seul l'être humain est capable de « volition volontaire ». L'animal non-humain n'a en effet accès qu'à des schémas d'action routinière (contention scheduling system) sélectionnées par inhibition mutuelle : la représentation la plus active supprime ses compétiteurs. Seul un organisme doué d'un « Système attentionnel de supervision » *(Supervisory Attentional System)*, capable de métareprésenter ses intentions et ses volitions, peut utiliser les programmes moteurs du système de base en se libérant des préférences motivationnelles dictées par l'environnement.

---

15 Comme toute definition téléologique, on peut estimer qu'une action reste une action si elle ne parvient pas à son but. Le guidage de l'action est également une condition qui ne peut qu'être imparfaitement remplie, comme les deux autres. Ce problème peut affecter la capacité d'identifier un mouvement singulier comme l'action de faire P sur la base de l'observation de l'action, voire du compte-rendu sincère qu'en donne l'agent. Mais cette indetermination reflète la nature causale et fonctionnelle du dispositif, et ne constitue pas une difficulté conceptuelle.

16 Cf. Shallice, (1988), Dienes & Perner (2002).

La théorie proposée, on l'a vu plus haut, ne reconnaît pas la possibilité d'une volition non volontaire ; elle n'a pas non plus à considérer que l'animal est incapable de vouloir, parce que le sens de l'effort ne dépend pas de la capacité de former des métareprésentations. [17] Quoique la distinction entre volontés plus ou moins volontaires ne puisse être fait, il est parfaitement cohérent de hiérarchiser les formes de contrôle qui peuvent intervenir dans la volition dans la condition 1, en distinguant, en particulier, les contrôles instrumentaux (moyen-fin), les contrôles sui-dirigés métacognitifs, (Proust, à paraître) et les contrôles métareprésentationnels, axés sur la construction explicite de la personne que l'on veut être. La hiérarchie ne traduit pas seulement une libération de plus en plus grande des influences de l'environnement présent ; elle exprime aussi l'enchâssement des contrôles d'ordre inférieur dans les contrôles supérieurs. Cet enchâssement est un effet de l'architecture mentale, et non l'expression d'une redescription métareprésentationnelle d'un niveau par le niveau suivant.

c) La déviance causale

Selon Laurier, la théorie volitionniste ne répond pas mieux au problème de la déviance causale que les autres théories causales de l'action. Qu'en penser ? On l'a dit dans le précis, la solution qu'apporte la théorie de la volition au problème de la déviance causale consiste d'une part à utiliser la contiguité causale-représentationnelle[18] et la réflexivité procédurale entre la commande et le suivi de la commande pour intégrer la possibilité de perturbations : celles-ci peuvent être neutralisées par le suivi volitif. Il est donc parfaitement possible, contrairement à ce que dit Laurier, de spécifier de manière positive ce que doit être une chaîne causale non-déviante. Si, une fois la volition formée, intervient une perturbation causale dans la commande qui reste maîtrisable par rétroaction, la perturbation ne porte pas atteinte à l'action volontaire concernée, puisque les trois conditions sont finalement satisfaites. En revanche, si la perturbation rompt irrécupérablement la contiguité représentationnelle entre commande et suivi, alors la troisième condition n'étant pas satisfaite, l'action n'a pas la structure de la volition : on n'a pas, par conséquent, à renoncer à la suffisance de la définition. Ce n'est pas par convention que l'on parvient à cette conclusion, mais uniquement en utilisant le principe de la contiguité causale entre la commande et son suivi (lequel forme le soubassement ontologique de la définition). Ce caractère réflexif de la volition, hérité de la structure en boucle du contrôle, est accessible au jeune enfant et à l'animal non langagier, dépourvus l'un et l'autre de capacités mentalisatrices (sans "mindreading"). Il constitue une condition architecturale de possibilité pour tout système doté d'un système autonome de pilotage, robot ou animal.

d) Principe téléologique et « définition » de l'action

Laurier ne voit « absolument pas en quoi l'analyse des mécanismes de contrôle de l'activité musculaire et de leur mode de recrutement en accord avec la loi de l'effet (…) peut compter comme une 'définition' de la volition. » Les conditions 1-3 ne fournissent à ses yeux, au mieux, que des « corrélations nomologiques », mais « ne peuvent jeter aucune espèce de lumière sur les concepts d'action de base ou de volition, et relèvent d'une forme de réductionnisme que Proust semblait pourtant vouloir répudier ». On répondra, d'abord, que ce type de définition est déjà présent dans la littérature, puisque les concepts de fonction, de but, et de signification d'un symbole ou d'un énoncé, en ont déjà fait l'objet. Une définition téléologique, rappelons-le, expose la structure dynamique - l'histoire sélective - au terme de laquelle un effet est répliqué et sélectionné. La justification de choisir ce type de définition est que la structure téléologique de la causalité à l'œuvre dans la sélection par les effets est mieux

17 Sur le sens de l'agir, cf. Proust $(2003)_b$ et (2006_b); comp. avec le sens de savoir: Proust (à paraître).
18 Sur cette notion, cf. Mellor (1991) et Proust (à paraître).

adaptée à la nature du *definiendum*. La définition étiologique est optimale pour caractériser la sélection exécutive d'une régulation dans un contexte donné sur la base de son évolution.

On a déjà souligné que les conditions 1-3 sont parfaitement générales, et déterminent conceptuellement de manière nécessaire et suffisante ce qu'est une action, de la plus élémentaire à la plus sophistiquée. La définition de la volition est certes, spécifiable dans l'idiome de la psychologie ou de la neurophysiologie « et tout le bataclan », comme le dit Laurier, mais ce n'est pas d'elles qu'elle tire son intelligibilité; également applicable en robotique, elle doit son universalité à la théorie mathématique des systèmes de contrôle adaptatifs, dont il sera plus longuement question dans la réponse à Pierre Livet.

L'usage critique du terme ambigu de "mécanisme" mérite d'être clarifié dans ce contexte ; dans l'emploi où il désigne des dispositions physiques réalisant un dispositif fonctionnel, il relève d'une science particulière, comme la neurophysiologie. Mais s'il désigne le dispositif fonctionnel lui-même, il relève de toute discipline s'intéressant au mental, ou cherchant à définir les états mentaux par leurs relations fonctionnelles, comme la philosophie de l'esprit, la psychologie « rationnelle » ou l'anatomie cérébrale fonctionnelle. Si les boucles de contrôle constituaient effectivement "la base corporelle" de l'esprit, on pourrait parler de réductionnisme pour caractériser la présente tentative de définir le concept de volition. Mais les boucles de contrôle constituent un mode d'organisation fonctionnel des neurones, pas une base de réduction.[19] Un philosophe peut donc avoir des raisons de s'intéresser à la psychologie ou aux neurosciences : non pas pour tirer servilement de ces disciplines les définitions dont il a besoin, mais à titre heuristique : pour identifier la structure fonctionnelle pertinente pour la définition recherchée.

e) Compatibilisme, incompatibilisme et liberté comparative

Le dilemme du compatibilisme est le suivant : soit le déterminisme est faux, et dans ce cas on ne peut plus construire la liberté de la volonté sur l'oganisation causale du vouloir ; soit il est vrai, et dans ce cas les volitions, comme les autres attitudes, sont le produit causal d'autres états, et ne peuvent être dits compatibles avec la liberté qu'en changeant le sens du mot « liberté » d'une manière à servir l'argument, c'est-à-dire par pétition de principe. Daniel Laurier trouve beaucoup à redire à cette formulation. Je traiterai des deux principales objections. La première est que « l'hypothèse indéterministe ne semble pas éliminer la pertinence de faire référence à l'organisation interne de la volonté de l'agent ». A ceci on répondra que si par organisation interne, on entend une structure causale liant une attitude et une exécution, on voit mal comment cette organisation sortirait indemne de la désorganisation causale associée à l'indéterminisme.

La seconde est que, ce même chapitre objectant au compatibilisme l'annexion indue du terme de liberté, il paraît curieux de le rétablir en parlant de liberté comparative. Le débat devient ici nominal, car l'ouvrage est clair sur l'interprétation qu'il convient de donner au terme de liberté comparative et sur la motivation de ne retenir qu'un tel sens comparatif, quand on soutient que l'agir n'échappe pas au déterminisme causal (étendu à ses formes probabilistes). Les pages 245-251 présentent les raisons à l'appui de ce choix terminologique ; elles clarifient en quel sens résiduel non illusoire on peut encore comprendre le terme, et le rôle causal que joue l'impression de liberté dans le cadre de l'apprentissage de l'autonomisation à l'égard des saillances présentes.

e) Sommes-nous des épiphénomènes ?

19 f. Proust et Pacherie, (à paraître).

Dans cet ouvrage, je défends de nouveau la thèse, déjà argumentée dans Proust (2000), que l'identité personnelle se construit dans un mode de contrôle particulier – l'évaluation et la révision de ses buts mentaux. A la faveur de ce contrôle, l'esprit se modifie lui-même et, simultanément, élabore l'identité de la personne. L'exercice de ces actions mentales, mis en œuvre au fil du temps, joue ici le rôle central. Daniel Laurier estime « assez troublante à première vue » la conclusion qu'il tire de ce dernier chapitre, selon laquelle « nous sommes davantage les produits d'actions impersonnelles que les auteurs de nos propres actions, et que nous ne faisons, littéralement, jamais rien ». Il joue évidemment ici sur un double registre : il retraduit les concepts dans un langage familier, celui de la psychologie ordinaire, et construit ainsi une version révisée des thèses, où elles perdent toute intelligibilité. En effet, dire que les actions sont « impersonnelles » suppose que cela ait un sens de dire qu'une action est soit personnelle, soit impersonnelle. Il existe une troisième possibilité, qui est de dire qu'une action n'est ni l'un ni l'autre, dans la mesure où la volition et la personne ne peuvent pas, en vertu de l'ontologie du mental, être dans un lien d'agentivité positive ou négative (de même qu'elles ne sont ni des nombres pairs ni des nombres impairs).

La notion d'« agent », quelque cruciale qu'elle soit pour communiquer avec nos semblables et leur attribuer des rôles et des responsabilités, comme l'ont remarqué beaucoup de philosophes de l'action contemporains,[20] n'est pas non plus la notion pertinente pour expliquer la cause d'une action. L'agent peut-être le siège, ou le représentant social, de la causalité des attitudes propositionnelles, mais il n'est pas, à strictement parler, la cause proximale de « son » action. Ici encore, les besoins de la psychologie ordinaire ne sont pas nécessairement ceux du philosophe, et la traduction du raisonnement philosophique en idiome du sens commun est semée de pièges. Mon propre texte n'y échappe pas, chaque fois que je sollicite la psychologie ordinaire à des fins de clarté (voir par exemple l'énoncé de la condition 3 mentionnant « l'agent »).

Réponse à Pierre Livet

Pierre Livet a commencé avant moi à s'intéresser aux analyses naturalistes de l'action, et il adhère aussi depuis longtemps au programme de recherche consistant à tirer conceptuellement parti de la théorie du contrôle. Relèvent entre autres de ce programme ses travaux anciens sur les systèmes auto-organisés et ses recherches sur la dynamique des émotions. Les questions qu'il pose sont très pertinentes, et montrent la profondeur de sa progression dans un territoire encore peu exploré philosophiquement. La première concerne l'interprétation qu'il convient de donner aux résultats de Patrick Haggard concernant la dynamique de la conscience d'agir. Pour que l'on puisse considérer que ces résultats réfutent la conception antiréaliste, attributive, de la conscience d'agir, il faudrait , dit Livet, « que la conscience d'être à l'origine de l'action puisse tenir une place comparable dans l'expérience de l'action volontaire et dans celle de l'action involontaire ». Si ces deux formes de conscience diffèrent phénoménalement, selon Livet, on peut rétrospectivement s'appuyer sur la différence pour s'attribuer ou non la volition correspondante. Dans les deux théories, ajoute Livet, « la conscience peut rester épiphénoménale. » Une première façon de comprendre l'argument mène droit à une difficulté: si la conscience est supposée rester épiphénoménale, elle ne peut pas jouer de rôle causal ; de ce fait, l'agent ne peut pas s'appuyer sur la différence

20 Voir entre autres, Davidson, 1980, Goldman, 1970, Dennett 2003.

de ce que cela fait d'avoir ou non préparé l'action. L'argument est probablement plutôt que la conscience pourrait ne jouer strictement aucun rôle dans les expériences de Haggard, et que seule l'activation neuronale de la préparation détermine la manière dont le sujet rapporte son expérience, et lie temporellement le début du mouvement et les effets observés du mouvement.

On peut répondre à cette objection en deux temps. Le premier temps concède l'argument, et admet qu'on ne peut rien conclure de l'expérience de Haggard sans rejeter la possibilité de l'épiphénoménalisme de la conscience. La stratégie de réponse pourrait être la suivante. L'épiphénoménalisme de la conscience – comme l'épiphénoménalisme du mental (voir la réponse à Daniel Laurier) - peut être structurant ou occurrent[21]. Les signaux conscients par lesquels nous sommes couplés à l'environnement sont de type structurant. Pour cette raison, ils ont un rôle causal à jouer et ne peuvent être épiphénoménaux. Deuxième temps : cet argument est-il effectivement nécessaire pour justifier l'interprétation par Haggard de ses résultats? En fait, on peut en douter. Si l'objectif de Haggard est seulement de révéler la structure causale de la volition, consciente ou non, ses résultats suffisent à appuyer sa conclusion : rien ne permet de comprendre la différence de liage dans les conditions active et passive sinon l'intervention active du contrôle volitif. La préparation de l'action change la réception perceptive des conséquences de l'action. La théorie attributive de Wegner, en revanche, suppose qu'aucune différence de liage n'intervienne *dès le début de l'action*, puisque l'agent est censé s'attribuer l'action sans s'appuyer sur l'effectuation proprement dite (mais seulement à partir d'indices attributifs « humiens »). Que les réponses faisant l'objet d'un rapport subjectif soient celles d'un zombie ou non, on peut donc trancher en faveur du rôle causal, et non seulement attributif/rétrospectif, du contrôle volitif.

A la différence de ce que suggère Pierre Livet, on ne peut opposer le contrôle et la volition, dans la théorie que je propose. Dans la mesure où la théorie de la volition met essentiellement en jeu les conditions 1-3, où est admise la possibilité d'un effort représenté, mais non conscient en vue d'atteindre P, l'épiphénoménalisme de la conscience ne constitue pas en soi une véritable difficulté pour la théorie. Si cet épiphénoménalisme est intenable, c'est pour d'autres raisons, que je ne peux pas exposer ici.[22]

La deuxième source de perplexité de Pierre Livet concerne la pertinence du schéma de contrôle propre à la motricité pour les autres niveaux de contrôle envisagés dans l'ouvrage, comme celui du contrôle épisodique de l'action, du contrôle métacognitif et du contrôle de ses propres engagements éthiques.[23] Je suis heureuse de pouvoir approfondir ici cette question, inabordable dans un petit ouvrage pédagogique. Il est vrai que la théorie de la viabilité constitue à mes yeux la base du raisonnement philosophique sur le contrôle ; cette théorie a la généralité requise pour pouvoir s'appliquer au-delà du cas moteur, aux formes les plus sophistiquées de raisonnement réflexif et de l'action mentale. Le premier avantage de cette théorie mathématique, est qu'elle fournit les principes de toute forme de régulation, en démontrant l'intervention constitutive, pour tout contrôle, de lois de régulation et de lois de rétro-action. Le deuxième est qu'elle permet de caractériser le couplage entre une entité et un environnement, l'un et l'autre changeants, en proposant une alternative globale aux calculs probabilistes généralement utilisés pour décrire chaque type de couplage. L'idée est en effet

21 Une théorie intéressante, quoique inhabituelle, consiste à distinguer la conscience phénoménale de la capacité représentationnelle. C'est pourquoi il convient d'étudier séparément le rôle causal et la réductibilité, dans le cas de la conscience et et dans celui des representations.

22 Sur cette question, cf. Proust & Frankowska, en préparation.

23 On trouvera trois de ces contrôles résumés dans la figure 7.1, p. 296 de l'ouvrage.

de s'intéresser au repérage, par un agent, des limites du noyau de viabilité,[24] soit de décrire l'ensemble des cas où la régulation doit être modifiée pour maintenir l'intégrité du système. Elle permet aussi de généraliser le problème du modèle inverse par la notion de bassin de capturabilité. Du fait que cette théorie décrit le processus de la prise de décision rationnelle sous incertitude, elle est en mesure de jouer un rôle conceptuel central dans la compréhension de toutes les formes d'action.

La théorisation dynamique qu'offre la théorie de la viabilité me paraît compatible avec le recours à des représentations pour mémoriser les régulations. Cette idée a déjà été défendue sur le fond par deux théoriciens du contrôle adaptatif « traditionnel », Conant et Ashby, en 1970. Dans un système optimal, ont-ils démontré, les actions du régulateur « ne sont rien d'autre que les actions du système considérées sous une certaine mise en correspondance ». Il s'ensuit que les lois de régulation gouvernent des simulations représentant les dynamiques possibles du système. Une idée voisine a, plus récemment, été explorée dans les neurosciences de l'action, sous l'appellation de "modèles internes". Le cerveau (cortex pariétal) anticipe par simulation les effets proximaux et distaux des mouvements. Pour préparer une action, un modèle inverse est calculé pour sélectionner la meilleure trajectoire, (ou une enveloppe satisfaisante de trajectoires), pour atteindre un but. Les modèles directs simulent ensuite le déroulement du programme d'action sélectionné pour comparer le feedback observé au feedback simulé.

On l'a vu plus haut, ces concepts sont d'une très large application, et ne se cantonnent pas à l'action motrice.[25] Ils s'appliquent aussi bien, à des échelles de temps très diverses, aux décisions mémorielles, perceptives, financières, amoureuses, etc. Ils apportent, à mon sens, une clarification notable au rôle des révisions présidant à la constitution du soi, et à la différence entre la dimension motivationnelle-volitive et la dimension épistémique du rapport à soi. Si Conant et Ashby ont raison, le format simulatoire des lois de régulation et de feedback devrait donc reparaître, avec des contenus différents, identiquement d'un niveau à l'autre. Par exemple, dans le contrôle métacognitif, le sujet qui s'interroge sur ce qu'il a en mémoire simule la récupération, et extrait de cette simulation l'information nécessaire à la décision de poursuivre ou non sa recherche. Il n'est pas exclu que cette action mentale suppose elle aussi un modèle inverse, c'est-à-dire un principe de sélection de la commande optimale pour récupérer le souvenir recherché étant donné le contexte. Une fois le modèle inverse trouvé (mettre ne oeuvre la bonne procédure étant donné l'objectif), reste à apprécier la faisabilité temporelle du souvenir dirigé. Il est possible qu'un modèle direct, représentant l'anticipation des contraintes de validité pertinentes, prenne ensuite le relais comme dans le cas de la motricité. Ce qui permet de faire cette hypothèse, c'est que les sentiments épistémiques qui structurent les rétroactions métacognitives, sont en tous points analogues aux marqueurs proprioceptifs qui structurent les rétroactions des actions corporelles. Les formes de contrôle les plus élevées, c'est-à-dire acquises les plus récemment, - comme celle qui préside à la stabilisation de l'identité personnelle - sont également plus lentes à s'établir, dans la mesure où les lois de rétro-action s'y installent moins par des contraintes développementales très précoces (comme le fait l'action motrice) que par le jeu d'influences complexes individualisées: apprentissage par observation, imitation, témoignage, roman, film, etc.

---

18 Le noyau de viabilité est l'ensemble des états initiaux d'où part au moins une évolution viable dans l'environnement concerné, une évolution viable étant une évolution compatible avec l'intégrité du système.

25 L'étude des perturbations du contrôle chez les patients atteints de schizophrénie montre que, malgré la différence entre l'agir moteur et l'agir mental, les analogies de contrôle laissent penser que le caractère moteur d'une commande ne constitue pas les limites d'une espèce naturelle. (cf. Proust, 2006). Le repérage actif/passif est une caractéristique architecturale commune toutes les formes de contrôle.

Pierre Livet pose des questions importantes, concernant la manière dont nous pouvons arbitrer entre des valeurs incompatibles entre elles. Il est vrai que ce territoire commence à peine à être exploré par les neurosciences, et que nous savons peu de choses sur la manière dont ces arbitrages sont conduits.[26] On ne peut évidemment pas conclure de notre présente ignorance du processus de décision en situation d'incertitude concernant les valeurs, à la liberté du sujet moral. Je suis en accord sur ce point avec Livet: un sujet dont le champ des valeurs est plus ouvert, est un sujet plus libre, de même qu'un homme qui peut marcher, est en un sens plus libre que celui qui n'a pas cette option. Ce n'est pas parce que la causalité devient probabiliste qu'elle cesse d'opérer,[27] et ce serait trahir Livet que de lui attribuer un tel argument, puisqu'il parle seulement de liberté "comparative". Le théoricien n'est pas outillé pour prédire si tel agent va opter pour telle décision; mais du fait qu'elles impliquent une forme de contrôle, les capacités de décision sont solidaires et de leur réalisation physiologique et des apprentissages informationnels/motivationnels qui les justifient. En insistant sur le fait que la hiérarchie des niveaux de contrôle n'est jamais "complètement déterminée", Pierre Livet, tel que je le comprends, souligne le fait que la prise de décision peut susciter à son tour des explorations nouvelles, au lieu de dépendre d'apprentissages passés non modifiables. Cette ouverture toutefois est elle-même apprise, une activité pour laquelle l'agent est redevable à son développement et à son environnement.

III - Réponse à Stéphane Chauvier[28]

La reconstruction par Stéphane Chauvier du parcours philosophique suivi dans *La nature de la volonté*, reflète une lecture très attentive. Elle est globalement juste, en dépit du fait qu'il n'en partage évidemment ni les prémisses, ni les conclusions - mais le diable est dans les détails ! Redessinant scrupuleusement la hiérarchie des formes de contrôle dans lequels peut s'exercer une volition, il glisse dans sa description des éléments parodiques savoureux. Il dépeint un univers de basculements mécaniques, un peu comme celui qui affecte la glande pinéale de Descartes: des basculements sans subjectivité, des basculements impersonnels et glacés: c'est du Hobbes, estime Chauvier, et l'on pourrait ajouter dans le même esprit, du Hobbes relu par La Mettrie avec un zeste de Skinner. L'enjeu est de taille, puisqu'il s'agit de sauver le sujet de ce qui apparaît comme sa disparition programmée. Je tâcherai dans ce qui suit de convaincre, sinon Stéphane Chauvier, du moins les lecteurs que le naturalisme ne "désenchante" pas, que les systèmes de contrôle ne sont pas des systèmes hobbesiens, et qu'en particulier, chez l'homme, ils offrent des moyens réflexifs puissants et efficaces d'auto-attribution de l'agir, bien avant que le concept de "je" soit maîtrisé, et que l'identité

---

26 Voir Sugrue et al. 2005

27 Le terme de "déterminisme" a plusieurs significations qui risquent de provoquer la confusion dans le présent débat. En théorie des systèmes de contrôle adaptatif, la notion de déterminisme est liée à la nature des évolutions. Soit un *système évolutionnaire,* c'est-à-dire un système qui associe à tout état initial un ensemble d'évolutions de variables d'état partant de cet état. On dit de ce système qu'il est « déterministe » si à chaque état initial correspond une et une seule évolution. Si plusieurs évolutions de variables d'état sont associées à chaque état initial, le système est dit « indéterministe ».(Voir Aubin et al., 2005). Mais cet indéterminisme ne fait que qualifier l'état de nos connaissances quant à l'évolution ultérieure du système. Les évolutions restent objectivement dépendantes de paramètres causaux. On reste donc dans le registre de la causalité probabiliste, et non dans celui de l'indéterminisme au sens philosophique.

28 Je remercie Frank Esken et Anna Loussouarn d'avoir participé à la réflexion préalable dont cette réponse est issue.

personnelle soit constituée. Je renvoie le lecteur intéressé par cette distinction à l'expérience de pensée des trois agents, présentée pp 275 et suivantes.

Revenons d'abord à des considérations de méthode. Les questions dont nous traitons, l'action, l'identité personnelle, sont au coeur d'un débat qui divise les philosophes analytiques en deux camps. Ceux qui, à la suite de McDowell, clament l'autonomie explicative du niveau personnel, et ceux qui, de Peacocke à Dennett, admettent que les états subpersonnels peuvent être porteurs de contenus mentaux, et rejettent donc l'autonomie du niveau personnel.[29] *La nature de la volonté* s'inscrit délibérément dans le camp des anti-autonomistes, et se donne des contraintes qui sont généralement étrangères aux autonomistes, en particulier de comprendre ce qui forme le socle de toutes les formes d'actions, humaine ou non: leur structure causale-téléologique. La définition proposée, évidemment, doit aussi satisfaire certaines des exigences issues du "niveau personnel": elle doit rendre compte de la conscience que l'agent prend de son action; ce versant de l'explication ne peut évidemment se borner à expliquer la conscience "normale" de l'action; elle doit inclure aussi les perturbations de la conscience d'agir, en particulier dans les troubles schizophréniques ou autistiques. Elle doit aussi être compatible avec le fait que l'action est rationnelle, et, dans le cas humain, est explicitement rationalisée.

Ces différences de méthode entraînent une conception différente de l'*explicandum*. Pour Stéphane Chauvier, la définition de l'action doit mettre en évidence "la pente qu'ont les agents humains à se poser comme les *auteurs* de leurs actions". Pour le naturaliste, ce que les humains pensent d'eux-mêmes lorsqu'ils agissent, la contribution que font les énoncés en première personne à cette pensée, et leur disposition à s'attribuer une action, doivent *aussi* être *expliqués – et non seulement mentionnés dans une explication "autonomiste" par les raisons d'agir*. On ne peut évidemment trouver dans ce petit livre qu'une esquisse des travaux philosophiques effectués dans ce domaine, relatifs à la proprioception de l'action, à l'image du corps, au sens de la possession subjective (sense of ownership), au sens de l'agir (sense of agency), et au rôle du langage dans la compréhension de soi. Mais ces travaux sont assez avancés pour qu'on puisse se former une idée relativement précise de la structure du vouloir et des formes de la conscience d'agir associées aux diverses organisations temporelles de l'action.

La section de l'ouvrage intulée "Qu'est-ce qui détermine la conscience de la réflexivité de l'action"? (p. 194) expose les conséquences sémantiques de la contiguité causale. Cette section permet de répondre en partie à la question ci-dessus, concernant l'auto-attribution de l'agir. Cette auto-attribution repose sur la réflexivité fondamentale de la volition, comme effort de parvenir à une cible en vertu de cet effort. Mais on doit bien convenir que cette auto-attribution n'est pas toujours *explicitable* ni même *revendiquée* par le sujet – soit parce qu'il est non-verbal, soit parce qu'il souffre d'une pathologie mentale, soit parce que l'action a été faite de manière automatique, sous une impulsion exogène. Il ne cesse pas pour autant d'effectuer le contrôle de son action de manière pratique, indice qu'il conserve la *motivation* d'atteindre le résultat et *sait comment* procéder aux ajustements nécessaires.[30]

---

29 Cf. Proust (2004). Nous revenons sur ce débat dans Proust et Pacherie, à paraître.

30 Chauvier commente ainsi la page 198 : "le sujet est pathologiquement concerné par ce qu'il fait". Il ne s'agit, peut-être pas, malgré les apparences, d'une observation ironique. Le problème que cherche à résoudre ce passage est qu'un patient atteint de schizophrénie peut nier être l'auteur d'une action, tout en conservant l'impression subjective de l'agir, liée au fait que sa volition déclenche des rétro-actions attendues et observées (dans sa proprioception, etc.). Sur le fond, la conclusion tirée par Chauvier ("un organisme vivant se tire mieux d'affaire qu'un zombie") suppose que la conscience de la motivation soit indissociable du rôle fonctionnel de la motivation. Cette position est peut-être correcte, mais elle ne fait pas l'unanimité.

A quelles conditions une volition est-elle alors revendiquée par l'agent comme étant "la sienne?" La phrase est très ambiguë: ainsi formulée, elle peut concerner autant l'effort d'une mouche pour rester orientée sur sa cible, que l'effort d'un politicien pour faire reconnaître ses propres mérites dans les affaires publiques. La théorie proposée consiste à montrer que l'auto-attribution se fait en diverses étapes de sophistication croissante, associées chacune à un système de contrôle spécialisé. Pour simplifier, la qualification d'une action comme "mienne" peut s'articuler primitivement par le couple actif-passif, puis par le couple planifiable (et révisable) - non planifiable (non révisable), enfin par l'opposition identitaire entre moi, et telle autre personne. Ces trois plans concernent respectivement le contrôle instrumental de l'action, le contrôle métacognitif, et le contrôle social.

La question se pose évidemment de savoir le rôle qui revient à la conscience dans le sens de l'agir *de chaque niveau*. J'ai montré ailleurs que la conscience phénoménale de la réflexivité et des intensités qualitatives doit minimalement être présente chaque fois qu'un contrôle de l'agir est en place,[31] mais l'ouvrage ne fournit pas les arguments qui étayent cette thèse. Je concède volontiers à Chauvier qu'il aurait été utile de montrer par des analyses de détail que le modèle volitionniste proposé s'applique aux actions essentiellement symboliques, comme "l'action de se marier ou de pacifier un conflit", qui font appel, en complément de la compréhension des relations instrumentales moyens-fins et de la maîtrise du contrôle métacognitif, à une théorisation sur l'esprit.

Il me reste à tenter de répondre à la question centrale du commentaire de Stéphane Chauvier: "Que peut bien signifier le fait que l'agent, non pas ait conscience de s'auto-affecter, mais ait conscience de pouvoir s'auto-affecter ?". Rappelons le contexte. Il s'agit de découvrir la condition de possiblité de la réflexivité forte, associée à l'usage quasi-indexical du mot "je" (p. 262) . La disposition métacognitive à évaluer ses propres capacités mentales et à les réviser, s'avère remplir la condition recherchée:

> "Le sens d'être soi (..) réside dans la conscience de pouvoir s'affecter, c'est-à-dire dans le souvenir de s'être affecté joint à la conscience d'être en mesure maintenant, de le faire" (p. 274).

Il faudrait en toute rigueur, se souvenir d'un aspect de l'argument que cette phrase risque de masquer, à savoir que le sens d'être soi ne dépend pas *seulement* de cette capacité métacognitive, sans quoi l'animal non-humain, qui est pourvu de cette capacité, aurait une conception de soi; il dépend aussi de la capacité de théoriser sur soi, que l'animal non-humain, pour autant qu'on le sache, n'a pas. Mais le chapitre développant plusieurs sortes de théories mémorielles, l'arrière plan métareprésentationnel (je me souviens que j'ai fait P en t) est commun à toutes les théories discutées, et ce qui est en jeu est de comprendre, non l'attribution de l'agir à soi, mais la conception de soi comme restant le même au fil des actions. C'est parce qu'on s'auto-affecte conformément à une norme, comme le suppose la structure de la métacognition, que l'on dispose, une fois la métareprésentation en place, de la possibilité de se (méta)représenter comme persistant au fil du temps. Pourquoi? Parce que la réflexivité de la modification fait que la référence du mot je est induplicable. On ne peut contrôler que ce qui est dans son propre espace de régulation, et l'on ne peut pas copier une régulation, parce qu'elle contient l'ensemble de son contexte, et qu'elle évolue en fonction de ses propres effets contigus.

Comment alors passer du *souvenir conscient de s'être auto-affecté* à la conscience de *pouvoir de le faire* ? Si l'on s'en tient à la séquence inférentielle entre ces deux états mentaux, on peut observer comment, d'une métareprésentation (mémorielle) sont dérivées successivement deux propositions:

1. Je me souviens de façon véridique m'être auto-affecté.

---

31 Cf. Proust (2006)$_b$

2. J'ai pu, dans le passé, m'auto-affecter
3. Je peux maintenant m'auto-affecter

La proposition (2) est une conséquence logique de (1), tandis que la proposition (3) est une généralisation empirique de (2), banale dans le raisonnement dispositionnel. Chauvier ne conteste pas la validité de cette généralisation d'ordre théorique, mais s'interroge sur le sens de l'énoncé "j'ai conscience de pouvoir m'auto-affecter". Il n'a de sens dans l'argument en cours, selon lui, que duement complété. Soit on entend par là: "j'ai conscience de pouvoir m'auto-affecter volontairement". Accorde-t-on cette interprétation, on est, selon Chauvier, obligé de convenir que l'agent dit déjà "je" et se pose déjà comme l'auteur de ses actions. Si en revanche, l'énoncé veut dire "j'ai conscience de pouvoir m'auto-affecter involontairement", alors on doit renoncer à la réalité de la notion de personne.

L'objection pourrait être dévastatrice si la volition de s'auto-affecter pouvait être volontaire ou involontaire; or on a vu plus haut (réponse à Laurier, section 2 b) que l'on ne peut dire qu'une volition soit autre que l'effort d'atteindre un résultat, et cela, en amont de toute considération attributive. En outre, on ne doit pas méconnaître l'erreur de perspective sur le problème posé que crée l'importation anticipée des métareprésentations dans les propositions 1-3. Il est plausible qu'un apprentissage implicite de la compétence métacognitive, par le truchement des sentiments épistémiques qui signalent le succès ou l'erreur dans ce domaine, précède (dans l'ontogenèse et dans la phylogenèse) l'auto-attribution explicite de souvenirs portant sur ses propres actions mentales. On peut ainsi former le savoir faire pratique lié, par exemple, à la métamémoire, sans disposer du concept de soi-même.

Comme j'ai tenté de l'expliquer dans le livre, la réflexivité mise en jeu dans la métacognition est un effet sémantique de l'architecture de contrôle: tout système ayant la propriété de contiguité causale – come celle qui relie la commande et le suivi -, produit, s'il est associé à une sémantique, des constituants inarticulés, à valeur réflexive, tels que "ici", "maintenant", "soi-même", ou, en théorie de l'action: "en vertu de cet effort/de cette intention/de cette volition" etc. L'auto-affection peut donc être pensée sans concept de soi, et l'auto-affection volontaire peut être le produit de la volition sans avoir à s'inscrire dans l'opposition conceptuelle volontaire/involontaire.

Stéphane Chauvier a raison: je ne peux pas me laisser enfermer dans cette antinomie. Etre une personne, selon moi, n'est pas une fiction. Pourtant, je n'attribue pas à la personne le rôle de l'agent. La personne n'agit pas plus que l'agent n'agit, si l'on entend par là qu'il/elle intervient en tant que véhicule dans un transfert d'énergie causant un résultat. La personne est la représentation dynamique qui forme *la cible* des volitions impliquées dans l'auto-affection. Cette représentation joue un rôle normatif et causal capital dans les délibérations, les planifications et les volitions concernant la vie sociale. La représentation par l'agent de soi-même comme personne identique au fil du temps unifie l'expérience mémorielle autour de valeurs. Elle conditionne et motive les actions "symboliques" ou stratégiques. En outre, elle survient sur des dispositions efficientes à agir mentalement. Un agent impulsif à la Frankfurt, - Oberov dans mon expérience de pensée des trois agents (p. 279),- n'a pas les moyens de former la représentation de soi-même. On ne peut construire de réflexivité personnelle là où n'existe pas de contrôle métacognitif.

Loin d'être un "*flatus vocis"*, le concept de personne est un déterminant causal de l'architecture mentale de première importance pour la régulation sociale. La difficulté de mon critique est d'accepter l'ontologie proposée. Si le "je" est une forme de régulation qui redécrit dans le langage du temps personnel la réflexivité propre au contrôle, pour le plus grand bénéfice individuel et collectif, il n'y a pas lieu de le soumettre au scalpel pour y retrouver la chair et le sang, ni de se plaindre de ne rien trouver. L'intelligibilité de la superposition des

contrôles est, indéniablement, moins immédiate, et moins riche en émotions associées, que celle de la psychologie ordinaire dans laquelle nous sommes bercés depuis l'enfance.

Références

Anscombe, G.E.M.1959. *Intention*, Oxford, Blackwell.

Aubin J.-P., Bayen, A., Bonneuil, N. & Saint-Pierre, P. .2005. *Viability, Control and Games: Regulation of Complex Evolutionary Systems Under Uncertainty and Viability Constraint*, Berlin, New-York: Springer.

Changeux, J.P. & Dehaene, S. 1989. Neuronal models of cognitive function. *Cognition*, 33, 63-109.

Conant, R. C., and Ashby, W. R. (1970). 'Every good regulator of a system must be a model of that system', *International Journal of Systems Science*, 1: 89-97.

Davidson, D..1980. *Essays on Actions and Events,* Oxford, Clarendon Press. ; trad. P. Engel.1993. *Actions et événements*, Paris, PUF.

Dennett, D.C. 2003. *Freedom Evolves.* New York, Viking Press.

Dienes, Z. & Perner,J. 2002. The metacognitive implications of the implicit-explicit distinction, in Izaute, M., Chambres, P., Marescaux, P.-J. (Eds), *Metacognition: Process, function, and use*. Dordrecht, Netherlands: Kluwer Academic Publishers.

Dretske, F. 1988. *Explaining Behavior : Reasons in a world of causes*. Cambridge : MIT Press.

Edelman, G. M. 1987. *Neural Darwinism : the theory of neuronal group selection*. New York: Basic Books.

Evans, G. 1982. *The Varieties of Reference,* Oxford, Clarendon Press.

Frankfurt H.G., 1988, *The importance of what we care about,* Cambridge, Cambridge University Press.

Goldman, A. 1970. *A Theory of Human Action*. Englewood Cliffs, NJ: Prentice-Hall.

Granger, G.G. 1979. *Langages et Epistémologie*. Paris, Klincksiek.

Haggard, P., 2003. Conscious awareness of intention and of action, in J. Roessler & N.Eilan (eds.), *Agency and self-awareness: issues in philosophy and psychology*, Oxford, Oxford University Press, 111-127.

Kim, J. 1993. *Supervenience and mind*. Cambridge, Cambridge University Press.

Kistler, M. 1999. Causalité et lois de la nature, Paris,Vrin ; version anglaise révisée, 2006 London, Routledge.

Locke, J., [1755]-1994 *An Essay concerning Human Understanding,* trad. Costes, *Essai Philosophique concernant l'Entendement Humain,* Paris, Vrin.

Malebranche.[1674-5]-1979. *De la recherche de la vérité*. Paris: Gallimard NRF-Pleiade.

Marteniuk, R.G., MacKenzie, C.L., Jeannerod, M., Athenes, S. 1987. Marteniuk, 1987 Constraints on human arm movement trajectories. - *Canadian Journal of Psychology,* 41(3): 365-78.

McCann, Hugh .1998. *The Works of Agency: On Human Action, Will and Freedom*. Ithaca, New York: Cornell University Press.

Mellor, H. 1991. *Matters of Metaphysics*. Cambridge: Cambridge University Press, 17-29.

Millikan, R. 1984. *Language, Thought and other biological categories, New Foundations for* Realism, Cambridge, MIT Press.

Peacocke, C., (1992a): Scenarios, Concepts and Perception, in T. Crane (dir.), The Contents of experience, Cambridge, Cambridge University Press, 105-135.

Proust, J. 2000. Les conditions de la connaissance de soi in *Philosophiques* 27, 1, 2000, 161-186.
Proust, J. 2001. A plea for mental acts, *Synthese*, 129, 105-128.
Proust, J. 2003a. Thinking of oneself as the same. *Consciousness and Cognition*, 12, 495-509.
Proust, J. 2003b. Perceiving intentions, in J. Roessler & N.Eilan (dirs.), *Agency and self-awareness: issues in philosophy and psychology*, Oxford, Oxford University Press, 2003, 296-320.
Proust, J. 2004. La philosophie de l'esprit a-t-elle besoin d'experts en cognition ? ? in : *La philosophie cognitive*, E. Pacherie & J. Proust (dirs.), Gap, Ophrys/ Paris, Editions de la MSH, 35-53.
Proust, J. 2006a. Rationality and metacognition in non-human animals, in S. Hurley & M. Nudds (dirs.) *Rational Animals* ?, Oxford, Oxford University Press, 247-274.
Proust, J. 2006b. Agency in schizophrenics from a control theory viewpoint, in W. Prinz & N. Sebanz .(dirs.) *Disorders of volition*, Cambridge, MIT Press, 87-118.
Proust, J. à paraître. Metacognition and metarepresentation. *Synthese*.
Proust, J. à paraître. What is a mental function ?, in : A. Brenner & J. Gayon (eds.), *French Philosophy of Science*, Boston Studies in the Philosophy of Science, Springer, 2007.
Proust, J. & Frankowska, H. en préparation. Phenomenal experience and viability theory.
Proust, J. & Pacherie, E. à paraître. Neurosciences et compréhension d'autrui. In *Des neurones à la philosophie : Neurophilosophie et philosophie des neurosciences*, E. Ennen, L. Faucher, P. Poirier et É. Racine, (dirs.), DeBoeck Université.
Quartz, S.R. & Sejnowski, T.J. 1997. The neuronal basis of cognitive development : a constructivist manifesto. *Behavioral and Brain Sciences,* 20, 537-596.
Searle, J.R..1983. *Intentionality, an Essay in the Philosophy of Mind,* Cambridge, Cambridge University Press. Trad. fr. par Cl. Pichevin, *L'Intentionnalité,* Paris, Ed. de Minuit, 1985.
Shallice, T., 1988. *From Neuropsychology to Mental Structure,* Cambridge, Cambridge University Press.
Sugrue, L.P., Corrado, G.S. & Newsome, W.T.: 2005, Choosing the greater of two goods : neural currencies for valuation and decision making, Nature Reviews Neuroscience, 6, 36375.
Wegner, D. 2002. The illusion of the conscious will. Cambridge : MIT Press.
Wittgenstein, L. 1953- 2005. *Recherches philosophiques*; trad. fr. par F. Dastur et al., Paris : Gallimard.