La *Belle au bois dormant* : débat autour d'un paradoxe

(Sleeping Beauty: Debate on a Paradox)

Laurent Delabre
laurentdelabre@yahoo.fr

Résumé : Je présente à nouveau le paradoxe probabiliste de la *Belle au bois dormant* et propose une vue d'ensemble des discussions engagées dans le but de résoudre le problème. Je résume et parfois critique brièvement les différentes vues : position neutre, tiérisme bayésien, tiérisme non bayésien, demisme traditionnel, ainsi que le nouveau demisme de la conservation de la croyance, un point de vue méconnu mais intéressant. J'essaie en même temps de clarifier quelques notions essentielles, de présenter les grands acteurs du débat, et d'anticiper l'évolution de ce dernier.

Abstract : I restate the Sleeping Beauty probabilistic paradox and offer an overview of the ongoing discussions that aim at resolving the problem. I summarize and eventually criticize briefly the various views: neutral position, Bayesian thirdism, non-Bayesian thirdism, traditional halfism, as well as the new halfism of credence conservation, a somewhat neglected but interesting point of view. I try at the same time to clarify some essential notions, to introduce the main actors of the debate, and to anticipate its evolution.

Sommaire :

Tel un monstre créé pour déconcerter les bons esprits, le paradoxe probabiliste de la *Belle au bois dormant* divise les spécialistes internationaux du *self-locating belief* depuis quelques années. Il sera sans aucun doute plus résistant que le fameux *paradoxe de Monty Hall*, qui n'effraie plus personne aujourd'hui. Le problème est pourtant très simple en apparence, moins abstrait que l'*Argument de l'Apocalypse* ou le *Pile ou face divin*, avec lesquels il est souvent comparé. Au-delà des cercles philosophiques, il peut aussi amuser les curieux ou surprendre les mathématiciens et les informaticiens les plus aguerris. Il ne faut pas le prendre à la légère : de grandes leçons peuvent être tirées de nos tentatives pour le résoudre, voire de sa résolution, laquelle nécessite un consensus pour l'instant inenvisageable. Sur le chemin de la connaissance se présentent ainsi des obstacles qui éprouvent la raison et engendrent d'interminables débats. Nous préférons parfois les contourner et les oublier, mais nos efforts pour les franchir ne sont jamais une perte de temps.

1. Un problème, une polémique.

Il est impossible de savoir avec certitude qui a imaginé le paradoxe et quelle forme celui-ci aurait pris par le passé. Le nom *Belle au bois dormant* lui aurait été donné par Robert Stalnaker. En 1997, il fit une apparition peu remarquée dans un article de Piccione et Rubinstein, et il fallut attendre encore quelques années pour que la communauté philosophique commence à le prendre au sérieux. Rappelons l'énoncé du problème et donnons une vue d'ensemble des différentes attitudes et opinions des chercheurs.

La Belle au bois dormant accepte de se soumettre à une expérience dont elle connaît toutes les règles. Le dimanche soir, alors qu'elle dort profondément, on lance une pièce de monnaie équitable. Si la pièce tombe côté face, on réveille la Belle le lendemain (lundi), on a un entretien avec elle, puis on la rendort en lui administrant un somnifère qui lui fait complètement oublier les événements de cette journée ; le mardi, on la laisse dormir. Si c'est pile, on la réveille le lundi et ont lieu de même entretien et prise du somnifère à effet amnésique, puis on la réveille une seconde fois le mardi selon les mêmes modalités. On la réveille enfin mercredi et l'expérience est terminée. Au cours de l'entretien, on demande à la Belle : « À quel degré devez-vous croire que le résultat du lancer de la pièce est face ? » (question parfois reformulée ainsi : « Quelle est la probabilité que la pièce soit tombée sur face ? »).

Voici qu'au cours de l'expérience, la Belle se réveille, incapable de déterminer si on est lundi ou mardi. On s'entretient avec elle et on finit par lui poser la question fatidique. Que doit-elle répondre ?

L'énoncé original s'arrête là, mais il est intéressant de le compléter ainsi : si on est lundi, dès que la Belle a répondu à la question de l'entretien, on lui annonce qu'on est lundi et on lui pose à nouveau la question. Que doit-elle répondre alors ?

Les réponses aux deux questions pourraient bien être : 1/3 (la réponse principale) puis 1/2 (après annonce de la date). Les nombreux partisans de ces résultats sont souvent appelés tiéristes (*thirders*). Leur première argumentation commence par un constat fréquentiste : si l'expérience était répétée de très nombreuses fois, il y aurait deux fois plus de réveils-pile que de réveils-face, et autant de réveils-pile le lundi que de réveils-face le lundi. La Belle devrait donc se dire qu'il y a une chance sur trois que son réveil soit un réveil-face, c'est-à-dire une chance sur trois que la pièce soit tombée sur face, puis ramener le degré de sa croyance à 1/2 après annonce de la date. Ce dernier résultat s'obtient aussi par une application du théorème de Bayes, si l'on prend soin, bien sûr, d'attribuer aux probabilités *a priori* P(Face) et P(Lundi) les valeurs recommandées par le tiérisme :

$$P(\text{Face} \mid \text{Lundi}) = P(\text{Face}).P(\text{Lundi} \mid \text{Face}) / P(\text{Lundi})$$
$$= (1/3).1 / (2/3)$$
$$= 1/2$$

Il conviendrait de distinguer au moins deux tiérismes. Adam Elga, qui, le premier, tente dans un article de résoudre la *Belle au bois dormant*, portant du même coup le problème à la connaissance de nombreux philosophes, prend finalement parti pour une thèse assez contestée : à son réveil, la Belle doit changer son degré de croyance initial 1/2 en 1/3 bien qu'étrangement elle ne gagne aucune information expliquant ce décalage. Un tiérisme plus récent, partagé par beaucoup, simplement qualifié de bayésien par Horgan, son principal représentant, essaie de corriger ce défaut et affirme que la Belle acquiert une nouvelle information. Ce problème de l'information est la principale difficulté dans l'argumentation tiériste.

De l'autre côté, les demistes (*halfers*) pensent que la réponse de la Belle avant l'éventuelle annonce de la date doit être 1/2. Leur premier raisonnement est celui-ci : dimanche soir, la Belle s'endort en croyant que la pièce a une chance sur deux de tomber sur face. Elle se réveille, incapable de se repérer précisément dans le temps, mais elle savait déjà qu'elle devait être réveillée pendant l'expérience, aussi elle ne gagne aucune information qui viendrait modifier son degré de croyance initial, 1/2. Par contre, selon les demistes de l'école de feu David Lewis, il se produit effectivement un décalage bayésien lorsque la Belle apprend qu'on est lundi. Si l'on attribue aux probabilités *a priori* P(Face) et P(Lundi) les valeurs appropriées, le calcul donne :

$$P(\text{Face} \mid \text{Lundi}) = P(\text{Face}).P(\text{Lundi} \mid \text{Face}) / P(\text{Lundi})$$
$$= (1/2).1 / (3/4)$$
$$= 2/3$$

Ce résultat, qui est donc la réponse à fournir après annonce de la date, est fortement contrintuitif et engage peu de chercheurs sur la voie du demisme traditionnel. Une autre thèse, selon laquelle la Belle doit répondre 1/2 lorsqu'elle sait qu'on est lundi, est née récemment. Défendu notamment par Nick Bostrom sous une autre appellation (ce philosophe parle d'un « modèle hybride » qu'il conçoit comme une synthèse du tiérisme et du demisme traditionnel), ce demisme où le degré de croyance initial est conservé lors du réveil de la Belle même après l'annonce de la date, n'est pas très apprécié lui non plus, car il *semble* ne pas respecter l'inférence bayésienne. Les arguments de ses rares avocats (dont je fais partie) sont difficiles mais passionnants.

Le paradoxe est redoutable : les solutions demiste et tiériste, qui se rejettent l'une l'autre, alternent dans l'esprit des indécis qui ne trouvent pas plus de cohérence dans l'une que dans l'autre. Heureux les indécis ! En goûtant les points de vue opposés, ils se préparent à construire une solution plus riche. Cela dit, le débat ne se réduit pas à un affrontement entre demistes et tiéristes. Parmi les chercheurs neutres, bien sûr, certains prennent le temps de réexaminer les principes, les modèles et les vieilles méthodes pour les adapter à ce défi d'un nouveau genre. Mais d'autres estiment que c'est inutile : la *Belle au bois dormant* est simplement insoluble dans son énoncé actuel. L'énoncé et, par suite, nos raisonnements et nos tentatives de réponse sont peut-être corrompus si, quelque part, un non-sens s'est déguisé en proposition sensée, ou si une proposition est interprétée de deux façons différentes par manque de précision. Selon Berry Groisman, l'approche épistémique des probabilités est inadéquate alors que le fréquentisme convient parfaitement pour dénouer le paradoxe, à la condition de préciser, si cela manque dans l'énoncé, ce qu'il faut compter pour établir la fréquence et donc la probabilité objective. Voilà qui est étrange ! Quand je dis que je suis dans un réveil-face, j'affirme principalement et clairement que la pièce est tombée sur face,

un réveil-face étant par définition un événement qui suit l'obtention de face. Les lancers de pièce et les réveils sont liés, et pourtant compter les uns ne conduit pas au même résultat que compter les autres puisque pile est suivi par deux réveils au lieu d'un.

Les discussions des philosophes étaient-elles vaines, voire trompeuses ? Est-ce maintenant la fin du cauchemar, pour reprendre le mot de Groisman[1] ? Nous aimerions être aussi rassurés que lui, mais ce n'est pas le cas. Il est possible, certes, que le paradoxe de la *Belle* superpose deux problèmes aux solutions bien différentes, et que des ajouts dans sa formulation permettent de trancher. Alors peut-on simplement se dire que la réponse est 1/2 si l'on considère tel événement et non tel autre, 1/3 si l'on considère tel univers et non tel autre ? Est-il satisfaisant de se dire qu'entre 1/2 et 1/3, il ne peut pas y avoir de vainqueur, et qu'il est inutile de débattre plus longtemps ? En relisant l'énoncé, nous constatons toute la malice des éléments du scénario, notamment de cette drogue à effet amnésique, mais en même temps nous constatons la simplicité des phrases et nous comprenons le sens, tout le sens, sans ambiguïté, comme s'il n'y avait rien à ajouter. Les mêmes questions resurgissent. Pourquoi la Belle ne pourrait-elle pas donner une seule réponse et prouver que c'est la bonne ? Que signifie, pour un *sujet*, préférer une solution à une autre ? Quelle est la cause des fluctuations de notre intuition quand, d'accord avec une solution, nous y renonçons soudain ? Si l'objectivisme ne parvient pas, seul, à donner une unique solution autrement qu'en risquant de transformer le problème en un autre et, au lieu d'affronter la difficulté, de la contourner comme si elle n'existait pas, alors il apparaît plutôt qu'il est insuffisant. Tout en tenant compte des mises en garde de Groisman, qui perçoit tout de même l'ambiguïté du comptage fréquentiste, nous devons reconnaître l'importance du subjectivisme et nous replonger dans les arguments des philosophes tiéristes et demistes.

2. Mondes centrés, principe de réflexion et anamorphose.

Il convient de faire intervenir la notion de monde centré (*centered world*). Engagée dans l'expérience, la Belle qui apprendrait que pile a été obtenu, c'est-à-dire qu'elle est dans le monde engendré par le fait que la pièce est tombée côté pile, serait incapable de se localiser dans le temps au jour près avant qu'on lui apprenne la date précise. Elle peut prétendre être la Belle du monde-pile, mais pas à la fois la Belle du lundi et la Belle du mardi. Ainsi, réfléchir au problème de la *Belle* requiert la distinction de trois parties temporelles du sujet de l'expérience, qui ne peuvent donc pas être présentes en même temps. On peut dire que celle qui est localisée dans le monde-face est présente si et seulement si est vraie la proposition centrée :

H_1 : « La pièce est tombée sur face et aujourd'hui est lundi »

Les deux autres, colocalisées dans le monde-pile, correspondent aux propositions centrées :

T_1 : « La pièce est tombée sur pile et aujourd'hui est lundi »
T_2 : « La pièce est tombée sur pile et aujourd'hui est mardi »[2]

Lorsque la Belle se réveille et cherche à estimer, grâce aux probabilités, sa place dans l'univers, elle voit que ces trois propositions sont possibles mais pas forcément équiprobables. Une et une seule est vraie mais… si H_1 est vraie, alors T_1 et T_2 ne l'ont jamais été et ne le seront jamais durant l'expérience ; si T_1 est vraie, alors H_1 ne l'a jamais été et ne le sera

[1] « The End of Sleeping Beauty's Nightmare » est le titre de son article publié récemment. Peu cité dans les bibliographies des bayésiens, il a pourtant inspiré des chercheurs très ouverts comme Franceschi.
[2] Ces symboles sont évidemment les conventions des anglophones : H est mis pour *Heads* et T pour *Tails*.

jamais durant l'expérience, et T_2 le sera demain ; si T_2 est vraie, alors H_1 ne l'a jamais été et ne le sera jamais durant l'expérience, et T_1 l'a été hier. Certes, la formulation de la question de l'entretien incite la Belle à n'envisager que deux mondes non centrés, mais le protocole de l'expérience est tel qu'en cas de pile, on lui pose la question plus souvent, et qu'elle ne peut pas se souvenir d'un précédent entretien. À quel point doit-elle prendre en compte le temps qui lui échappe ? On le voit : la distribution des probabilités est un vrai casse-tête. La Belle ignore le résultat du lancer de la pièce ainsi que sa localisation temporelle, il est évident qu'elle doit croire que l'hypothèse T_1 est aussi probable que sa voisine, l'hypothèse T_2 (respect du principe d'indifférence partagé par Adam Elga et David Lewis). Mais doit-elle croire que l'hypothèse H_1, en quelque sorte plus étrangère aux deux autres hypothèses que ces dernières ne le sont l'une pour l'autre, est aussi probable que T_1 (point de vue tiériste) ou bien aussi probable que $T_1 \vee T_2$ (point de vue demiste) ? Quelles sont les règles du jeu, quels sont les bons branchements dans cet entrecroisement d'hypothèses qu'un sujet émet sur sa situation temporelle et sur le résultat d'un tirage aléatoire ?

Le tiériste Elga raisonne ainsi[3] : si la Belle apprend qu'on est lundi, alors elle doit estimer la probabilité de face à 1/2, qui est tout simplement la probabilité qu'une pièce équitable, récemment lancée, soit tombée sur face. En effet, la voici réveillée lundi, événement qui devait assurément arriver ; la pensée que demain peut avoir lieu un réveil n'a aucune importance. Or, cette probabilité de face sachant lundi n'est pas autre chose que $P(H_1 \mid H_1 \vee T_1)$, qui est égale à $P(H_1) / [P(H_1) + P(T_1)]$. Il s'ensuit que $P(H_1) = P(T_1)$. Comme $P(T_1) = P(T_2)$ et $P(H_1) + P(T_1) + P(T_2) = 1$, on arrive à $P(H_1) = P(T_1) = P(T_2) = 1/3$. Autrement dit, la Belle encore ignorante de la date doit estimer à 1/3 la probabilité de face.

Le demiste Lewis argumente ainsi[4] : la Belle n'acquiert à son réveil aucune information susceptible de modifier sa croyance en l'obtention de face. Plus exactement, elle ne gagne aucune nouvelle évidence non centrée entre le moment où, sur le point de s'engager dans l'expérience, elle croit que P(Face) = 1/2, et le moment où, ignorante de la date, elle estime à nouveau la probabilité de face ; elle ne gagne qu'une évidence centrée sans aucune force, $(H_1 \vee T_1 \vee T_2)$. La probabilité de face étant inchangée, $P(H_1) = P(T_1 \vee T_2) = 1/2$ et par conséquent $P(T_1) = P(T_2) = 1/4$. Lewis explique en outre qu'Elga part d'une prémisse qu'il croit sûre alors qu'elle n'est qu'intuitive, et l'intuition nous leurre parfois : le raisonnement du tiériste, qui prend en effet appui sur la croyance insuffisamment justifiée $P(H_1 \mid H_1 \vee T_1) = 1/2$, semble défectueux. D'après Lewis, $P(H_1 \mid H_1 \vee T_1) = 2/3$, résultat contrintuitif et pourtant plus sûr, calculé grâce à l'application (demiste) du théorème de Bayes vue plus haut.[5]

Elga affirme que la *Belle au bois dormant* fournit un nouveau contre-exemple au principe de réflexion de van Fraassen, dont il retient l'énoncé suivant :

« Un sujet, certain qu'il attribuera demain le degré de croyance *x* à la proposition R (à moins qu'il reçoive une nouvelle information ou subisse des incidents cognitifs entre-temps), doit *maintenant* attribuer le degré de croyance *x* à R. »

En effet, le dimanche, veille de l'expérience, un sujet qui accepte le tiérisme non bayésien d'Elga sait qu'il estimera demain à 1/3 la probabilité de face, alors qu'il ne recevra entre-temps aucune information nouvelle et n'aura pas encore pris la drogue à effet amnésique, et pourtant il continue d'estimer à 1/2 cette probabilité tant qu'il n'est pas entré dans l'expérience. Bien sûr, pour Lewis, la thèse d'Elga ne met pas en défaut le principe de réflexion : c'est le principe qui met en défaut la thèse. Toutefois, nous pourrions objecter à ces deux chercheurs (et à beaucoup d'autres) que le conflit entre principe et thèse est discutable,

[3] Elga 2000.
[4] Lewis D. 2001.
[5] Meacham 2008 détaille tous les principes qui fondent les argumentations d'Elga et de Lewis.

car dès le lundi, la Belle a perdu la capacité de se repérer dans le temps au jour près. Sa mémoire n'a pas encore été altérée, mais elle n'en sait rien ; il lui semble qu'elle est réveillée le lendemain du dimanche, mais elle se dit que c'est peut-être déjà mardi. Elle est en proie au doute et finalement dans le même état cognitif qu'après l'absorption de la drogue, comme si celle-ci avait un effet rétroactif. Les observateurs extérieurs savent qu'elle n'a pas subi un « incident cognitif », mais pour elle, subjectivement, être droguée ou non droguée est la même chose : elle est dans l'expérience, à l'intérieur d'une sorte de boîte, un lieu où les choses ne se passent pas comme à l'extérieur[6].

Cette image de la boîte est par ailleurs une des métaphores utilisées par Jean-Paul Delahaye. Selon ce tiériste, le protocole de l'expérience est une boîte qui agit sur celui qui y est enfermé, plus exactement sur sa perception des probabilités. Le sujet engagé dans l'expérience, et lui seul, regarde soudain la probabilité objective 1/2 de pile à travers un miroir déformant. Cette distorsion de l'espace de probabilité, ou *anamorphose probabiliste*, est ici un effet de loupe : la Belle sait qu'elle existe en double dans le monde-pile, aussi il lui apparaît que la probabilité de pile grossit, comme une tête d'épingle vue à travers une loupe. Finalement, le degré de la croyance en l'obtention de pile atteint 2/3 (il doit être le double du degré de croyance en l'obtention de face). L'information ($H_1 \vee T_1 \vee T_2$), que Lewis prive de toute efficacité, a donc une importance chez Delahaye, qui la retraduirait ainsi : « Le protocole est maintenant enclenché. » Son acquisition est prévisible dès la veille de l'expérience et on ne voit pas comment l'employer dans une application du théorème de Bayes pour modifier des croyances. Pourtant, la certitude que l'on n'est pas en dehors mais dans l'expérience (dont on connaît déjà les règles), est suffisant pour transformer la probabilité subjective de face.

L'anamorphose est une sorte de généralisation de l'inférence bayésienne, elle traite des cas où conditionaliser semble impossible. Ses résultats ne sont pas contestables quand ils peuvent être retrouvés grâce à l'inférence classique. Par exemple, l'effet de filtre, anamorphose opposée à l'effet de loupe, que Delahaye décrit dans sa variante du *Protocole sans mardi* (l'expérience ne dure qu'un jour et en cas de pile, cette fois-ci, la Belle peut ne pas être réveillée), modifie comme il faut les probabilités[7]. Mais je crois que les démonstrations très illustrées qui accompagnent l'effet de loupe ne sont pas assez convaincantes. Pour les demistes qui peuvent, eux aussi, réfléchir aux effets de sélection, aux erreurs dans la collecte d'informations et autres « biais anthropiques »[8], le protocole de l'expérience originale est particulier, il n'est pas un miroir déformant, les deux réveils-pile ne changent rien ; ils voient cela aussi clairement que le tiériste français perçoit un changement de perspective du sujet, et argumentent avec autant de force.

Avant de présenter, justement, le demisme traditionnel, signalons rapidement que l'argument des paris, parfois utilisé par les tiéristes non bayésiens, et toujours décisif, selon Delahaye, lors de la résolution d'un problème probabiliste, est de plus en plus critiqué. Plusieurs chercheurs essaient de montrer qu'en raison de la nature du paradoxe, le système de récompenses qui vient spontanément à l'esprit, celui où la Belle parie sur le résultat du lancer de la pièce à chaque réveil et gagne une même somme d'argent à chaque bonne réponse (donc deux fois cette somme en cas de paris gagnants sur pile), ne prouve pas que le tiérisme a vu juste. Des systèmes plus complexes de *Dutch books* sont élaborés, puis eux aussi attaqués. La controverse ressemble à un jeu où le plus astucieux l'emportera[9].

---

[6] Monton 2002 discute aussi, avec d'autres arguments liés à la perte de mémoire, cette violation apparente du principe de réflexion par les tiéristes non bayésiens.

[7] Delahaye 2003.

[8] Bostrom 2002 est l'ouvrage de référence pour se familiariser avec les effets de sélection.

[9] Pour plus de détails, on lira avec intérêt Arntzenius 2002, Hitchcock 2004 ou encore Pust 2008b.

3. Aspects du demisme traditionnel.

Les nombreux tiéristes arrivés au secours d'Elga ont raison de prétendre que les demistes sont minoritaires, mais ils n'avaient parfois lu que le court article de Lewis qui, malheureusement décédé, ne peut plus se défendre. Aujourd'hui, ils connaissent un peu mieux les positions de Leslie, White ou Bradley, pour ne citer que ceux-ci. Franceschi, lui aussi, fut demiste pendant plusieurs années avant d'essayer de rassembler les vues opposées dans une solution unique.

Il semble que la plupart des demistes, en réaction au comptage fréquentiste des réveils élémentaires, considèrent des séries de réveils, groupent les réveils dans l'unité de l'expérience. John Leslie est d'avis que le tiérisme qu'il combat naît à la pensée de la répétition de l'expérience. Cette simulation mentale est salutaire dans la plupart des cas, pas avec la *Belle au bois dormant*. En effet, lorsque l'expérience est reproduite de très nombreuses fois, les réveils-pile comme les réveils-face deviennent *effectifs* et on trouve une majorité de réveils-pile : les résultats tiéristes semblent alors l'emporter. Mais si l'on est sûr qu'elle n'a lieu qu'une fois, alors un réveil-face est seulement *possible*, et ce n'est pas un réveil-pile mais une série de deux réveils-pile qui est, elle aussi, seulement *possible*. Dans ce cas, la Belle qui, d'une part, sait que la pièce de monnaie décide équitablement s'il y aura un seul réveil ou une série de deux réveils, et qui, d'autre part, n'a pas le souvenir d'avoir déjà été réveillée au cours de l'expérience, n'a pas de raison de croire qu'être dans le monde-pile (dans la série de réveils-pile) est plus probable qu'être dans le monde-face. Et le fait qu'en cas de pile il y ait deux entretiens et donc que la question lui soit posée deux fois (on ne compte pas la question subsidiaire du lundi) n'est pas perturbant. Quelle serait la probabilité de face si l'expérience était répétée un petit nombre de fois ? À ma connaissance, Leslie ne propose pas de formule mathématique pour la calculer[10].

Paul Franceschi, influencé par Leslie mais conscient de l'étrangeté de sa distinction entre une expérience unique et une expérience renouvelée, résume ainsi sa manière de corriger ce qui est pour lui l'erreur tiériste : « On ne peut pas additionner les réveils-face le lundi et les réveils-pile le lundi, car il ne s'agit pas du même objet. Les réveils-pile le lundi sont indissociables des réveils-pile le mardi : on ne peut avoir un réveil-pile le lundi sans un réveil-pile le mardi. Pour cette raison, alors que les réveils-face le lundi comptent 1 (1 objet), les réveils-pile le lundi et les réveils-pile le mardi ne comptent qu'1/2 (1/2 objet). »[11] Un réveil-face n'est pas le même objet qu'un réveil-pile car c'est un groupe dont la particularité est de n'avoir qu'un membre ; c'est une série au même titre que la série des deux réveils-pile indissociables. Franceschi ne critique pas le principe du comptage fréquentiste, il reprécise ce que doivent être les objets comptés : ses groupements d'un ou de deux réveils ne sont pas arbitraires mais au contraire exigés par le protocole. De son point de vue, même si l'expérience était répétée, la Belle devrait toujours estimer à 1/2 la probabilité de face.

Roger White sait que les tiéristes sont préoccupés par le problème de l'information nouvelle acquise entre le moment où la Belle s'endort dimanche soir et le moment où elle se réveille. Peu convaincu par les propositions tiéristes, il montre la pertinence d'une autre information favorisant le demisme :

W : « Je suis réveillée au moins une fois durant l'expérience »

Dans des variantes où la Belle peut ne pas être réveillée durant l'expérience, W semble plus appropriée que toute autre information pour produire de justes modifications de

[10] Je remercie beaucoup Jean-Paul Delahaye, qui m'a fait parvenir une correspondance avec Leslie où celui-ci livre son précieux avis.

[11] Extrait d'une correspondance personnelle. Franceschi 2008 affine cet argument « ontologique ».

croyances[12]. Dans le cas du problème original où la Belle est toujours réveillée, W laisse évidemment les probabilités inchangées. Il est inutile d'expliquer davantage le défi que White lance aux partisans du 1/3, il suffit de constater qu'à une localisation du sujet dans un jour, il oppose une localisation dans l'expérience quel que soit ce jour : « au moins une fois durant l'expérience » est la manière circonspecte de dire simplement « dans l'expérience ».

En cherchant à se repérer dans le temps, un sujet se situe-t-il plus facilement dans un jour qu'il ne peut pas dater que dans l'expérience datable ? La question doit être débattue. Établir une probabilité-fréquence s'avère soudain moins simple qu'il n'y paraissait (lorsque le caractère élémentaire d'un réveil en faisait l'objet à compter par excellence). Si une localisation dans le jour correspond à un comptage de réveils élémentaires menant aux résultats tiéristes, une localisation dans l'expérience correspond à un comptage de séries de réveils consolidant le demisme. De Leslie qui refuse encore d'envisager la répétition de l'expérience jusqu'aux travaux les plus jeunes, nous constatons la prise de conscience progressive de la possibilité d'un argument fréquentiste soutenant, contre toute attente, l'approche épistémique demiste.

Darren Bradley, qui s'investit beaucoup dans le débat, demeure malgré tout un bayésien convaincu et n'a pas recours au fréquentisme. D'après lui, les tiéristes (principalement bayésiens) négligent un effet de sélection important, également rencontré dans le problème appelé *Certain Aces*[13] : Alice s'apprête à tirer une ou deux cartes d'un paquet dont vous seul connaissez le contenu. La première carte tirée sera un as, l'éventuelle seconde carte sera un roi. Vous demandez à Alice de lancer secrètement une pièce équitable et de tirer une seule carte en cas de face, deux en cas de pile. Ceci fait, elle vous révèle une de ses cartes : « J'ai un as. » Cette annonce confirme-t-elle, c'est-à-dire rend-elle plus probable l'obtention de face ? C'est difficile à dire car vous ne disposez pas de toutes les informations : Alice a forcément un as en main mais vous ignorez si, en cas de pile, elle avait l'intention d'annoncer une carte choisie au hasard parmi les deux. En effet, peut-être aurait-elle de toute façon annoncé son as, comme attirée par la valeur de la carte, ou bien voulait-elle dévoiler la première carte tirée. Si son procédé de sélection est aléatoire, alors son annonce confirme face. Si elle suit un « procédé persistant » biaisé en faveur de l'as, alors il n'y a aucune confirmation.

Bradley décrit aussi le problème *Uncertain Aces*, où l'as n'est pas nécessairement tiré et où l'on retrouve l'effet de sélection, mais selon le demiste, c'est *Certain Aces* et le procédé persistant qui modélisent le mieux la *Belle* : le réveil est un événement certain, et on ne place pas aléatoirement la Belle dans un état de veille ou de sommeil avant de lui demander la probabilité de face, la Belle n'estime des probabilités que lors d'un réveil. Lorsqu'elle ne sait rien d'autre que le fait qu'elle est réveillée, elle doit donc croire que ne sont confirmés ni pile ni face. Ce raisonnement éveille pourtant des soupçons. *Certain Aces* évoque plus facilement la situation de la Belle qui apprend qu'on est lundi : la révélation « J'ai un as » suppose qu'Alice s'est concentrée sur la première carte, pas la seconde (s'il y en a une) ; « Je suis réveillée lundi » suppose qu'on est le premier jour et pas le second. Là-dessus, Bradley est silencieux, alors que son argument pourrait se retourner contre lui et profiter à l'autre camp. Néanmoins, il met en évidence un effet de sélection qui pourrait jouer un grand rôle.

La tâche la plus ardue du demisme traditionnel est la défense du résultat contrintuitif $P(\text{Face} \mid \text{Lundi}) = P(H_1 \mid H_1 \vee T_1) = 2/3$. Les importants travaux de Franceschi sur la modélisation d'expériences de pensée dans les n-univers l'avaient amené, dans un premier

[12] Dans son papier de 2006, White décrit un « problème de la *Belle au bois dormant* généralisé », où chaque réveil élémentaire est remplacé par une possibilité de réveil : un générateur aléatoire ajusté sur une probabilité *c* décide s'il y a ou non réveil. Si la Belle prend en compte l'information W au moment de son réveil, elle peut calculer une probabilité de face qui est fonction de *c*.

[13] Bradley 2007b, pp. 11s.

temps, à défendre une analogie entre l'expérience de la Belle et une expérience avec une urne un peu spéciale[14]. L'urne contenait une boule rouge et une boule verte, laquelle était masquée lors d'un tirage sur deux (une pièce est secrètement lancée, en cas de face on ne peut ni voir ni sentir la boule verte). On tire de l'urne une boule, et si elle est rouge, tout le monde accepte l'idée que la probabilité de face passe de 1/2 à 2/3. Or, cette probabilité n'est pas étonnante d'un point de vue fréquentiste, elle se retrouve facilement par un comptage, ce qui n'est pas le cas du résultat demiste. Certes, le mardi en cas de face, le réveil de la Belle est, lui aussi, « masqué », mais l'analogie s'arrête là. Les boules sont des objets qui seront *peut-être* tirés de l'urne, elles ne correspondraient, dans l'expérience de la Belle, qu'à des *possibilités* de réveil, pas à des réveils effectifs. Un seul tirage, une seule boule tirée, cela ne peut pas évoquer un réveil de la Belle qui, amnésique, ne peut se rappeler un éventuel autre réveil. Franceschi semble aujourd'hui avoir abandonné cette analogie en même temps que le résultat P(Face | Lundi) = 2/3, mais il espère toujours modéliser au mieux l'expérience de la *Belle*[15].

Les variations qui discréditent la probabilité 2/3 sont nombreuses. Imaginons par exemple une expérience semblable à l'originale, où le réveil du mardi en cas de pile est remplacé par un événement quelconque, comme la dissimulation d'une rose derrière l'oreiller de la Belle endormie. Voici que la Belle, qui connaît cette nouvelle règle, se réveille lundi. Elle doit évidemment attribuer la probabilité 1/2 à H_1 et aussi à T_1, donc elle croit qu'il y a une chance sur deux pour qu'à la fin de l'expérience elle trouve une rose derrière son oreiller. À partir de là, revenons à un problème proche de l'original : supposons que l'événement du mardi en cas de pile soit un réveil de la Belle, laquelle, même amnésique, peut toujours prendre connaissance de la date du jour si elle le désire. Engagée dans l'expérience, la Belle est réveillée lundi et informée de cette date, peu importe comment. Il sera bien difficile de lui reprocher d'attribuer la probabilité 1/2 à H_1 comme à T_1. Devrait-elle changer sa croyance parce que l'événement éventuel du lendemain sera similaire à celui de ce jour, c'est-à-dire sera un réveil, un état de conscience où elle sera capable d'estimer des probabilités ? Devrait-elle croire que cet événement du lendemain est plus incertain, a une probabilité de 1/3 seulement ? Nous sommes tentés de répondre non.

4. Aspects du tiérisme bayésien.

Pour la plupart des tiéristes, un incident cognitif, même joint à la certitude que l'expérience est en cours, ne peut pas modifier la croyance en l'obtention de pile ou de face. La modification est nécessairement une révision bayésienne. Afin de mettre en évidence l'information nouvelle, plusieurs variantes de l'expérience de la Belle ont été proposées. Je donne ici celle de Cian Dorr, la plus commentée de toutes[16]. La Belle est réveillée le lundi et le mardi quel que soit le résultat du lancer de la pièce. Si pile, on lui administre à chaque fois la drogue à effet amnésique de l'expérience originale. Si face, on lui administre le lundi une drogue plus faible : la Belle est amnésique durant la première minute de son état de veille du mardi, puis elle recouvre la mémoire, et elle est alors sûre qu'on est mardi et que la pièce est tombée sur face. L'analyse de la variante nécessite de distinguer non plus trois mais quatre parties temporelles du sujet, associées aux propositions centrées :

H_1 : « La pièce est tombée sur face et aujourd'hui est lundi »
H_2 : « La pièce est tombée sur face et aujourd'hui est mardi »
T_1 : « La pièce est tombée sur pile et aujourd'hui est lundi »
T_2 : « La pièce est tombée sur pile et aujourd'hui est mardi »

[14] Franceschi 2005, pp. 3s.
[15] Franceschi 2008 défend une nouvelle analogie. On lira avec intérêt les stupéfiantes conclusions de l'analyse.
[16] Dorr 2002. Neal 2006 et surtout Finkelstein 2008 présentent également des variantes tiéristes ingénieuses.

Soit $P_n(R)$ le degré de croyance attribué à une quelconque proposition R par le sujet engagé dans l'expérience au milieu de la $n^{\text{ième}}$ minute de son état de veille. Au moment où la Belle se réveille, les quatre possibilités centrées sont pour elle équiprobables de toute évidence, donc $P_1(\text{Face}) = P_1(H_1) + P_1(H_2) = 1/4 + 1/4 = 1/2$. La probabilité de H_2 s'annule au bout d'une minute si la Belle ne se souvient pas d'un précédent réveil ; selon Dorr, rien ne vient favoriser ou discriminer une des trois hypothèses restantes, elles sont toujours équiprobables, donc $P_2(\text{Face}) = P_2(H_1) = 1/3$. La Belle a diminué le degré de sa croyance en l'obtention de face, ce qui semble normal puisqu'elle l'aurait au contraire augmenté (jusqu'à 1) si elle s'était souvenu d'un précédent réveil.

Dorr pense qu'il n'y a pas de différence significative entre sa variante et le problème original, ce que conteste Bradley. Pour le demiste, dans le problème original, si l'on excepte le moment où l'on apprend à la Belle qu'on est lundi, celle-ci n'a jamais l'occasion de croire face moins probable, et surtout pas plus probable, que pile, il n'y a pas un seul moment de son état de veille où arrive une information qui met à jour ses croyances, et cela fait toute la différence[17]. L'analogie entre les deux problèmes serait indéfendable. L'ingénieux tiériste Terry Horgan pense au contraire que la variante éloigne un peu dans le temps, ordonne et ainsi met en évidence certains mouvements dans l'esprit de la Belle, qui, nous allons le voir, ont lieu simultanément dans l'expérience originale.

Apportons avant tout une précision. Les propositions centrées font intervenir un indexical, c'est-à-dire, dans notre cas, un mot ou une expression comme « aujourd'hui » indiquant un moment *relatif* au sujet, et sont elles-mêmes dites *indexicales*. Selon Horgan, l'information W (« Je suis réveillée dans l'expérience »), est sans aucune force dans le problème de la *Belle*, voire inappropriée, notamment parce qu'elle n'est pas indexicale tant que le sujet sait *absolument* dans quelle expérience il se trouve, c'est-à-dire sait dater l'expérience. La clé du problème est dans l'information rivale, plus spécifique :

V : « Je suis réveillée aujourd'hui »[18]

Le dimanche soir, le protocole de l'expérience originale en tête, la Belle peut déjà songer à ce qui va lui arriver, à ses futurs moments de veille et de sommeil, elle a conscience qu'elle peut être inconsciente mardi. W est une certitude, mais V n'est pas une certitude car « aujourd'hui », le jour que la Belle anticipe, peut être mardi. Bien sûr, elle estime à 1/2 la probabilité que la pièce de monnaie tombe sur face.

Engagée dans l'expérience, voici que la Belle se réveille en perdant une information concernant sa localisation temporelle : elle ignore si ce jour est lundi ou mardi. Elle ignore également de quel côté est tombée la pièce. Aussi lui apparaissent les quatre possibilités H_1, H_2, T_1 et T_2, *essentiellement* indexicales, dans le sens où la Belle ne peut exprimer ce jour qu'à l'aide d'un indexical. Ces quatre hypothèses sont conformes à ses certitudes du dimanche. Elle attribue à chacune d'elles la probabilité *a priori* 1/4. Elle est également sûre des quatre probabilités conditionnelles correspondantes : $P(H_2 \mid V) = 0$, $P(H_1 \mid V) = P(T_1 \mid V) = P(T_2 \mid V) = 1/3$. Or, la perte d'information est accompagnée par un gain : V prend tout son sens, est une certitude maintenant que la Belle est consciente. V est une évidence essentiellement indexicale qui met à jour les probabilités subjectives. La Belle est en mesure de donner les probabilités *a posteriori*[19] des quatre hypothèses, et ces probabilités, bien sûr, sont égales aux probabilités conditionnelles correspondantes. Il en résulte la nouvelle probabilité que la pièce soit tombée sur face : 1/3.

---

[17] Bradley 2003.

[18] Horgan s'explique en deux temps, d'abord dans un article de 2004, puis dans un article de 2007. Dans ce dernier, la proposition « Je suis réveillée aujourd'hui » est symbolisée par A_{Today}. J'utilise la lettre V à la place, pensant conserver ce choix commode dans mes papiers ultérieurs.

[19] Il ne faut surtout pas entendre par là que cette estimation de probabilités est postérieure (dans le temps).

Horgan décrit là un glissement bayésien insolite. Les quatre possibilités centrées auxquelles la Belle assigne les probabilités *a priori* surviennent après le réveil parmi plusieurs événements mentaux synchrones. De plus, la Belle n'a évidemment jamais la possibilité, durant l'expérience, de reconnaître qu'elle n'est pas réveillée, alors qu'elle a toujours l'occasion de se dire qu'elle est réveillée. Pour cette dernière raison, Bradley juge le raisonnement du tiériste faillible[20].

Dans le tiérisme bayésien, la proposition « Je suis réveillée aujourd'hui » (parfois formulée « Je suis réveillée maintenant ») n'est pas la seule à jouer le rôle de l'information nouvelle. Sans doute dans le souci de proposer une information complète, précise, reflétant en quelque sorte tout le débat, un article récent, produit d'une rencontre de plusieurs philosophes, dont Horgan[21], utilise même une arme de la théorie objectiviste : la probabilité indéfinie (ou générale). Supposons que la Belle, engagée à nouveau dans l'expérience originale, réfléchissant au fait qu'elle vient d'être réveillée, essaie d'estimer au mieux l'instant du réveil. Par exemple, elle estime que c'était il y a plus de neuf minutes mais moins de dix minutes. Elle suggère donc un intervalle de temps Δ durant lequel on l'a assurément réveillée. Autrement dit, si W(*t*, *s*) signifie « La Belle a été réveillée dans l'expérience *s* quelque part durant l'intervalle Δ (relatif à l'instant *t*) et ne s'est pas souvenu d'un précédent réveil durant *s* » (les variables *t* et *s* sont libres), alors la proposition W(*maintenant*, σ), où σ est l'expérience particulière actuelle, est une certitude pour la Belle. Eh bien, l'article prétend démontrer que W(*maintenant*, σ) est une information capable de faire passer la probabilité de face de 1/2 à 1/3 ! Comment cela ?

Soyons large et disons qu'une expérience dure 72 heures, de dimanche midi à mercredi midi. Si B(*t*, *s*) signifie « *t* est un instant dans l'expérience *s* », si Toss(*x*, *s*) signifie « *x* est la pièce lancée dans l'expérience *s* », si H*x* signifie « *x* tombe sur face », si δ est la durée de Δ exprimée en heures, et si nous supposons une distribution uniforme des probabilités dans le temps, nous pouvons calculer les probabilités indéfinies suivantes :

$$\text{prob}(W(t, s) \mid Hx \wedge B(t, s) \wedge \text{Toss}(x, s)) = \delta / 72$$
$$\text{prob}(W(t, s) \mid \neg Hx \wedge B(t, s) \wedge \text{Toss}(x, s)) = 2\delta / 72$$

Remarquons que la seconde est le double de la première, puisqu'en cas de pile, il y a deux intervalles de durée δ pour lesquels W(*t*, *s*) est vraie. Le théorème de Bayes et un calcul que nous ne détaillerons pas nous conduisent donc à un résultat prévisible :

$$\text{prob}(Hx \mid W(t, s) \wedge B(t, s) \wedge \text{Toss}(x, s)) = 1/3$$

En se réveillant, la Belle apprend W(*maintenant*, σ) $\wedge$ B(*maintenant*, σ) $\wedge$ Toss(τ, σ) et peut directement inférer la probabilité définitive PROB(Hτ) = 1/3. La démonstration est tout de même discutable, à mon sens. Apparemment, la durée de l'expérience est sans importance : 50 heures ou 80 heures ne modifieront pas le résultat final. Mais pourquoi l'expérience devrait-elle avoir la même durée en cas de pile et en cas de face ? Changeons un peu les règles : la Belle participe à plusieurs expériences consécutives qui prennent toujours fin le lendemain du dernier réveil, ainsi elle ne reste jamais endormie un jour entier et, lorsqu'une expérience est terminée, elle s'engage dans la suivante sans attendre qu'un jour entier s'écoule. Dans ces conditions, un tiériste croit toujours que la Belle qui se réveille au milieu d'une expérience doit estimer à 1/3 la probabilité de face si elle ne peut pas dater le jour. Pourtant, indéniablement, une expérience-pile dure un jour de plus qu'une expérience-

[20] Bradley 2007b détaille cette objection. Pust 2008a ajoute, plus sévèrement, que Horgan maltraite la théorie épistémique : la Belle n'estime des probabilités qu'en se supposant consciente, réveillée, et il est impensable qu'elle n'attribue pas à H_2 la probabilité nulle préalablement à toute révision bayésienne.

[21] OSCAR Seminar 2008.

face, et les calculs précédents mènent forcément à une probabilité supérieure à 1/3. Pouvons-nous alors leur faire confiance ? L'arbitraire semble bien présent dans l'argument objectiviste, de sorte qu'on pourrait conclure n'importe quoi avec un peu d'astuce.

5. Le demisme de la conservation de la croyance.

Les chercheurs Monton et Kierland appliquent une méthode de minimisation de l'inexactitude de croyances pour tenter de résoudre le problème de la *Belle*. Ils concluent qu'à la première question de l'entretien, la Belle peut répondre 1/3 et 1/2, les deux probabilités sont acceptables ; mais dès qu'elle apprend qu'on est lundi, elle doit préférer 1/2 à 2/3. Ils ne voient pas de contradiction dans le fait que la Belle conserve le degré de croyance 1/2 malgré le gain d'information, et disent seulement que la conditionalisation bayésienne est parfois inappropriée dans les problèmes de ce genre[22]. Un représentant de ce que j'appelle le *demisme de la conservation de la croyance* ne serait pas satisfait par une si faible explication.

Que cette appellation ne nous leurre pas ! Un tel demiste n'est absolument pas un farouche adversaire du tiérisme, auquel il emprunte certaines vues comme il en emprunte au demisme traditionnel. Certes, il pense que la Belle qui ignore la date du jour doit préférer la réponse 1/2, mais il pense aussi que sa réponse ne change pas lorsqu'elle sait que ce jour est lundi. Il ne défend pas un principe de conservation à tout prix de la probabilité 1/2, faisant fi des informations mises à disposition, mais son raisonnement appliqué au problème de la *Belle* le conduit à défendre deux réponses qui se trouvent être identiques. Par contre, il peut adapter la première réponse dans un scénario où l'expérience est répétée, à certaines conditions ; nous avons vu que cette idée n'est pas partagée par tous les demistes traditionnels.

Parmi ces nouveaux demistes, Christopher Meacham est particulier. Pour analyser certains problèmes de *self-location* où le temps joue un rôle crucial, il propose de remplacer la conditionalisation centrée classique par ce qu'il appelle une *conditionalisation compartimentée*[23]. Nous avons jusqu'à présent conditionalisé de manière classique, lorsque nous avons étudié la variante de Dorr, ou plus simplement lorsque nous avons aidé la Belle qui apprend qu'on est lundi à mettre à jour ses croyances : après élimination d'une possibilité centrée, nous avons réparti sa probabilité entre les possibilités centrées restantes. La conditionalisation compartimentée consiste à ne distribuer la probabilité qu'entre les possibilités centrées du même monde non centré (à moins qu'il n'y en ait plus, auquel cas on se tourne vers les autres mondes). Par exemple, admettons qu'au réveil la Belle estime ces probabilités : $P(H_1) = 1/2$ et $P(T_1) = P(T_2) = 1/4$. Lorsque T_2 est éliminée (après le gain de l'information « Aujourd'hui est lundi »), seule T_1 doit en profiter car il ne faut considérer que le monde-pile. On arrive ainsi à $P(T_2) = 0$ et $P(H_1) = P(T_1) = 1/2$. La probabilité de face reste donc inchangée malgré le gain d'information. Meacham reconnaît que ces nouvelles règles de mise à jour des croyances sont problématiques dans bien des situations et heurteront nombre de bayésiens[24], mais il les défend au mieux contre les critiques qu'il arrive à prévoir.

Les autres demistes de la conservation de la croyance assurent ne pas maltraiter l'inférence bayésienne. Sur ce point, il est vrai que leur argumentation est parfois difficile à suivre. La Belle incapable de dater son réveil serait en quelque sorte différente de la Belle qui en est capable, et les croyances de la première Belle ne pourraient pas servir à déduire celles de la seconde. Bostrom, notamment, explique qu'il ne faudrait pas distinguer trois parties temporelles du sujet de l'expérience, mais plutôt cinq : deux dans le monde-face, en incluant la partie de la Belle qui, le lundi, sait qu'on est lundi, et de même trois dans le monde-pile. À

[22] Monton 2005.
[23] Meacham 2008.
[24] Bradley 2008, notamment, se penche sur le cas de Meacham.

chaque partie correspond une proposition centrée précisant si le sujet connaît ou ignore la date. Avant le gain de l'information « Aujourd'hui est lundi », les trois propositions qui précisent que le sujet ignore la date sont possibles et les deux autres impossibles (leur probabilité est nulle), et après le gain, bien entendu, celles qui étaient possibles deviennent impossibles et inversement. Cela rend inefficace le théorème de Bayes (en voulant l'appliquer avec des probabilités nulles, on aboutit à une expression indéfinie), et pourtant la démarche logique des bayésiens est tout à fait respectée. L'équiprobabilité de pile et de face avant comme après l'annonce de la date devient plausible. Elle n'est malheureusement pas démontrable par le calcul mais par l'étude de variantes extrêmes (où l'expérience peut durer un million de jours par exemple), dont les seules solutions envisageables sont celles qui ne heurtent pas fortement l'intuition[25].

Les explications de Bostrom paraissent insuffisantes. Il est vrai qu'il n'analyse pas assez la structure du paradoxe et ne montre pas pourquoi celui-ci diffère d'autres problèmes où la distinction de nombreuses parties temporelles n'est pas nécessaire[26]. La solution que je propose essaie tant bien que mal de gommer ces défauts. Nous avons déjà remarqué que la Belle a la possibilité de se localiser, à son réveil, soit dans un jour, soit dans l'expérience tout entière (ou, disons, l'intervalle de temps comprenant lundi et mardi). Dans beaucoup d'autres problèmes, cette distinction de deux localisations n'a aucune importance. Mais dans notre problème, le comptage fréquentiste des jours, qui appuie l'approche épistémique tiériste, ne conduit pas aux mêmes résultats que le comptage des expériences préféré par les demistes. Si une localisation est privilégiée par le sujet de l'expérience, il faut dire laquelle et pourquoi.

À mon sens, lorsque la Belle se réveille et perd la possibilité de se repérer au jour près, lorsqu'on violente sa perception du temps et la sereine succession des jours, qu'elle perd des morceaux de sa mémoire et donc de son existence, elle se localise plus naturellement dans une expérience datable que dans un jour que seul un indexical comme « aujourd'hui » peut exprimer. Elle fait taire cette inclination fragile dès qu'elle s'interroge sur la date du jour, et par exemple, si on lui apprend que la pièce est tombée sur pile, elle peut se demander quelle est la probabilité qu'elle soit la Belle du lundi plutôt que la Belle du mardi, et alors compter des jours. Mais si elle ignore tout et cherche à estimer la probabilité de face, c'est bien le résultat demiste qui va s'imposer en son esprit, parce qu'elle se situe d'abord dans l'expérience, pas dans le jour. Par contre, dès qu'elle apprend qu'on est lundi, son esprit est comme réorienté : alors qu'une association d'idées comme « je suis dans une expérience où je suis réveillée » avait du sens, « je suis dans une expérience où aujourd'hui est lundi » est un non-sens que sa raison refuse de former. Aussi, la localisation dans le jour reprend le dessus et la raison peut à nouveau travailler sereinement. La Belle capable de dater son réveil ne considère plus les mêmes objets que la Belle qui en était incapable : il est impossible de calculer grâce au théorème de Bayes les probabilités *a posteriori* aperçues par la Belle qui se localise dans le jour, à partir des probabilités *a priori* aperçues par la Belle qui se localise dans l'expérience. On ne peut être tenté d'appliquer le théorème que si l'on est encouragé par l'habitude, et peut-être trompé par le formalisme et éloigné de sa signification.

Les tiéristes trouveront peut-être cette solution compliquée, ils diront que la Belle serait plus tranquille en focalisant son esprit sur le jour à tout moment de son état de veille, et jamais sur l'expérience. Il est étrange que, dans l'incapacité de se situer exactement dans le temps, elle continue d'estimer à 1/2 la probabilité de face, alors qu'elle sait qu'elle annulerait

---

[25] Bostrom 2007, pp. 68ss.

[26] Bradley 2007a (pp. 136-141) est sévère. Le raisonnement de Bostrom revient à dire que, lorsque la Belle apprend qu'on est lundi, elle apprend aussi qu'elle vient d'apprendre qu'on est lundi, autrement dit elle gagne une information centrée qui la renseigne davantage sur sa localisation temporelle. Cette argutie cache le fait que Bostrom ne propose finalement aucune règle pour déterminer des degrés de croyance dans un problème comme la *Belle*, ce qui lui permet de faire passer la solution qu'il souhaite.

cette probabilité en apprenant ou en déduisant qu'on est mardi : le degré de croyance 1/2 devrait lui sembler trop grand dès qu'elle anticipe son proche avenir. Mais je crois que la Belle récupère, en apprenant la date, les repères temporels que le protocole de l'expérience lui a pris, bien plus qu'elle ne reçoit une information. Je le répète : elle fait plus que mettre à jour ses croyances, elle réoriente aussi sa raison, elle change de perspective et considère de nouveaux objets. Elle devient presque un autre sujet, suffisamment pour que le bayésianisme lui-même interdise une conditionalisation imprudente et insensée.

Afin de rendre l'entière signification de la notion de « modèle hybride » chère à Bostrom, ajoutons que ce philosophe cherche à concilier l'intuition de la probabilité 1/2 qui, chez le sujet incapable de dater son réveil, est un sentiment assez fort lorsque l'expérience originale n'a lieu qu'une fois, et l'intuition de la probabilité 1/3 qui naît à la pensée de la répétition de l'expérience. Il analyse plusieurs variantes, les n-*fold Sleeping Beauty Problems*, où l'expérience de la Belle, qui s'étend sur *n* semaines consécutives, consiste en une répétition de l'expérience originale (*n* lancers de pièce, *n* séries d'un ou deux réveils…) sans interruption (pas de réveil libérateur les mercredis, ce qui empêche la Belle de savoir, non seulement quel jour de la semaine, mais aussi quelle semaine elle est réveillée). Selon Bostrom, plus l'expérience originale est ainsi répétée, plus le degré de croyance en l'obtention de face est proche de 1/3. Il calcule des probabilités dans l'intervalle]1/3 ; 1/2[, ce que Leslie n'a pas réussi à faire. Par exemple, le sujet engagé dans l'expérience du *2-fold* et incapable de se repérer dans le temps doit estimer à 5/12 la probabilité que la pièce soit tombée sur face lors de son dernier lancer. Remarquons que les tiéristes (non bayésiens) Vaidman et Saunders anticipent et critiquent ce résultat six ans plus tôt, dans un court article[27] : contrairement à Bostrom, ils pensent que le demisme montre son inconsistance lorsqu'il en vient à calculer de telles probabilités intermédiaires.

Conclusion.

Dans la plupart des problèmes de probabilités où elles ont leur place, les conceptions épistémiques et fréquentistes se complètent et préconisent une solution unique. Au moins pour cette raison, les spécialistes du paradoxe de la *Belle au bois dormant* refusent souvent de croire que les conceptions sont devenues incompatibles. Les affrontements entre demistes subjectivistes et tiéristes objectivistes sont rares. Bien plutôt, celui qui n'omet aucune des approches possibles a des chances de convaincre. Ainsi, les demistes cherchent les objets à compter pour établir une fréquence égale au degré de croyance qu'ils défendent, et les tiéristes cherchent une information modificatrice de croyances afin de vérifier leur intuition fréquentiste de départ. Nous avons vu qu'un débat complexe s'ensuit, des principes sont revisités, des notions sont rediscutées. Le chercheur de vérité y prend goût, élabore des variantes de plus en plus astucieuses et engage toutes ses armes dans la bataille.

Comment ce débat peut-il évoluer dans les prochaines années ? Des papiers très récents, signés Peter Lewis, Papineau et Durà-Vilà entre autres, lui donnent une fière allure puisque, ni plus ni moins, l'interprétation de la mécanique quantique d'Everett est appelée à la rescousse. Une question délicate surgit : peut-on être tiériste tout en soutenant la théorie des états relatifs (les fameux mondes multiples), laquelle *semble* favoriser le demisme ? Nous ne parlerons pas maintenant de cette controverse difficile et loin d'être close, mais nous devons remarquer que l'intérêt porté à la *Belle* croît dans le monde de la philosophie des sciences.

Nous ne pouvons pas non plus étudier la critique du célèbre *Argument de l'Apocalypse* cher à Leslie. Rappelons quand même que des demistes aussi bien que des tiéristes ont comparé ce problème avec la *Belle* et trouvé des similitudes dans les structures. Comme pour

[27] Vaidman et Saunders 2001.

modérer leur audace, quelques-uns préfèrent défendre un parallélisme plutôt qu'une identité, avouant que subsistent des dissemblances dont ils n'ont pas mesuré l'importance. Pour les autres, les problèmes ne peuvent pas être résolus de la même manière, quoique la comparaison soit pertinente et enrichissante. Si nous devons distinguer les deux paradoxes, il n'est pas dit que nous distinguerons toujours les deux débats qui les concernent. Ces débats qui, bien qu'ils aient parfois croisé leurs routes, se sont pour l'instant surtout organisés indépendamment et différemment, ont-ils avantage à s'unir ?[28]

Références bibliographiques.

ARNTZENIUS, F.
2002 "Reflections on Sleeping Beauty", *Analysis* 62: 53-62
2003 "Some Problems for Conditionalization and Reflection", *Journal of Philosophy* 100: 356-370

BOSTROM, N.
2002 *Anthropic Bias: Observation Selection Effects in Science and Philosophy*, New York, Routledge
2007 "Sleeping Beauty and Self-Location: A Hybrid Model", *Synthese* 157: 59-78 (et http://www.anthropic-principle.com/preprints.html#beauty sous le titre "Sleeping Beauty: A Synthesis of Views")

BRADLEY, D. J.
2003 "Sleeping Beauty: A Note on Dorr's Argument for 1/3", *Analysis* 63: 266-268
2007a *Bayesianism and Self-Locating Beliefs or Tom Bayes meets John Perry*, mémoire
2007b "Four Problems about Self-Locating Evidence", *PhilSci Archive* 3344
2008 "Dynamic Beliefs", manuscrit, http://faculty.arts.ubc.ca/dbradley/

DELAHAYE, J.-P.
2003 "La Belle au bois dormant, la fin du monde et les extraterrestres", *Pour la Science*, n° 309, pp. 98-103 (et http://www.anthropic-principle.com/preprints.html#doomsday)

DORR, C.
2002 "Sleeping Beauty: in Defence of Elga", *Analysis* 62: 292-296

ELGA, A.
2000 "Self-Locating Belief and the Sleeping Beauty Problem", *Analysis* 60: 143-147 (et http://www.princeton.edu/~adame/)

FINKELSTEIN, J.
2008 "Sleeping Beauty: Theme and Variations", *PhilSci Archive* 4318

FRANCESCHI, P.
2005 "Sleeping Beauty and the Problem of World Reduction", *PhilSci Archive* 2572 (ancienne version de l'article de 2008, aux différences significatives)
2008 "A Two-Sided Ontological Solution to the Sleeping Beauty Problem", *PhilSci Archive* 4376
2009 *Les enfants d'Eubulide. Dialogue autour des paradoxes philosophiques*, à paraître

[28] Je remercie Jean-Paul Delahaye, toujours disponible pour des commentaires et des discussions très précieuses. Je remercie vivement Paul Franceschi pour son soutien chaleureux, ses commentaires et nos discussions.

GROISMAN, B.
2008 "The End of Sleeping Beauty's Nightmare", *The British Journal for the Philosophy of Science* 59: 409-416 (et *PhilSci Archive* 3624)

HITCHCOCK, C.
2004 "Beauty and the Bets", *Synthese* 139: 405-420

HORGAN, T.
2004 "Sleeping Beauty Awakened: New Odds at the Dawn of the New Day", *Analysis* 64: 10-21
2007 "Synchronic Bayesian Updating and the Generalized Sleeping Beauty Problem", *Analysis* 67: 50-59

LESLIE, J.
1996 *The End of the World. The Science and Ethics of Human Extinction*, London, Routledge

LEWIS, D.
1980 "A Subjectivist Guide to Objective Chance", *Studies in Inductive Logic and Probability*, R.C. Jeffrey, Vol. 2, pp. 263-293, Berkeley, University of California Press
2001 "Sleeping Beauty: Reply to Elga", *Analysis* 61: 171-176

LEWIS, P. J.
2007 "Quantum Sleeping Beauty", *Analysis* 67: 59-65 (et *PhilSci Archive* 2715)
2008 "Probability, Self-Location and Quantum Branching", *PhilSci Archive* 4309
2009 "Reply to Papineau and Durà-Vilà", *PhilSci Archive* 3990

MEACHAM, C.
2008 "Sleeping Beauty and the Dynamics of *De Se* Beliefs", *Philosophical Studies* 138: 245-269

MONTON, B.
2002 "Sleeping Beauty and the Forgetful Bayesian", *Analysis* 62: 47-53
2005 (avec B. Kierland) "Minimizing Inaccuracy for Self-Locating Beliefs", *Philosophy and Phenomenological Research* 70: 384-395 (et *PhilSci Archive* 1224)

NEAL, R. M.
2006 "Puzzles of Anthropic Reasoning Resolved Using Full Non-indexical Conditioning", *PhilSci Archive* 2888 (et http://www.cs.toronto.edu/~radford/papers-online.html)

OSCAR SEMINAR
2008 "An Objectivist Argument for Thirdism", *Analysis* 68: 149-155 (et http://oscarhome.soc-sci.arizona.edu/ftp/publications.html)

PAPINEAU, D. et DURÀ-VILÀ, V.
2008 "A Thirder and an Everettian: a Reply to Lewis's 'Quantum Sleeping Beauty' ", *PhilSci Archive* 3912
2009 "Reply to Lewis: Metaphysics versus Epistemology", *PhilSci Archive* 4118

PICCIONE, M. et RUBINSTEIN, A.
1997 "On the Interpretation of Decision Problems with Imperfect Recall", *Games and Economic Behavior* 20: 3-24

PUST, J.
2008a "Horgan on Sleeping Beauty", *Synthese* 160: 97-101
2008b (avec K. Draper) "Diachronic Dutch Books and Sleeping Beauty", *Synthese* 164: 281-287

VAIDMAN, L. et SAUNDERS, S.
2001 "On Sleeping Beauty Controversy", *PhilSci Archive* 324

VAN FRAASSEN, B. C.
1984 "Belief and the Will", *Journal of Philosophy* 81: 235-256
1995 "Belief and the Problem of Ulysses and the Sirens", *Philosophical Studies* 77: 7-37 (et http://www.princeton.edu/~fraassen/abstract/index.htm)

WHITE, R.
2006 "The Generalized Sleeping Beauty Problem: A Challenge for Thirders", *Analysis* 66: 114-119