Imitation et authentification

Le test de Turing et la cryptologie

Bastien Guerry

DEA de sciences cognitives
Discipline : philosophie
Directeur de stage : Daniel Andler
Année 2001-2002

Département d'Etudes Cognitives
45, rue d'Ulm, 75005 Paris

Ecole des Hautes Etudes en Sciences Sociales
56, boulevard Raspail, 75006 Paris

« Que sauriez-vous de moi si on vous disait, par exemple, que je suis d'une bonne famille, que mon père est un homme honorable, mes frères, des garçons sérieux et pleins d'avenir, et moi-même quelqu'un de très capable, d'un peu dissipé, sans doute, mais prometteur, néanmoins, de sorte qu'on pourrait dans une certaine mesure me faire confiance et cætera, et cætera. Vous ne sauriez rien de moi et vous n'auriez aucune raison après cela d'être plus tranquille si vous deviez m'engager comme vendeur dans votre magasin. »

Les enfants Tanner, Robert Walser (1878 – 1956)

Introduction

Ce mémoire est une enquête sur les notions d'imitation et d'authentification. Comme une telle enquête ne saurait être purement conceptuelle, nous arpentons essentiellement deux terrains : celui du test de Turing et celui de la cryptologie. Nous nous concentrons d'abord sur le test de Turing pour montrer qu'une analyse de l'imitation et de l'authentification permet de dégager quelques arguments en faveur de sa validité. Puis nous étudions la manière dont les travaux de Turing en cryptanalyse ont pu l'amener à l'hypothèse de la pensée des machines et à cette formulation précise du jeu de l'imitation. Enfin, nous étudions les logiques de l'authentification, lesquelles sont destinées à l'analyse des protocoles cryptographiques d'identification.

Les questions techniques de cryptologie et les questions philosophiques concernant la pensée des machines sont hétérogènes. Il ne s'agit pas d'importer en philosophie de l'esprit des notions issues de la cryptologie, encore moins d'aborder la cryptologie à la lumière de nos problèmes philosophiques. Toutefois, nous espérons montrer comment ces questions ont pu se rencontrer dans le travail de Turing, rencontre qui n'est pas étrangère aux ambitions de la « première cybernétique ». L'analyse historique et philosophique de cette rencontre constitue l'épicentre de notre mémoire. A partir de là, nous creusons les problèmes philosophiques posés par le test de Turing, puis nous cherchons dans l'analyse des protocoles cryptographiques les outils pour préciser les problèmes soulevés par la notion d'authentification.

Nous n'entrons pas dans les débats contemporains sur le rôle de la simulation dans l'acquisition ou l'utilisation d'une théorie de l'esprit. La seule conviction que nous avons gagnée à lire les théories à ce sujet, c'est que les notions d'imitation et d'authentification, toujours sollicitées, sont délicates à manipuler. Or, comme le dit si bien Goodman, « l'utilité d'une notion témoigne non pas de sa clarté, mais plutôt de l'urgence philosophique qu'il y a à la clarifier. » (Goodman, 1984, p. 54)

Partie 1

Jeu de l'imitation et dispositions

1.1 Présentation et problème

1.1.1 Présentation du jeu de l'imitation

Dans son article de 1950, Turing propose de remplacer la question « une machine peut-elle penser ? » par un jeu qu'il appelle le jeu de l'imitation. Celui-ci se joue à trois : un homme (A), une femme (B) et un interrogateur (C). L'homme et la femme peuvent chacun communiquer avec l'interrogateur par l'intermédiaire d'un télétype, mais ils ne peuvent pas communiquer entre eux. Le jeu consiste en un ensemble de questions que l'interrogateur pose tour à tour aux deux autres joueurs, chacun d'entre eux répondant selon le but qui lui est assigné. Le but de l'homme est de se faire passer pour la femme, le but de la femme est de voir son identité reconnue, le but de l'interrogateur est de déterminer l'identité des deux autres joueurs.

La question « les machines peuvent-elles penser ? » est remplacée par la question : « un ordinateur est-il capable de jouer aussi efficacement que l'homme au jeu de l'imitation ? » Est-il capable d'imiter la femme de manière telle qu'un interrogateur se trompera aussi souvent qu'avec un homme en A ? La réponse de Turing est que « dans une cinquantaine d'années il sera possible de programmer des ordinateurs, avec une capacité de mémoire d'à peu près 10^9, pour les faire si bien jouer au jeu de l'imitation qu'un interrogateur moyen n'aura pas plus de 70 pour cent de chances de procéder à l'identification exacte après cinq minutes d'interrogation. » (Anderson, 1964, p. 49). Cette réponse est pour nous un programme d'explication : nous devons expliquer le remplacement de la notion initiale de « machine » par celle d'ordinateur ; nous devons ensuite expliquer le rôle que jouent les probabilités dans le jeu de l'imitation ; nous aurons enfin à détailler les difficultés liées à la notion d'identification.

Ce jeu de l'imitation a été ultérieurement désigné comme le « test de Turing »[1]. Il soulève deux questions : (i) une machine est-elle réellement capable de le franchir avec succès ? (ii) ce

test est-il un « bon » test pour savoir si une machine est capable de penser? L'article de Turing vise à répondre positivement à ces deux questions. Mais une réponse positive à l'une n'implique pas une réponse positive à l'autre. Nous présentons en annexe (section 4.5 page 72) un aperçu des tests de Turing effectués jusqu'à présent. Dans ce qui suit, nous nous intéresserons surtout à la *validité* du test plutôt qu'au fait qu'une machine puisse le franchir avec succès.

1.1.2 Le test est-il un bon test ?

Avant d'examiner cette question, précisons ce que nous attendons d'ordinaire d'un test. Que nous testions des objets, des événements ou des comportements, le but général d'un test est de dire quels sont les éléments d'un ensemble qui tombent sous un prédicat donné : nous testons une solution pour savoir si elle est basique ou acide, nous lançons un dé un millier de fois pour savoir si la distribution des résultats est aléatoire ou non, nous testons les réactions d'un adulte face à des images pour savoir s'il en éprouve du plaisir ou non, etc. On peut bien sûr se servir d'une hiérarchie de prédicats ou d'une échelle graduée pour des mesures, mais le but est toujours de ranger une entité sous une étiquette donnée. Prenons la situation la plus simple où nous voulons tester si un x est P. Nous dirons d'un test qu'il est *cohérent* s'il exclut tous les individus ne tombant pas sous le prédicat en question. Nous dirons qu'il est *complet* s'il inclut tous les individus tombant sous le prédicat. Nous dirons qu'il est *adéquat* s'il est cohérent et complet, c.-à-d. s'il exclut tous les individus ne tombant pas sous le prédicat et s'il inclut tous les individus tombant sous le prédicat. Il peut y avoir des degrés d'incohérence et d'incomplétude : selon qu'un test inclut une certaine proportion d'individus ne tombant pas sous le prédicat ou selon qu'il exclut une certaine proportion d'individus censés tomber sous le prédicat, le test sera plus ou moins incohérent, plus ou moins incomplet. Nous dirons enfin d'un test qu'il est *fiable* si l'hypothèse de son adéquation a de bonnes chances d'être confirmée[2].

Supposons maintenant que tous les individus d'une catégorie donnée appartiennent exclusivement à P ou à $\neg P$. Soit T un test permettant de savoir si un individu de cette catégorie est P ou $\neg P$. Si T est cohérent, alors il suit immédiatement que T est aussi complet et adéquat. De même, si T est complet, alors T est aussi cohérent et adéquat. T est donc un test idéal. Mais de nombreuses difficultés nous éloignent de cet idéal. Nous en distinguons trois principales :

1. Le prédicat que nous testons est-il bien déterminé?
2. La présence ou l'absence du prédicat opère-t-elle une bipartition dans l'ensemble d'individus qui nous intéresse?
3. Que puis-je conclure d'un échec au test ?

Les remarques précèdentes ont pour fonction de nous aider à savoir ce que nous pouvons attendre du test de Turing. Nous répondons d'abord aux trois questions posées ci-dessus, puis

nous précisons l'endroit où nous comptons faire opérer notre argumentation pour défendre sa validité.

Le prédicat que nous testons est-il bien déterminé? L'article de 1950 contient un léger flottement de vocabulaire : Turing parle tantôt d'« intelligence », tantôt de « pensée ». Nous pensons que ce flottement indique seulement que Turing prend ces termes dans leur acception ordinaire, hors de toute définition philosophique ou scientifique. Aussi emploierons-nous ces deux termes indifféremment, préférant employer le terme générique d'« état mental » lorsque cela sera sans ambiguïté.

On nous objectera que ce dernier terme est tout aussi mal déterminé que ceux de « pensée » et d'« intelligence », que nous n'en donnons aucune définition philosophique rigoureuse. Nous répondrons que Turing ne s'interroge pas sur ce qu'*est* un état mental pour ensuite se demander si les machines peuvent en avoir ; il part de ce que nous supposons ordinairement être un état mental chez l'homme et se demande ensuite si nous pouvons en supposer chez les machines. Par définition, tous les êtres humains doués de parole franchissent avec succès le test de Turing et l'on s'accorde à dire que ces êtres humains « pensent » : cette définition pragmatique donne à la notion de « pensée » une définition suffisamment déterminée dans le cadre du test de Turing, car celui-ci ne teste pas en la machine la présence de quelque entité philosophiquement insaisissable, mais la présence de ce que nous attribuons chaque jour à autrui sous l'étiquette « état mental ». Une définition philosophique n'est pas un préalable nécessaire à l'analyse du test de Turing : le rôle de ce test est justement d'extraire la question de la pensée des machines hors du champ spéculatif pour y répondre par la mise en scène de notre usage commun du langage.

La présence ou l'absence du prédicat opère-t-elle une bipartition dans l'ensemble d'individus qui nous intéresse? Nous pouvons répondre positivement à cette question, pour autant que nous précisons ce que nous entendons par « l'ensemble des individus qui nous intéresse ». Les conditions du test de Turing sont telles que nous ne nous intéressons pas aux êtres humains en état d'ébriété, ni aux machines à coudre, ni aux animaux : le test a lieu dans les conditions ordinaires d'attribution d'un état mental à autrui, pour autant que celui-ci est capable de s'exprimer verbalement. Une fois que nous restreignons ainsi l'ensemble des individus qui « nous intéressent », il est raisonnable de penser que nous attribuerons des états mentaux à certains de ces individus, que nous les refuserons aux autres, et qu'il n'y aura pas de troisième sous-ensemble définitivement indéterminé.

Que puis-je conclure d'un échec au test de Turing? Rien. Le fait que le prédicat dont nous testons la présence opère une bipartition dans un ensemble d'éléments n'implique pas que

le test permette effectivement d'opérer cette bipartition. Le test de Turing vise la cohérence plutôt que la complétude. Qu'une entité échoue à se faire passer pour un être humain ne nous dit rien sur les capacités de cette entité. Ce qui est arrivé au cours du prix Lœbner de 1991 l'illustre bien : un homme a été identifié comme étant une machine, mais rien n'a été conclu au sujet des capacités de cet homme à penser.

Précisons maintenant ce que nous pouvons attendre du test de Turing, et comment nous allons nous y prendre pour prouver sa validité. Si le test de Turing est cohérent, les machines qui le franchiront avec succès pourront toutes être dites « pensantes » ; s'il est complet, toutes les machines « pensantes » le franchiront avec succès. Savoir aujourd'hui si le test est fiable, c'est savoir s'il a de bonnes chances d'être adéquat. Bien sûr, l'adéquation ne peut jamais être prouvée dès aujourd'hui, car le requisit de complétude impose que l'on ait fait passer le test à toutes les entités susceptibles de le franchir avec succès. Mais même en l'absence de certitude sur l'adéquation d'un test, on peut avoir de bonnes raisons de croire en sa fiabilité, si toutes les entités susceptibles de le franchir avec succès l'ont, à ce jour, franchi avec succès.

Il va de soi que le test de Turing n'est pas complet. A supposer certains individus capables d'avoir des états mentaux mais incapables de les verbaliser, ceux-ci ne pourront pas passer le test. Mais si nous nous restreignons à la classe d'individus que nous avons mentionnée plus haut, le problème de la complétude peut être temporairement mis de côté. En revanche, le test de Turing a de bonnes chances d'être cohérent et *de le devenir de plus en plus*. Si nous arrivons à montrer qu'il peut devenir de plus en plus cohérent, et si nous mettons entre parenthèses la question de sa complétude, nous aurons alors montré qu'il a de bonne chance d'être et de devenir de plus en plus adéquat.

Le fait que nous n'ayons ici aucune certitude n'est pas gênant. Turing ne prétendait pas résoudre un problème, mais simplement lui donner une tournure expérimentale. Il comptait éclaircir les conditions dans lesquelles la question des machines pensantes pouvait se poser, non résoudre une énigme philosophique. Si nous arrivons à justifier la cohérence du test de Turing, nous ne répondrons pas plus que lui aux questions qu'il voulait éviter, mais nous aurons rendu justice à sa démarche de clarification.

1.2 Objections et réponses

1.2.1 Objections et réponses classiques

Objections portant sur l'incomplétude du test

Une première objection[3] accuse le test de Turing d'exclure d'emblée toute une classe d'êtres vivants pour lesquels on s'accorde en général à dire qu'ils pensent : les chimpanzés, les chiens ou les enfants ne sachant pas encore parler[4]. Nous avons déjà répondu à cette objection quand nous avons délimité l'ensemble des individus pour lesquels le test de Turing est pertinent.

Une deuxième objection reproche au test de se concentrer uniquement sur des capacités verbales, de ne s'intéresser qu'à une définition désincarnée de l'intelligence ; sans organes sensoriels, comment une machine pourra-t-elle attacher un sens à des propositions dont les constituants font référence au monde ? Nous pourrions répondre à cette objection en affirmant que les termes du langage faisant directement référence au monde ne sont pas les plus communs. Mais cette réponse n'est pas seulement douteuse, elle est rendue inutile par l'argument suivant. Tant que les problèmes théoriques liés à la référence ne sont pas résolus dans le cas de l'homme, on ne peut pas faire de la capacité référentielle un critère permettant de décerner le label « intelligent ». Lorsque nous attribuons des états mentaux à autrui, nous ne soulevons ni ne résolvons la question de savoir si celui-ci possède bien la capacité à faire référence au monde. Puisque le jeu de l'imitation n'engage que la capacité d'une machine à se voir attribuer des états mentaux par un être humain, et puisque cette capacité est ce qui fait d'ordinaire qu'on dit d'un homme qu'il pense, alors nous n'avons pas à tenir compte des problèmes liés à la référence pour savoir si le test nous donne le droit d'attribuer la pensée aux machines.

Objection de la simulation

La troisième objection, que Copeland nomme « l'objection de la simulation », consiste à affirmer qu'un X simulé n'est pas un X. Une imitation d'état mental n'est pas un état mental. Turing incite lui-même à ce qu'on lui oppose cette objection lorsqu'il expose le jeu de l'imitation entre l'homme et la femme : le fait qu'un homme soit capable de simuler la féminité ne fait pas de lui un être féminin. Pour répondre à cette question, Copeland distingue entre deux sens du mot « simulation ».

Selon une première acception, une entité en simule une autre si elle lui ressemble, même sans en posséder les traits constitutifs : l'état de mort simulée est un état qui ressemble à la mort sans en possèder les traits constitutifs. Dans un deuxième sens, X' simule X si X' ressemble à X et si X' possède les traits constitutifs de X, même si les conditions de production de X' diffèrent des conditions de production originales de X : une protéine artificielle est une imitation de protéine

naturelle non seulement parce qu'elle lui ressemble, mais surtout parce qu'elle en possède les propriétés constitutives, même si elle n'a pas été synthétisée de manière naturelle. Autant la première relation de simulation ne préserve pas l'authenticité, autant la seconde la préserve : pour avoir été synthétisée de manière artificielle, une protéine n'en est pas moins naturelle. Selon Copeland, puisque la simulation de l'intelligence est une simulation du deuxième type, alors l'objection ne vaut pas. Un comportement est qualifié d'« intelligent » selon qu'il possède certains traits constitutifs, non selon la manière dont il est produit.

Nous insisterons plus loin (section 1.2.2 page 9) sur les insuffisances de cette réponse, mais nous pouvons déjà signaler une assomption que Copeland passe sous silence. Les « traits constitutifs » qu'il évoque ne peuvent pas être des traits seulement *prototypiques* : si tel était le cas, la première définition de la simulation deviendrait obscure (car on ne voit pas comment une chose peut ressembler à une autre sans qu'il existe au moins un prototype dont les deux expriment certaines propriétés) et la deuxième définition deviendrait insuffisante (car de ce que deux choses possèdent certains traits prototypiques en commun nous ne pouvons pas conclure que ces deux choses sont des exemplaires authentiques d'une même catégorie.) Si nous entendons propriétés « nécessaires et suffisantes » là où Copeland parle de traits « constitutifs », alors les deux définitions sont valables. Mais l'hypothèse selon laquelle il existe des propriétés nécessaires et suffisantes définissant l'intelligence d'un comportement – hypothèse nécessaire à la réponse de Copeland – est pour le moins douteuse.

Objection de la boîte noire

La quatrième objection tire parti de la formulation béhavioriste du test de Turing pour imaginer une situation dans laquelle sa portée sera invalidée. Le jeu de l'imitation traite les joueurs A et B comme des boîtes noires. L'attribution d'états mentaux à ces deux joueurs par l'interrogateur ne s'appuie que sur l'observation de leur comportement verbal. Aussi peut-on imaginer une machine construite de telle manière qu'elle passe avec succès le test de Turing, mais dont nous saurions *à l'avance* qu'elle ne possède aucun état mental. Le nombre de mots d'une langue donnée est fini. A partir de cet ensemble fini de mot, nous ne pouvons construire qu'un ensemble fini de phrases de moins de cent lettres. A l'intérieur de cet ensemble fini d'énoncés, il n'y a qu'un sous-ensemble fini d'énoncés syntaxiquement corrects, à l'intérieur duquel il n'y a qu'un sous-ensemble fini d'énoncés sensés. A partir de cet ensemble fini d'énoncés ayant du sens, nous ne pouvons construire qu'un ensemble fini de couples question-réponse ayant eux-mêmes du sens. Puisque cet ensemble de couples question-réponse est fini, on peut imaginer un super-ordinateur les ayant tous gardés en mémoire. Cette machine trompera parfaitement un interrogateur, mais nous savons que ce n'est qu'une boîte à « réflexes ».

On peut objecter à cette expérience de pensée que la limite des cent lettres est arbitraire :

mais les conditions de réalisation du test imposent qu'une limite soit choisie, même arbitrairement. De même, supposer que l'ensemble des énoncés ayant un sens est un sous-ensemble des énoncés syntaxiquement corrects revient à ignorer une large part du sens que nous trouvons dans certains écrits poétiques : mais encore une fois, les contraintes pratiques du test sont telles que nous pouvons nous permettre d'ignorer ce fait marginal. Turing ne parle-t-il pas lui-même d'un « interrogateur moyen » (Anderson, 1964, p. 49) ? Nous ne pouvons pas répondre directement à cette objection. Il nous faut d'abord voir quelles assomptions implicites ont été faites pour construire cette expérience de pensée, et c'est en montrant que certaines de ces assomptions sont illégitimes que nous refuserons d'en accepter les conclusions.

1.2.2 Problèmes concernant l'objection de la simulation

De la réponse de Copeland à l'objection de la simulation, nous conserverons la distinction entre les objets dont l'authenticité dépend des conditions de leur production (comme les fourrures) et les objets dont l'authenticité ne dépend que de la possession de certaines propriétés constitutives (comme les protéines) – à condition toutefois de préciser ce que nous entendons par « constitutives ». Cette distinction peut s'éclairer par la distinction entre *autographique* et *allographique* que propose Goodman dans (Goodman, 1968, p. 147). Nous l'introduisons ici, car nous l'utiliserons ensuite dans notre tentative de répondre à l'objection de la simulation (cf. section 1.5.3 page 22).

Se posant la question des critères d'authenticité d'une œuvre d'art, Goodman remarque que la distinction entre un original et une contrefaçon ne se fait pas de la même manière selon les différents arts. Si l'authenticité d'une œuvre dépend de ses conditions de production, elle sera dite *autographique*. Si une œuvre d'art ne consiste qu'en la classe des exécutions fidèles à une partition[5], l'œuvre sera dite *allographique*. Le premier cas correspond typiquement aux œuvres picturales, le second aux œuvres musicales. Ce qui fait l'authenticité d'une exécution de la dixième symphonie de Mahler est la fidélité de l'orchestre à la partition de cette symphonie, la correction orthographique de la transcription sonore. Une exécution orthographiquement incorrecte ne rend pas l'œuvre moins authentique, elle la fait sortir de la classe d'exécution définissant *cette* œuvre. Dans le cas de peinture, l'authenticité de l'œuvre est entièrement déterminée par ses conditions de production : il suffit que l'une de ces conditions ne soit pas respectée pour que nous soyons en présence d'une contrefaçon[6].

En usant librement de la terminologie goodmanienne, nous pouvons dire que la fourrure est un objet autographique : son authenticité est garantie par ses conditions de production, par le fait qu'elle provient d'un animal. En revanche, une protéine est un objet allographique : la question de l'authenticité entre une protéine naturelle et une protéine artificielle ne se pose pas, car nous n'accordons l'étiquette « protéine » qu'en vertu de la conformité d'une substance

à sa description scientifique[7]. D'après le raisonnement suivi par Copeland, l'intelligence d'un comportement ou d'un individu serait allographique : les conditions de production du comportement intelligent ne sont pas pertinentes pour juger de son authenticité, seule importerait la possession des traits constitutifs de l'intelligence.

Nous sommes en mesure de formuler les questions pour lesquelles cette réponse reste muette. (i) Est-ce qu'un état mental est définissable en termes de « traits constitutifs » ? (ii) Si oui, ces « traits constitutifs » sont-ils des traits prototypiques ou des propriétés nécessaires et suffisantes[8] ? (iii) Quel est le rapport d'un état mental à ses conditions de production[9] ? Pour répondre complètement à l'objection de la simulation, nous devrons répondre à ces questions en montrant dans chaque cas ce qu'il advient de la portée du jeu de l'imitation.

1.3 Analyse de l'imitation

Dans cette section, nous proposons une définition générique de l'imitation et nous tentons de comprendre ce qui se passe lorsque celle-ci concerne spécifiquement un objet, un comportement ou une disposition. La distinction de ces trois cas prépare l'analyse du rôle de l'imitation dans le test de Turing. Nous soulignons à la fin ce qui fait la difficulté de l'imitation et de l'authentification des dispositions.

1.3.1 L'imitation comme double référence

Il a souvent été montré que la relation de ressemblance ne suffisait pas à définir la relation de représentation ; dire qu'une chose en représente une autre *parce qu'*elle lui ressemble est insuffisant, car cela ne dit pas de quel point de vue est émis le jugement de ressemblance. Mais on a plus rarement souligné le fait que la relation de ressemblance ne suffit pas non plus à définir la relation d'imitation. C'est la thèse que nous voudrions maintenant défendre, et à partir de laquelle nous construirons notre définition de l'imitation.

Représentation et ressemblance

Précisons d'abord la distinction entre représentation et ressemblance. Quand nous cherchons à définir la notion de représentation à l'aide de la notion de ressemblance, nous confondons deux modes de référence distincts : la dénotation et l'expression. Ce qu'un symbole *dénote*, c'est ce dont il prend la place : il n'y a pas nécessairement de lien de ressemblance entre le symbole et ce qu'il symbolise, comme le montre le cas souvent évoqué des symboles de la langue écrite. En suivant Goodman, nous parlerons de ce qu'un symbole *exprime* pour désigner les propriétés qu'il exemplifie. Soit une crucifixion du Greco : le tableau dénote la scène de la mise en croix du

Christ et exprime certaines nuances de gris-chair, des formes tortueuses, de la douleur. Ce que ce tableau exprime est ce à quoi il renvoie non du fait de ce qu'il représente, mais du fait d'être classé parmi les objets possédant la propriété d'être colorés en nuances de gris-chair, d'exhiber des formes tortueuses, de susciter de la tristesse[10].

La ressemblance entre deux objets provient du fait qu'ils expriment certaines propriétés en commun, non de ce qu'ils dénotent tous les deux la même chose. Deux maisons peuvent se ressembler du fait qu'elles ont toutes les deux un toit de chaume ; ces maisons ne dénotent rien. Deux mots de langages différents peuvent tous deux désigner le même oiseau et ne se ressembler en rien. Quand nous disons de deux tableaux de crucifixion qu'ils se ressemblent, nous confondons souvent la similitude entre les scènes qu'ils dénotent et la similitude entre les propriétés qu'ils expriment, alors que celle-ci n'est qu'une conséquence indirecte de celle-là. Deux tableaux pourraient représenter la mise en croix du Christ sans se ressembler du tout. Il est certes courant qu'un symbole exprime certaines propriétés de ce qu'il dénote. Une poupée vaudou doit exprimer certains traits de l'homme dont elle prend symboliquement la place, notamment ceux de ces traits sur lesquels l'action magique doit opérer. Confrontés à des symboles de ce genre, nous sommes conduits à confondre représentation et ressemblance. Mais ne pas garder cette distinction à l'esprit mettrait en péril notre tentative de définition de l'imitation.

Imitation et référence

La définition de l'imitation fait nécessairement appel à la notion de ressemblance. Mais de ce que X ressemble à Y, il ne s'en suit pas nécessairement que X est une imitation de Y. Que manque-t-il ? Nous réservons en général le terme d'« imitation » ou de « simulation »[11] à des situations dans lesquelles la relation de ressemblance à été voulue : un imitateur prend intentionnellement la voix de Chirac, un copiste peint volontairement une réplique de l'original, un faussaire a consciemment pour but de produire la plus grande ressemblance possible entre vrais et faux billets. Dans ces trois cas, le fait que l'imitation soit intentionnelle crée, en plus du rapport de ressemblance, un lien de représentation entre ce qui imite et ce qui est imité. C'est parce que l'acteur prend intentionnellement la voix de Chirac que nous savons que ce n'est pas sa propre voix, et que la voix qu'il émet représente la voix de Chirac. De même, la copie d'un tableau dénote l'original en ceci que, sans l'original, l'intention du copiste n'aurait pu s'appuyer sur rien. Enfin, la volonté du faussaire s'appuie nécessairement sur la représentation d'un authentique billet de banque, représentation que nous retrouvons dans le fait que le faux billet dénote le vrai. Sans référence dénotative entre ce qui imite et ce qui est imité, l'imitation n'est qu'une reproduction aveugle ; sans référence expressive entre l'imitation et l'original, l'imitation n'est qu'une intention vide.

L'analyse de ce qu'il y a d'intentionnel dans une imitation permet d'éclairer la nécessité de la relation de dénotation existant entre l'imitation et ce qui est imité. Mais l'intentionnalité peut être engagée dans l'imitation de différentes manières. Il peut y avoir des différences temporelles : l'imitation de Chirac est contemporaine de l'acte d'imiter ; une contrefaçon de Munch survit à l'original. Il peut y avoir des différences d'intensités : l'intention d'imiter peut être plus ou moins forte. Troisièmement, l'intention d'imiter peut être manifeste ou cachée : le faussaire veut produire la ressemblance et cacher son intention de la produire ; l'imitateur veut produire la ressemblance mais ne se préoccupe pas nécessairement de cacher cette intention. Enfin, il y a des cas où nous voudrions parler d'imitation sans savoir cependant si la relation de représentation est intentionnelle. Dans le cas d'un phasme imitant une brindille, il s'agit bien d'une imitation : le phasme ressemble à la brindille *et* la représente pour le système cognitif du prédateur. Mais il serait périlleux de s'interroger sur son intention d'imiter.

Puisque nous cherchons une définition générique de l'imitation, nous ne souhaitons pas tenir compte dans un premier temps de toutes ces nuances liées à l'engagement de l'intention dans l'imitation. Il nous suffira de dire que toute imitation est définie non seulement par la relation expressive de ressemblance avec ce qui est imité, mais aussi par la relation dénotative avec ce que l'imitation représente. Nous aboutissons donc à la définition suivante :

Imitation X imite Y si et seulement si X exemplifie certaines propriétés de Y et si X dénote Y.

Ceci nous permet en retour de préciser la distinction que nous faisons entre *reconnaissance*, *identification* et *authentification*. Reconnaître une entité n'est rien d'autre que la classer correctement. Authentifier une entité consiste non seulement à la classer correctement, mais à reconnaître en plus la nécessité de ce classement. Alors que la reconnaissance n'implique pas de retour réflexif sur les pratiques de classification, l'authentification met en jeu cette pratique soit dans une justification (« C'était bien du cuir. ») soit dans un démenti (« Ce n'était pas de la fourrure. »). Une ressemblance entre deux entités peut nous égarer et perturber nos habitudes de classification, mais elle ne peut pas nous *tromper*. En revanche, une imitation peut nous tromper en voilant le fait qu'elle renvoie à autre chose qu'elle-même, c.-à-d. en rendant impossible la reconnaissance de la nécessité d'un classement.

L'identification est à part. D'un côté, elle est cas de reconnaissance où la classe considérée ne contient qu'un individu. De l'autre, elle semble être un cas d'authentification, car la reconnaissance d'une entité comme étant elle-même semble impliquer que l'on reconnaisse la nécessité de cette relation d'identité. Nous ne débattrons pas de ces difficultés qui, de Frege à Perry (cf. annexe, section 4.6 page 73), ne cessent d'intriguer. Dans le doute, nous considérerons l'identification comme un cas particulier de reconnaissance.

1.3.2 L'objet de l'imitation

La définition que nous venons de donner de l'imitation est une définition générique qui ne préjuge pas de la nature des entités imitées. Nous regardons maintenant ce qui se passe lorsqu'une imitation renvoie à un objet, à un comportement ou à une disposition. Dans ces trois cas, nous supposerons qu'il y a un sens à vouloir authentifier l'original d'une contrefaçon. Autrement dit, nous ferons comme si l'authenticité d'un objet (un cuir), d'un comportement (une manie) ou d'une disposition (la fragilité) résidait dans ses conditions d'existence. La dépendance aux « conditions d'existence » recouvre aussi bien la dépendance à une chaîne causale (conditions de production) que la dépendance à des éléments concomitants. Pour le cuir, ce sera par exemple le fait de provenir d'un animal ; pour le comportement maniaque, ce pourra être le fait d'être causé par un dérèglement hormonal ; pour la fragilité, cela sera par exemple le fait d'être en rapport avec une structure cristalline. Notre but est de montrer que, dans le cas des dispositions, cette supposition mène à une impasse : nous ne pouvons pas nous référer de manière exhaustive aux conditions d'existence d'une disposition pour juger de son authenticité.

Dans chacun des trois cas d'imitation, le but est de partager avec ce qui est imité le plus grand nombre possible de propriétés. Comme nous ne nous intéressons qu'à des exemples d'imitation « autographique », les propriétés à partager ne sont pas des propriétés nécessaires et suffisantes (car celles-ci nous fourniraient la « partition » d'une entité), mais des propriétés seulement prototypiques. Par exemple, un faux cuir reproduira la rigidité et la brillance d'un vrai cuir ; une simulation de comportement maniaque copiera l'ensemble des gestes liés au comportement maniaque ; un verre en plastique exemplifiera la transparence et la finesse pour nous faire croire à sa fragilité. Si l'imitation a en plus pour fonction de se faire passer pour l'original, alors elle cherchera à ce que ses propriétés correspondent si étroitement aux propriétés de ce qu'elle imite que le lien de dénotation entre l'une et l'autre entité soit inaccessible à un spectateur ordinaire[12].

Dans une imitation d'objet, les propriétés exemplifiées par l'imitation sont des propriétés observables : l'imitation sera comprise dans la plupart des classes de prédicats manifestes applicables à ce qui est imité. Le faux cuir pourra être aussi souple et brillant que le vrai cuir. La possession des mêmes traits prototypiques fera illusion si le client n'a pas accès à la différence entre les conditions de production du vrai et du faux cuir, c.-à-d. s'il n'a pas accès au lien de dénotation qui existe entre les deux cuirs (lien par lequel le faux n'est pas seulement une reproduction mais bien une contrefaçon.) Mais les différences entre ces conditions de productions sont des différences *observables*. S'il ne peut pas observer les conditions de production d'un cuir pendant qu'il est dans le magasin, le client sait néanmoins que ces conditions de production ont pu être observées par un organisme chargé de garantir l'authenticité des cuirs.

Dans une simulation de comportement, les propriétés exemplifiées par la simulation sont

aussi des propriétés observables. Bien qu'il soit plus difficile d'observer les conditions de production d'un comportement, cette observation reste possible. Si je ne connais pas bien la personne que j'ai en face de moi, je m'apercevrai difficilement que son comportement maniaque n'est qu'une simulation, mais je peux toujours me procurer des relevés médicaux qui me diront si elle est victime ou non du dérèglement hormonal authentifiant la maladie. En m'apercevant qu'elle n'en est pas victime, je me rends compte que son comportement n'était qu'une simulation. Je réalise alors que sa simulation ne faisait que ressembler au comportement authentiquement maniaque, et qu'elle n'a pu être entreprise que si la personne s'est d'abord représenté les traits prototypiques d'un dérèglement maniaque. Autrement dit, l'accès aux relevés médicaux me permet de découvrir le lien dénotationnel existant entre le comportement simulé et le comportement authentique, lien dont j'impute l'origine à l'intention qu'avait la personne de me tromper.

Dans le cas de la fragilité du verre ou de la fiabilité d'une personne, les données du problème sont différentes. Concernant l'observation des conditions d'existence d'une disposition, nous ne sommes plus en face d'une difficulté seulement pratique, mais en face d'une impossibilité théorique. En effet, comme nous le détaillons en annexe (section 4.7 page 76), la réduction d'un terme dispositionnel à un ensemble fini de propriétés observables est problématique. La transparence est certes un bon indice de ce que le verre possède une structure cristalline, et la présence de cette structure cristalline est un bon indice de la fragilité du verre, mais nous ne disposons pas de critère ultime nous permettant de dire qu'un ensemble de propriétés *suffit* à définir une disposition.

Dans le cas d'un objet et d'un comportement, l'authenticité est déterminable par la référence à un nombre fini de conditions d'existence observables ; mais un nombre fini de propriétés observables ne nous garantira jamais qu'un verre est authentiquement fragile ou qu'une personne est authentiquement fiable. Même la possession d'un ensemble de propriétés jugées fondamentales pour la possession d'une disposition ne suffisent pas à l'authentifier, pas plus que la possession d'un ensemble prototypique de traits ne permet d'évaluer avec certitude l'authenticité d'un objet ou d'un comportement.

Nous avons dit au début de cette section que nous ne nous intéresserions qu'à des objets, comportements ou dispositions pour lesquels l'entreprise d'authentification avait un sens, c.-à-d. – suivant la terminologie de Goodman – à des entités autographiques. Cette précaution n'était pas utile pour les dispositions : étant donné qu'une disposition n'est jamais descriptible à partir de propriétés nécessaires et suffisantes, elle ne peut pas jamais être allographique. En même temps qu'une disposition pose le problème de son authentification, elle nous confisque les moyens de le résoudre.

1.4 Béhaviorisme logique et dispositions

Dans ce qui suit, nous rappelons certaines thèses du béhaviorisme logique qui nous semblent éclairer le jeu de l'imitation. Nous montrons le rapport entre le problème de la description des états mentaux en termes de dispositions et le problème de la réduction des dispositions à des propriétés observables. Enfin, nous montrons comment le test de Turing ne donne pas seulement à l'interrogateur pour tâche de reconnaître des comportements, mais aussi d'authentifier des dispositions, étape nécessaire aux arguments que nous formulons pour soutenir la validité du test.

1.4.1 Etats mentaux et dispositions

On doit à Gilbert Ryle (Ryle, 1949) d'avoir insisté sur ce qu'il baptise les « erreurs de catégories ». Nous commettons une erreur de catégorie lorsque nous attendons de plusieurs entités hétérogènes qu'elles se soumettent de la même manière aux mêmes formes de jugement. Par exemple, un étranger à qui l'on aura fait visiter tous les collèges de Cambridge pourra demander : « Ces collèges sont très beaux, mais où est l'université ? » L'étranger commet une erreur de catégorie du fait qu'il inclut l'université dans la classe des bâtiments scolaires, alors qu'elle n'est que l'entité abstraite organisant l'ensemble des collèges. Cet exemple montre que la gravité d'une erreur de catégorie ne provient pas tant de la réunion dans une même classe d'entités de type hétérogène, mais surtout de ce qu'on attend de ces entités hétérogènes qu'elles se comportent comme des entités correctement classées ou, d'après l'expression que Ryle aime employer, qu'elles fassent « le même travail ».

A partir de la définition de l'erreur de catégorie, Ryle entend dissoudre la plupart des faux problèmes qui apparaissent dans les analyses des phénomènes mentaux. Il montre comment le mythe intellectualiste d'un monde mental derrière le monde physique naît de ce qu'on attend de l'esprit le même genre de pouvoirs causaux que ceux attribués à la matière depuis Newton. Contre cette idée, il insiste sur le fait que la pensée ne fait pas partie d'un règne privé, imperméable à toute observation extérieure, et accessible depuis le seul point de vue du sujet. La pensée ne doit pas s'analyser en terme de propriétés invisibles, mais en terme de dispositions[13]. Ceci implique que les énoncés sur les états mentaux ne sont pas catégoriques, mais hypothétiques : un état mental ne nous est accessible que comme un ensemble de comportements observables associés aux conditions hypothétiques dans lesquels ils surviennent. Aussi, quand nous rendons compte d'un comportement comme étant « intelligent », nous ne désignons pas réellement une *cause* de ce comportement, nous ne faisons qu'en rendre logiquement raison d'après nos observations.

1.4.2 Difficultés de la définition des états mentaux comme dispositions

L'analyse des états mentaux en termes dispositionnels ne va pas sans difficultés. Distinguons parmi elles la difficulté de *traduire* un état mental comme un ensemble fini de dispositions et celle de *réduire* une disposition à un ensemble fini de comportements observables. Pour le détail des difficultés liées à cette réduction, voir section 4.7 page 76. Nous tenterons à la fin de déceler l'origine commune de ces deux types de difficultés.

En faisant de toute attitude propositionnelle une disposition, nous sommes dès à présent sûrs de ne pas pouvoir embrasser l'ensemble des comportements observables définissant cette attitude propositionnelle, à cause de l'impossibilité de définir une disposition comme une conjonction finie de propriétés observables. A partir de la définition des dispositions comme disjonction de propriétés, nous pourrions être tentés de définir une attitude propositionnelle comme une disjonction de propriétés observables liées à la classe d'individus pour lesquels ces propriétés sont caractéristiques de l'attitude considérée. Si, lorsqu'ils croient qu'il va pleuvoir, les Corses prennent leur parapluie et les Bretons mettent une vareuse, alors la croyance « Il va pleuvoir » pourrait être définie comme $\langle$ *les-Corses-prennent-leur-parapluie* $\vee$ *les-Bretons-mettent-leur-vareuse* $\rangle$. Mais, en plus de l'aspect peu maniable de cette définition, il est douteux que nous puissions par cette voie obtenir une définition complète.

En plus de ces difficultés liées à la réduction des dispositions, la traduction des états mentaux en termes de dispositions est elle-même problématique. Soit par exemple l'attitude propositionnelle[14]:

> (1) Jean croit que la neige est blanche.
>
> (1a) Jean est disposé à énoncer la phrase « La neige est blanche » ou un équivalent logique.
>
> (1b) Jean désire communiquer sa croyance que la neige est blanche.
>
> (1c) Jean croit qu'en énonçant la phrase « La neige est blanche », il communiquera sa croyance que la neige est blanche.

Nous pourrons tenter de traduire (1) à l'aide de l'énoncé dispositionnel (1a). Mais l'évaluation de la vérité de (1a) nécessite que l'on évalue la vérité des autres énoncés (1b) et (1c). Or, d'une part cet ensemble d'énoncés à évaluer afin d'évaluer (1a) n'est pas clos, d'autre part (1b) et (1c) sont eux-mêmes des énoncés dispositionnels qui exigeront l'évaluation de la vérité d'autres énoncés dispositionnels, etc.

Cette difficulté de « traduction » est-elle propre aux états mentaux? Nous voudrions suggérer que non. La distinction entre les difficultés de traduction et de réduction est pratique, mais elle n'est pas fondamentale. Dans les deux cas, la difficulté vient de ce que les jugements dispositionnels voilent des jugements hypothétiques, lesquels expriment les limites de ce que

nous pouvons connaître d'une chose ou d'un individu. Ces limites font que nous ne pouvons pas spécifier l'ensemble des conditions dans lesquelles une disposition se réduirait à des propriétés observables : mais les conditions que nous ignorons peuvent aussi bien porter sur des observations qui nous sont inaccessibles (et nous retrouvons alors le problème de la réduction) que sur d'autres dispositions (et nous retrouvons le problème de la « traduction »). La fragilité d'un verre peut aussi bien reposer sur la présence indécelable à l'œil nu d'une structure cristalline particulière que sur sa fusibilité à telle température.

1.4.3 Mise en jeu des dispositions

D'après une interprétation rudimentaire du test de Turing, l'interrogateur range les différents comportements verbaux du joueur qu'il interroge soit dans la case « comportement-d'être-humain » soit dans la case « comportement-de-machine » : une erreur de rangement prouve seulement que le comportement du joueur simule bien l'intelligence, non qu'il est réellement intelligent. Cette vue nous semble réductrice, et l'insistance sur les conditions originales du test de Turing peut nous aider à voir pourquoi. Dans le test original, (i) la comparaison a lieu en temps réel entre deux joueurs ; (ii) le jeu de l'imitation comporte une phase où les deux joueurs sont des êtres humains, la machine venant après pour remplacer l'un d'eux ; (iii) la machine doit être capable de simuler *n'importe quel* comportement sur lequel l'interrogateur juge bon de la tester.

Nous voudrions montrer que les deux premières conditions ont pour effet de transformer la tâche de reconnaissance des comportements en problème d'authentification de dispositions, et que la troisième condition implique qu'il est impossible de spécifier à l'avance l'ensemble des comportements que la machine sera « disposée » à imiter si elle franchit avec succès le test. Pour abréger la discussion, nous désignerons par PV_i le point de vue de l'interrogateur, PV_m celui de la machine et PV_e celui d'un observateur extérieur, capable d'observer les trois joueurs. Considérons maintenant l'effet que cela fait d'être un interrogateur, une fois que l'on garde en tête les conditions originales du jeu de l'imitation.

Plaçons-nous d'abord dans le « premier » jeu : l'interrogateur doit déterminer lequel, de l'homme ou de la femme, est vraiment la femme. Si l'interrogateur n'était en face que d'une entité, il ne pourrait déterminer son sexe qu'en comparant ses réponses à un modèle de comportement qu'il tiendrait pour typiquement[15] féminin. Le but de l'homme ne serait alors que d'imiter le comportement féminin, de préférence celui correspondant le mieux au modèle probable d'un interrogateur moyen.

Le fait de se retrouver en face de deux joueurs implique pour l'interrogateur que son modèle de comportement féminin n'est plus la seule référence qu'il possède pour évaluer la féminité d'un comportement. Il sait la pauvreté de ce modèle par rapport au comportement authentiquement

féminin de l'un des deux joueurs. Autrement dit, son but n'est plus de *reconnaître* un comportement relativement à une norme fixe, mais d'*authentifier* l'identité de la femme, cette identité étant attachée à un ensemble non spécifiable de comportements – soit à une disposition. Depuis PV_i le problème est de savoir si les deux comportements observés sont des échantillons d'une disposition liée à l'identité féminine ou d'une disposition à imiter la féminité. Depuis PV_m ou depuis PV_e, nous percevons cette disposition comme une *pré*disposition ; mais ce n'est pas ce qui est authentifiable depuis PV_i, car ce point de vue implique des limites dans l'ensemble des comportements que l'interrogateur observe. Le fait qu'il y ait deux joueurs introduit un élément dynamique dans les comparaisons que peut faire l'interrogateur, cet élément dynamique étant ce qui fait qu'il se voit obligé de comparer des dispositions et non seulement des comportements observables.

Qu'introduit de plus le fait que la machine remplace l'homme ? Pourquoi ne pas avoir directement décrit le jeu comme se déroulant entre une machine et une femme ? Comme le remarque J.-P. Dupuy dans (Dupuy, 1999, p. 32) : « la machine doit désormais simuler [...] la capacité de simulation de [l'homme]. » Mais ce redoublement du mot « simulation » risque ici d'être trompeur : la machine ne doit pas imiter l'homme en train d'imiter, elle doit seulement imiter la femme aussi bien que lui. L'intérêt du remplacement de l'homme par la machine doit se juger du point de vue de l'interrogateur : puisque faire la différence entre l'homme et la femme reviendra pour lui à comparer la disposition à être une femme à la disposition à seulement imiter une femme, alors ce sera encore cette comparaison entre des dispositions que fera l'interrogateur quand on remplacera (à son insu) l'homme par une machine – quand bien même, du point de vue de la machine, il paraît « déviant » de dire qu'elle imite une disposition.

La dernière condition est certainement la plus importante et la plus négligée dans les tests de Turing que nous sommes actuellement en mesure de faire passer à une machine (cf. annexe, section 4.5 page 72.) Les machines que Turing juge être de bonnes candidates au jeu de l'imitation sont des machines universelles. « Machine universelle » renvoie ici au concept élaboré par Turing dans (Turing et Girard, 1995) : il définit une machine telle que le fonctionnement de n'importe quelle machine particulière puisse être reproduit en elle par un programme approprié. Ceci ne signifie pas qu'une machine universelle puisse être programmée pour reproduire *tous* les comportements possibles : pour jouer au jeu de l'imitation, il suffit qu'elle soit programmée pour adopter ceux des comportements qu'un interrogateur moyen juge « intelligents ».

C'est ici qu'intervient la distinction que nous avons faite entre PV_i et PV_e. Si un observateur extérieur au test de Turing pouvait connaître l'ensemble des comportements que l'interrogateur juge intelligents, alors l'imitation effectuée par la machine ne serait que l'imitation de l'un de ces comportements. Mais le fait de dire qu'une bonne candidate au test est une machine capable d'imiter tous les comportements que l'interrogateur est *disposé* à tenir pour intelligents

indique qu'il n'est pas possible, pour un observateur extérieur, de déterminer à l'avance et de façon exhaustive l'ensemble de ces comportements. Pour que ceci soit possible, il faudrait que soit aussi possible un point de vue *absolument* extérieur nous indiquant l'ensemble des comportements que nous jugeons intelligents.

Nous pouvons résumer ainsi la dimension nouvelle prise par le test de Turing lorsque nous y introduisons les dispositions : le jeu de l'imitation ne permet pas de décider *directement* de la présence ou de l'absence des états mentaux dans une machine, mais de la présence ou de l'absence de la disposition à imiter ce qu'un interrogateur est disposé à tenir pour un état mental.

1.5 Réponses aux objections

1.5.1 Forme générale de l'argument

Dans (Goodman, 1968, p. 135–161), Goodman demande à quelle condition une différence esthétique entre un tableau et sa copie parfaite existe. « Parfaite » signifie ici « imperceptible *maintenant* pour *ce* spectateur. » La réponse de Goodman est qu'une différence esthétique importe pour *ce* spectateur dans *ces* circonstances si et seulement s'il sait qu'une différence dans les conditions de production de ces deux tableaux est perceptible pour un *certain* observateur dans *certaines* circonstances. Il n'est pas nécessaire que le spectateur sache quel expert ni dans quelles circonstances. Cet expert peut être déjà mort ou naître dans un million d'années, il suffit qu'on soit assuré de son existence et de son pouvoir d'expertise pour justifier en toutes circonstances la recherche d'une différence esthétique entre deux tableaux apparemment identiques. Inversement, le moindre doute quant à la possibilité de s'assurer de l'existence passée, présente ou future d'un tel expert[16] ruine complètement l'entreprise de distinction esthétique.

Si les états mentaux sont autographiques, un problème semblable peut se poser. A quelle condition existe-t-il une différence entre un état mental et sa simulation parfaite ? De même que le terme « parfait » était relatif aux circonstances particulières dans lesquelles un spectateur cherchait à comparer deux tableaux, il signifie ici « imperceptible *maintenant* pour *ce* spectateur. » Puisque les états mentaux ont été supposés autographiques, nous devons faire la même réponse que celle que nous avons faite au sujet des tableaux : il y a une différence maintenant pour ce spectateur entre un état mental et sa copie si et seulement s'il sait qu'une différence dans les conditions d'existence de ces deux états mentaux est perceptible pour un *certain* observateur dans *certaines* circonstances. Notre spectateur n'a pas à savoir quel est cet observateur expert ni dans quelles circonstances son expertise est possible : il suffit qu'il soit assuré de son existence.

Supposer que les tableaux et les états mentaux sont des entités autographiques n'implique

pas que l'authentification procède de la même manière dans les deux cas, aussi devons-nous marquer deux différences importantes. D'une part un tableau est un objet persistant, ce qui rend pertinente la référence à un expert passé ou futur ; les états mentaux étant transitoires, l'expert nous garantissant la possibilité d'une discrimination entre état mental authentique et simulé devra être présent. D'autre part, les conditions de production d'un tableau sont par principe observables, ce qui rend délicate l'entreprise de montrer qu'un observateur de ces conditions de production – du moins de celles qui sont pertinentes pour établir la différence avec une contrefaçon – peut ne pas exister. A l'inverse, la question de l'observabilité des conditions d'existence d'un état mental est problématique.

Avant de formuler notre argument général, circonscrivons sa portée. Voici l'état du problème. (i) Un état mental est-il autographique ou allographique ? Si un état mental est allographique, il n'y a pas de problème d'authentification à proprement parler, seulement un problème de reconnaissance. Si un état mental est autographique, nous devons nous poser la question : (ii) comment avons-nous accès à ses conditions d'existence ? Si on soutient qu'elles sont observables, alors authentifier un état mental revient à vérifier par l'observation la présence des conditions qui le rendent authentique. Si on soutient au contraire que les conditions d'existence d'un état mental ne sont pas observables, on dira (par exemple) que certaines propriétés d'un état mental comme les propriétés phénoménales ne sont accessibles qu'à la première personne. Si, sortant de l'opposition entre observable et non observable, nous décrivons les états mentaux en termes de dispositions, que devient l'authentification ? Comme nous l'avons vu plus haut (cf. section 1.3.2 page 13), l'authentification d'une disposition est à la fois légitime et impossible. En cherchant à authentifier un état mental défini comme disposition, nous pourrons repérer par l'observation un certain nombre de traits comportementaux qui sont de bons indices de cet état mental, mais nous n'aurons pas d'indice décisif.

Si – et c'est là que nous répondons aux objections – un état mental est défini en termes dispositionnels, il restera donc toujours une incertitude sur l'existence d'un expert en états mentaux. Refuser cet élément d'incertitude, c'est croire qu'il existe un point de vue à partir duquel la disposition définissant un état mental peut être parfaitement « résorbée » dans un ensemble de comportements observables. Les deux objections que nous discutons relèvent du même procédé consistant à se placer depuis deux points de vue incompatibles et, sans reconnaître l'incompatibilité de ces points de vue, à la projeter sur ce qui est observé. Quand nous nous tenons depuis PV_m et PV_i, nous réduisons à une prédisposition ce qui n'apparaît à l'homme que comme une disposition. Donc nous jugeons avec certitude que l'homme est trompé *et* que la simulation n'est qu'un simulacre. Quand nous nous tenons depuis PV_i et PV_e, nous réduisons à une prédisposition ce qui ne peut apparaître à un observateur extérieur que comme une disposition – à savoir la disposition d'un interrogateur à qualifier d' « intelligent »

un comportement. Nous jugeons alors qu'il y existe un comportement seulement mécanique capable de tromper cet interrogateur.

1.5.2 Réponse à l'objection de la boîte noire

L'argument de la boîte noire consistait à montrer que pour tout ensemble de comportements verbaux qu'un interrogateur tient pour « intelligents » (ou conformes à ce qu'il attend d'un être humain), il existe au moins une machine de Turing dont le programme permet de reproduire ce comportement de la façon la plus mécanique et la moins intelligente qui soit. Dans (Copeland, 1993), l'auteur attribue la création de ce super-ordinateur à des Martiens. Ici, les humains tiennent lieu d'interrogateurs potentiels et les Martiens d'observateurs potentiels : ceux-ci connaissent à l'avance le corpus de comportement que les humains tiennent pour intelligents. Notre réponse à l'objection consistera à dire que nous ne pouvons pas être des êtres humains *et* des Martiens, même en imagination. Nous ne pouvons pas sortir de notre pratique d'attribution des états mentaux pour en prendre un « cliché », pour déterminer de l'extérieur l'ensemble de nos prédispositions à attribuer tel ou tel état mental dans telle circonstance.

Une condition essentielle du test de Turing était que l'interrogateur était libre de poser n'importe quelle question aux deux joueurs. La réfutation de l'argument de la boîte noire ne consiste pas à dire que nous n'avons pas le droit de ne considérer qu'un nombre fini de lettres, de mots et phrases pour construire une machine à donner les réponses attendues, mais à dire que nous ne sommes pas capables de sortir du langage pour déterminer si une réponse est sensée, de sortir de la pratique par laquelle nous qualifions un comportement d'intelligent pour constituer une théorie parfaitement codifiée de cette pratique. Sans cette supposition, il existe toujours un doute pour l'interrogateur quant à l'existence d'un expert lui permettant de dire que son attribution d'état mental était justifiée. Après tout, nous ne pouvons pas être des machines, des hommes et des Martiens.

En faisant l'hypothèse que des Martiens ont pu déterminer l'ensemble des comportements verbaux que les humains jugent « intelligents », qu'ils ont pu déterminer notre disposition à attribuer des états mentaux comme une *prédisposition* mécanique, nous violons cette impossibilité. En résorbant l'élément d'indétermination qui existe nécessairement du point de vue de l'interrogateur quant à ce qu'il est disposé à juger intelligent, nous figeons l'intelligence (relative à un ensemble d'interrogateurs) dans un ensemble prédéterminé de comportements ; une fois que nous figeons ainsi l'intelligence telle qu'elle doit apparaître à un ensemble d'interrogateurs, il n'est plus difficile de construire un super-ordinateur reproduisant cette intelligence figée. L'hypothèse des Martiens ne sert pas seulement à expliquer la construction d'un super-ordinateur ; elle sert surtout à supposer possible la réduction à une prédisposition déterminée notre disposition indéterminable à attribuer des états mentaux.

1.5.3 Réponse à l'objection de la simulation

Nous avons vu qu'accepter la réponse de Copeland à l'objection de la simulation revenait à supposer que les états mentaux sont définissables en terme de propriétés nécessaires et suffisantes (i.e. est « allographiques »). Nous répondons maintenant à l'objection de la simulation sans faire cette supposition, c.-à-d. en considérant possible que l'authenticité d'un état mental soit déterminable d'après ses conditions réelles d'existence (i.e. est « autographique »). Pour cela, nous devons montrer que, depuis PV_i, il subsiste toujours un doute quant à l'existence d'un « expert en états mentaux » capable de justifier l'attribution qu'il effectue.

De même que la distinction entre PV_i et PV_e nous a permis de répondre à l'objection de la boîte noire, la distinction entre PV_i et PV_m nous sert à répondre à l'objection de la simulation. Si l'interrogateur n'a pas le droit de recourir à une expertise extérieure, il est seul avec ses évaluations de dispositions. Or, du fait qu'il compare deux joueurs en temps réel sans pouvoir accéder à leur identité, l'interrogateur est contraint de faire des hypothèses non seulement sur des comportements mais aussi sur des dispositions : lequel des deux joueurs est disposé à imiter l'autre ? Or, comme nous l'avons vu plus haut (cf. section 1.3.2 page 13), une disposition n'est pas authentifiable de manière absolument certaine. Donc l'interrogateur peut douter de la solubilité de son problème, ce qui justifie indirectement l'hypothèse de la pensée des machines.

Certes, de PV_m ou de PV_e, ce qui n'est accessible à l'interrogateur que comme une disposition pourra être perçu comme une prédisposition, un ensemble de comportements parfaitement réglés. Mais (i) nous n'avons pas le droit de nous placer depuis PV_m et (ii) le fait de nous placer depuis PV_e ne pourrait annuler l'incertitude résiduelle attachée à PV_i que si ce point de vue extérieur connaissait d'avance l'ensemble des comportements qu'un interrogateur moyen est prédisposé à juger intelligents.

Partie 2

Cryptanalyse et Test de Turing

Dans cette partie, nous montrons l'intérêt que peut présenter l'étude des travaux de Turing en cryptologie pour la compréhension du jeu de l'imitation. La cryptologie recouvre deux domaines : celui de la cryptographie (science et technique du chiffrement de l'information) et celui de la cryptanalyse (science et technique du déchiffrement de l'information). Un cryptosystème est un ensemble de règles permettant de produire un texte chiffré à partir d'un texte clair. Nous renvoyons au glossaire pour les autres définitions.

Tout d'abord, nous exposons ce que Putnam nomme le *modèle cryptographique* du rapport entre le langage et les représentations mentales. Nous montrons comment le mariage entre ce modèle et la conception de la communication comme codage mène aux impasses philosophiques que nous avons discutées avec Ryle, mais justifie la tentative d'une détermination statistique de la signification sous-jacente à un texte et des états mentaux sous-jacents à un comportement. Ensuite, nous donnons une esquisse des travaux de Turing en cryptanalyse. Nous expliquons l'analogie qu'il propose entre les méthodes de la cryptanalyse et celles de la confirmation d'hypothèses en physique, puis nous étendons cette analogie au problème de la confirmation d'hypothèses faites sur des états mentaux à partir du comportement. Cette extension de l'analogie proposée par Turing nous semble pouvoir éclairer l'article de 1950 sur au moins deux points : le rôle que joue la notion de « surprise » dans la réponse de Turing à l'objection de Lady Lovelace (Anderson, 1964, p. 57) et le rapport entre machine et hasard. Nous insistons sur le parallèle entre les différents rôles que joue le hasard dans le jeu de l'imitation et en cryptanalyse. Enfin, nous nous appuyons sur ces analyses pour interpréter le jeu de l'imitation comme un problème de projection d'hypothèses, ce qui nous permettra de défendre l'idée d'une fiabilité *évolutive* du test de Turing.

2.1 Cryptographie, signification, communication

2.1.1 Le « modèle cryptographique » de la signification

Dans (Putnam, 1988, p. 49–81), Putnam présente et réfute ce qu'il appelle le « modèle cryptographique » de la signification, modèle dont il fait remonter l'exposition la plus claire à Aristote. D'après celui-ci, chaque mot du langage renvoie à un concept, lequel est une représentation mentale de ce que désigne le mot. Cette conception est *cryptographique* pour autant que la version mentale de la représention correspond au message clair, tandis que sa version publique (linguistique) correspond au message chiffré. Voici les trois présupposés qui composent ce schème[17] :

1. Tout mot qu'il emploie est associé dans l'esprit du locuteur à [au moins] une certaine représentation mentale ;
2. Deux mots ne sont synonymes (n'ont la même signification) que s'ils sont associés à la *même* représentation mentale par les locuteurs qui emploient ces mots ;
3. La représentation mentale détermine, à tout le moins, ce que le mot désigne.

En supposant que l'expression linguistique est définissable comme une application $\mathcal{E}$ allant d'un ensemble de représentations à un ensemble de mots, alors :

- $\mathcal{E}$ est injective : à chaque mot correspond au plus une représentation mentale (sinon la représentation mentale ne *déterminerait* pas ce que le mot désigne et nous n'aurions pas 3.) ;
- La réciproque de $\mathcal{E}$ est surjective : à chaque représentation mentale correspond au moins un mot, possiblement plusieurs (cf. 1. et 2.) ;
- $\mathcal{E}$ n'est donc pas bijective : deux mots peuvent être synonymes (cf. 2).

Ces trois conditions sont celles exigées pour qu'un cryptosystème soit utilisable. A chaque texte chiffré ne doit correspondre qu'un seul texte clair. A chaque texte clair peuvent correspondre plusieurs textes chiffrés (ceci renforçant même l'efficacité du cryptosystème). Enfin, un cryptosystème ne consiste pas nécessairement en l'application d'une fonction bijective. Le nom que donne Putnam à la théorie de la signification qu'il discute est donc parfaitement justifié.

2.1.2 La communication comme codage

Ce modèle « cryptographique » de la signification épouse habituellement le modèle de la communication comme codage. D'après ce modèle communicationnel, le locuteur encode le contenu de ses représentations mentales dans une forme linguistique, transmet ce message chiffré au destinataire, lequel peut – s'il possède le code adéquat – décoder le message reçu pour accéder à son contenu, i.e. à la représentation mentale du locuteur.

Les problèmes posés par la conjugaison de ces deux modèles sont innombrables (cf. (Sperber et Wilson, 1989, p. 11–49) pour une discussion complète.) De manière générale, ils viennent de ce que les modèles cryptographiques de la signification et de la communication supposent d'un côté le tracé d'une frontière entre le privé et le public, entre messages clairs et messages chiffrés, mais suppriment de l'autre les moyens de la tracer. En même temps que le langage est conçu comme le véhicule d'une signification privé, le statut secret de cette signification rend problématique la manière dont nous y accédons. En effet, si la forme linguistique n'est que le véhicule de nos représentations mentales, on peut de même supposer que la forme picturale n'est que le véhicule des représentations imagées qui traversent mon esprit. Et si les formes linguistiques et picturales ne sont que les formes chiffrées d'un message clair sous-jacent, la question de l'accès à ce « message » sous-jacent devient très problématique, pour autrui *comme pour moi-même*. Je n'ai accès à la signification de mes pensées qu'en les exprimant verbalement, même si cette expression reste muette ; et je n'ai accès à ce que représente une image qu'en la regardant, même si je la « regarde » à travers ma mémoire. Toutes ces difficultés sont parentes de celles décelées par Ryle.

Ce qui a longtemps donné l'avantage à cette explication, c'est la justification qu'elle donne du rapport *arbitraire* entre les mots et les représentations, entre la forme de la communication et le contenu effectivement communiqué. Mais l'aspect de ce modèle qui nous intéresse plus particulièrement ici est le suivant : si nous disposons d'une théorie de la communication nous permettant de mesurer le rapport entre la quantité d'information contenue dans un message chiffré et la quantité d'information contenue dans le message clair correspondant, alors nous pouvons espérer une détermination quantitative de la signification d'un message à partir de sa forme linguistique, et plus généralement d'un état mental à partir d'un comportement. Cet espoir est celui incarné par les ambitions de la première cybernétique, et il est clairement exprimé par W. Weaver dans sa note introductive à la théorie de la communication de C. Shannon (Shannon et Weaver, 1963). Il y distingue trois niveaux distincts au sujet de la communication :

Niveau A Avec quelle précision les symboles de la communication sont-ils transmis ?

Niveau B Avec quelle précision les symboles transmis véhiculent-ils la signification désirée ?

Niveau C La signification reçue affecte-t-elle le comportement de la manière désirée ?

Ces niveaux forment un ordre : la solution des problèmes liés à l'effectivité de la communication (niveau C) n'est pas pertinente pour résoudre le problème sémantique (niveau B), et la solution de ce dernier problème n'a aucune pertinence pour la détermination de la quantité d'information transmise (niveau A). Weaver espère cependant que la méthode statistique utilisée pour résoudre les problèmes du premier niveau permettra d'aborder de manière quantitative les problèmes des deux autres niveaux. Pour répondre au problème sémantique, Weaver envisage

par exemple d'ajouter une boîte « destinataire sémantique » entre l'instance technique de décodage du message et l'instance finale de reception de la signification. D'après cette hypothèse, le décodage sémantique reposerait sur les mêmes principes statistiques que le décodage littéral.

Le décodage littéral d'un message chiffré dépend de notre capacité à rapporter les probabilités d'occurrences de chaque lettre du texte chiffré aux probabilités d'occurrence des lettres du texte clair (c.-à-d. aux probabilités liées au comportement de l'émetteur). De même, le décodage de la signification d'un message dépendrait de notre capacité à rapporter les différentes probabilités d'occurrence d'une série de mots aux probabilités d'occurrences des représentations mentales qui leur sont associées. Et rien n'interdit en principe de vouloir étendre ce genre d'explication au domaine pragmatique : si nous connaissons la probabilité pour qu'un état mental soit associé à un comportement donné, alors l'interprétation de ce comportement est accessible à un calcul statistique.

2.1.3 Le poids des probabilités

La conviction que les machines peuvent penser est certainement née de la rencontre entre (i) cette approche statistique de la signification et de la détermination des états mentaux et (ii) l'idée qu'il était possible de doter les machines d'un système d'évaluation des probabilités attachées à une hypothèse. Turing et la cryptographie sont au cœur de cette découverte, comme nous l'exposons maintenant.

Ainsi, alors que Peirce[18] insistait sur l'inaptitude des machines à évaluer des probabilités pour montrer les limites de l'analogie entre le raisonnement humain et raisonnement mécanique, c'est justement sur cette aptitude que Mackay insistera pour défendre l'idée d'un comportement mécanique « intelligent ». Une large partie de (Mackay, 1951) est en effet consacrée à montrer qu'une détermination probabiliste des seuils au-delà desquels changeront certains comportements routiniers permettra de doter la machine d'un comportement qui aura toute l'imprédictibilité du comportement humain : « 'Thinking' in such a mechanism becomes a stochastic process, proceeding along paths determined only statistically. »

Dans un article seulement publié en 1968 (Meltzer et Michie, 1969, p. 3–23) et qui s'inscrit directement dans la conception cybernétique que nous venons de mentionner, Turing pose les bases de sa défense des machines pensantes. Il est intéressant de comparer cet article avec celui qui sera effectivement publié en 1950. D'une part Turing y mentionne la cryptographie comme champ dans lequel il sera possible d'exercer la machine, soulignant le parallèle entre les procédés probabilistes employés en cryptanalyse et ceux employés pour la confirmation d'hypothèses en physique ; d'autre part, il s'y étend plus longuement sur l'intégration d'un élément de hasard dans les machines – intégration dont Mackay souligne aussi la nécessité. Dans la section suivante, nous expliquons et développons l'analogie qu'il suggère entre l'usage

des probabilités en cryptographie et en physique. La section 2.3.3 page 33 étudiera les différents rôles assignés au hasard dans la conception des machines.

2.2 L'analogie cryptographique

Nous faisons d'abord un rapide parcours des travaux de Turing en cryptographie, puis nous analysons la comparaison qu'il suggère entre système cryptographique et lois de l'univers. Nous montrons comment cette comparaison peut s'étendre à son approche des lois du comportement. Si cette extension de l'analogie est valable, elle justifie une approche probabiliste de l'authentification des états mentaux.

2.2.1 Turing et la mécanisation de la cryptanalyse

Turing est contemporain d'une période où la cryptologie subit une double transformation : elle devient une théorie mathématique, et ses procédés se mécanisent. La formalisation mathématique de la cryptologie a surtout été le fait de L. Hill et de C. Shannon. Sa mécanisation a essentiellement consisté en la mécanisation de la cryptographie avec la machine *Enigma*, puis en celle de la cryptanalyse avec les *bombes*. Ces bombes ont originairement été construites par les Polonais avant d'être reproduites et perfectionnées par les Anglais à Bletchley Park.

Si Shannon s'est intéressé à la formalisation mathématique de la cryptographie, Turing s'est en revanche intéressé à la mécanisation de la cryptanalyse. Mais ces deux approches convergaient toutes deux vers une définition du rapport mathématique entre le contenu en information d'un message chiffré et le contenu en information du message clair correspondant. Chez Shannon, la définition de ce rapport aboutit à sa théorie de l'information ; chez Turing, elle l'amena à définir les unités de *ban* et de *déciban*[19]. Shannon avait conscience des implications possibles de sa théorie de l'information dans le domaine du calcul par ordinateur ; mais Turing conçut sa théorie des décibans *précisément* en vue de résoudre un problème mécanique, celui du décodage d'*Enigma*.

Le problème quotidien que devait affronter Turing à Bletchley Park peut se formuler ainsi : étant donné le fonctionnement de la machine *Enigma* et étant donné tel message chiffré, comment trouver le message clair correspondant au message chiffré ? Trouver la solution de ce problème revenait à trouver la clef de chiffrement pour ce message, c.-à-d. retrouver la position des rotors à l'intérieur de la machine *Enigma* au moment où celle-ci chiffrait le message clair. La solution qu'élabora Turing s'appuyait sur le calcul de probabilités, comme n'importe quelle cryptanalyse classique ; mais les spécificités d'*Enigma* et la mécanisation de sa cryptanalyse l'amenèrent à manipuler les probabilités d'une nouvelle manière.

De nombreuses méthodes reposaient déjà sur le fait que la distribution statistique des lettres

de l'alphabet dans la langue écrite ordinaire est relativement fixe. La connaissance de cette distribution permet, une fois rapprochée de la distribution statistique des lettres d'un message chiffré, de mettre en rapport certaines lettres du message chiffré avec certaines lettres du message clair. La probabilité qu'une lettre du message clair corresponde à une lettre du message chiffré dépend de la corrélation entre le nombre respectif d'apparition de ces lettres dans les messages clairs et chiffrés. L'usage de ces probabilités est au fondement des méthodes de cryptanalyse des chiffres monoalphabétiques comme des chiffres polyalphabétiques (voir glossaire.)

Mais les probabilités intervenaient aussi *avant* l'analyse statistique des messages : on peut faire des conjectures sur le contenu du message clair non seulement à partir de la distribution des lettres du message chiffré, mais aussi à partir de ce que l'on sait des conditions d'émission du message. Selon ces conditions, les messages seront plus ou moins stéréotypés, et cette stéréotypie permettra en retour de faire des conjectures probables sur le contenu du message. Par exemple, en temps de guerre, les Alliés savaient que les Allemands envoyaient un message météo à 6h00 du matin tous les jours. L'interception d'un message à cette heure rendait plausible l'apparition de certains mots comme « Wetter » à l'intérieur du message. La présence de ces mots, baptisés « *cribs* » par Turing, était une hypothèse à confirmer.

Dans le cas de la machine *Enigma*, une analyse par la méthode statistique classique ne pouvait pas fonctionner utilement, en raison du trop grand nombre de clefs possibles. Aussi Turing entreprit-il d'accélérer la procédure en utilisant la méthode des mots probables. Pour que cette méthode soit mécanisable, il lui fallait d'abord un formalisme dans lequel exprimer la probabilité qu'un mot soit présent dans un message étant données certaines hypothèses faites sur le cryptosystème. Les bombes implémentaient un système d'évaluation probabiliste de ce genre d'hypothèses, d'où le rapprochement que fait Turing entre les problèmes du cryptanalyste et ceux de l'expérimentateur en physique.

2.2.2 Cryptosystème, lois de l'univers et lois du comportement

Dans (Meltzer et Michie, 1969), Turing s'interroge sur les moyens de rendre une machine plus intelligente. Il cite dans l'ordre différents jeux, l'apprentissage des langues, la traduction des langues, la cryptographie, les mathématiques[20]. Les recherches du siècle dernier en intelligence artificielle nous ont habitués à voir les machines explorer ces différents domaines, sauf en ce qui concerne celui de la cryptographie[21]. Pourtant, c'est celui dont Turing affirme qu'il sera peut-être le plus gratifiant.

> « Le champ de la cryptographie sera peut-être le plus gratifiant. Il y a un parallèle remarquable entre les problèmes du physicien et ceux du cryptographe. Le système sur lequel un message est chiffré correspond aux lois de l'univers, les messages interceptés correspondent aux preuves empiriques disponibles, les clefs pour un jour ou pour un message

aux constantes importantes à déterminer[22]. »

De même que l'expérimentateur formule des hypothèses sur les lois de l'univers, le cryptanalyste formule des hypothèses sur le fonctionnement du système cryptographique. De même que chaque message chiffré permet au cryptanalyste de mieux évaluer la vraisemblance d'une hypothèse faite sur le système cryptographique, l'expérimentateur utilise les résultats de son expérimentation pour tenter de corroborer les hypothèses qu'il fait sur les lois de l'univers. De même que les lois de l'univers expriment la nécessité du lien entre l'état initial et l'état observé d'un système physique, les lois d'un cryptosystème expriment la nécessité du rapport entre message clair et message chiffré. Dans les deux domaines que l'analogie met en rapport, nous avons donc affaire à l'évaluation de la relation entre probabilités *a priori* et *a posteriori* d'une hypothèse.

Le parallèle entre clef d'un cryptosystème et constante universelle semble étrange : les constantes sont parfaitement déterminées, alors que les clefs d'un cryptosystème exigent d'être choisies aléatoirement. Mais cette différence n'est pas pertinente pour notre analogie. Ce qui importe ici, c'est l'aspect *arbitraire* des constantes universelles et des clefs, l'indépendance de leur valeur par rapport à l'expression formelle des lois de l'univers ou du cryptosystème. Les deux systèmes de lois pourraient schématiquement être décrits comme des fonctions prenant deux arguments : une loi de l'univers prendrait en argument la description d'un état initial et la valeur des constantes universelles ; un cryptosystème prendrait en argument un message clair et la valeur de la clef.

L'évaluation par la machine de la crédibilité d'une hypothèse faite sur un système cryptographique est donc comparable à l'évaluation par le physicien de la crédibilité d'une hypothèse scientifique. Cette analogie peut s'étendre au cas de l'analyse d'un comportement : un observateur n'ayant accès qu'à une partie du comportement extérieur d'une entité évalue la crédibilité d'hypothèses faites sur ses états mentaux en fonction des données comportementales qu'il recueille. D'après l'extension de cette analogie, un système comportemental pourrait aussi être schématiquement décrit comme une fonction à deux arguments : le premier serait la description de l'état mental lié à une entité, le deuxième serait l'analogue de la clef (nous laissons provisoirement sa détermination en suspens)[23]. La figure 2.1 page 29 résume l'analogie et son extension.

	Cryptologie	**Physique**	**Comportement**
Entrée	Message clair	Etat initial du système	Etat mental
Clef	Clef aléatoire	Constantes universelles	[?]
Système	Cryptosystème	Lois de l'univers	Lois du comportement
Sortie	Message intercepté (chiffré)	Etat final du système	Comportement observé

FIG. 2.1 *Analogie entre lois d'un cryptosystème, de la physique, d'un comportement*

2.2.3 Confidentialité parfaite, imitation et authentification

L'avancée majeure de Shannon en cryptologie a été de définir rigoureusement la perfection d'un système cryptographique. Soit un cryptosystème défini comme la donnée de l'ensemble $\mathcal{M}$ des messages clairs possibles, l'ensemble $\mathcal{C}$ des messages chiffrés possibles, et l'ensemble $\mathcal{K}$ des clefs possibles, chaque clef définissant une transformation $M \rightarrow C$. Le cryptanalyste connaît $\mathcal{M}$, $\mathcal{C}$ et $\mathcal{K}$, mais non M et C. L'originalité de l'approche de Shannon consiste à traiter M, C et K comme s'il s'agissait de variables aléatoires. Soit $H(M|C)$ la mesure de l'incertitude[24] qu'il reste sur le message clair M lorsqu'on connaît le message chiffré C. Le système cryptographique $\langle \mathcal{M}, \mathcal{C}, \mathcal{K} \rangle$ est parfait si et seulement si $H(M|C) = H(M)$, autrement dit si nous en savons toujours aussi peu sur le message clair après réception du message chiffré correspondant.

Cette définition donne les circonstances dans lesquelles nous sommes assurés de la confidentialité parfaite d'un message. En poursuivant notre analogie, elle nous permet de déterminer les conditions pour qu'un état mental (ou l'identité d'une entité déterminée par la présence de cet état mental) soit parfaitement confidentiel[25] : il suffit pour cela que l'ensemble des informations dont nous disposons sur cet état mental soit exactement le même après observation du comportement de l'entité. Dans le cadre du jeu de l'imitation, cela signifie qu'un joueur *imitera* d'autant mieux l'autre que l'observation de leurs deux comportements n'apportera aucune information permettant de confirmer ou d'infirmer les hypothèses faites sur leurs états mentaux respectifs (et donc, dans le cadre du jeu de l'imitation, sur leur identité.)

2.3 La surprise et le hasard

2.3.1 Définition probabiliste de la surprise

Comment définir le fait qu'un événement soit surprenant ? La première idée est de dire qu'un événement est surprenant s'il est improbable. Mais l'improbabilité d'un événement n'est qu'une condition nécessaire, non suffisante. Comment distinguer parmi les événements improbables ceux qui sont surprenants de ceux qui ne le sont pas? Remarquons que la probabilité que nous assignons à un événement dépend des circonstances dans lesquelles il survient. Quand je jette une pièce en l'air cent fois, j'assigne une faible probabilité au fait qu'elle puisse retomber cent fois du côté pile parce que je crois que la pièce est normale. Si la pièce tombait effectivement cent fois du côté pile, la surprise ne viendrait pas directement de la faible probabilité de cet événement, mais de la faible probabilité d'un ensemble de circonstances exceptionnelles au regard desquelles l'événement est très probable.

Soit C un ensemble de circonstances tenues pour très probables et E un énoncé possiblement surprenant. Pour que E soit surprenant, il est nécessaire que E soit peu probable, soit $P(E) \approx 0$.

Cependant, si E est vrai, la surprise liée à cet événement ne tient pas à sa seule improbabilité, mais au fait que cet événement ait surgi au milieu des circonstances C, circonstances au regard desquelles il était jugé très peu probable. Pour que E soit surprenant, il faut donc que $P(C|E) \ll P(C)$, c.-à-d. que la probabilité d'avoir les circonstances C sachant que E est vrai soit très inférieure à la seule probabilité d'avoir les circonstances C.

Comment nous retrouver dans cette situation? Nous avons dit que la probabilité de C était très grande (soit très proche de 1). Considérons donc la probabilité d'un autre ensemble de circonstances que nous appelerons K. Par définition, nous tiendrons $P(K)$ pour très petit : cela pourra être, par exemple, la probabilité que ma pièce soit dans un métal tel que le côté face de la pièce est aimanté par le sol. Nous n'avons que trois ensembles de circonstances à prendre en considération pour couvrir l'ensemble des événements possibles : ou bien la pièce est ordinaire (C), ou bien son côté face est aimanté par le sol (K), ou bien elle est truquée d'une autre manière ($\neg C, \neg K$). Par le théorème de Bayes, nous avons

$$P(C|E) = \frac{P(E|C)P(C)}{P(E)} \tag{2.1}$$

La probabilité que la pièce tombe cent fois du côté pile correspond à la probabilité de cet événement relativement à chacun des trois ensembles de circonstances cités. Soit :

$$P(E) = P(C)P(E|C) + P(K)P(E|K) + P(\neg C, \neg K)P(E|\neg C, \neg K) \tag{2.2}$$

En vertu de (2.1), $P(C|E) \ll P(C)$ n'est vrai que si le facteur $\frac{P(E|C)}{P(E)}$ est très petit. En vertu de (2.2), $P(E)$ et $P(\neg C,\neg K)$ étant tenus pour fixes, plus $P(E|C)$ est petit, plus $P(E|K)$ sera grand. Autrement dit, la probabilité de E relativement aux circonstances C est d'autant plus petite que la probabilité de E relativement aux circonstances K est grande, soit $P(K)P(E|K) \gg P(C)P(E|C)$. Le fait que E soit surprenant peut alors s'exprimer en disant que (i) E est très improbable relativement à un ensemble très probable de circonstances C *et* (ii) E est très probable relativement à un ensemble très peu probable de circonstances K .

2.3.2 La réponse à l'argument de Lady Lovelace

Cette définition probabiliste de la surprise peut éclairer la réponse que donne Turing à l'objection de Lady Lovelace (Anderson, 1964, p. 57) Cette objection consiste à dire que la machine ne sait faire que ce qu'on lui ordonne, qu'elle est incapable d'initiative et d'invention.

Turing répond en deux temps. D'abord il montre que cette objection ne saurait valoir dans l'absolu, sauf à supposer résolue la question du déterminisme de nos propres actions. Tant que la possibilité d'une illusion sur notre liberté existe, cette liberté hypothétique ne pourra être alléguée dans aucun argument contre la liberté des machines. Et surtout, tant que la relativité

de notre point de vue nous empêchera de trancher la question du fondement de notre liberté, nous n'aurons pas le droit de faire comme si nous pouvions nous échapper de cette relativité pour juger, dans l'absolu, du déterminisme parfait de la machine. Mais le but de Turing n'est pas ici de s'engager dans une discussion métaphysique[26]. Il veut seulement dégager de l'objection de Lady Lovelace les éléments susceptibles de recevoir une réponse. Dans un deuxième temps, Turing considère donc une variante de l'objection tenant compte de la relativité de notre point de vue. Cette variante énonce que les machines sont incapables de nous « prendre par surprise. »

Cette nouvelle formulation se place sur le terrain psychologique. Aussi Turing y répond-il en rapportant sa propre expérience des machines : les machines le « prennent fréquemment par surprise. » Comment cette référence à une expérience personnelle peut-elle servir d'argument ? L'explication que donne Turing suit de près la définition probabiliste de la surprise: la surprise qu'il éprouve devant le résultat que lui fournit la machine ne vient pas directement de ce que ce résultat était improbable, mais de ce qu'en tentant d'anticiper ce résultat, Turing a « oublié » le caractère seulement probable des circonstances qu'il supposait. En oubliant les hypothèses sur lesquelles repose son anticipation du résultat, Turing s'interdit de considérer un autre ensemble de circonstances au regard duquel le résultat effectivement produit par la machine est le plus probable[27].

Pour affirmer définitivement qu'une machine ne peut pas nous surprendre, il ne suffirait donc pas d'affirmer qu'elle est parfaitement déterminée, il faudrait encore montrer que (i) nous connaissons parfaitement ses déterminations ; (ii) au cas où nous ne connaîtrions certaines de ses déterminations que de manière probable, alors nous connaîtrions néanmoins parfaitement la probabilité que nous attachons à nos suppositions ; (iii) au cas où nous ne connaîtrions pas parfaitement la probabilité attachée à nos suppositions, alors nous connaîtrions néanmoins parfaitement la probabilité que nous avons de ne pas parfaitement connaître la probabilité que nous attachons aux déterminations de la machine, etc.

Le fait que la machine ne puisse pas nous surprendre ne tient pas seulement à la limite de nos capacités de calcul, mais elle tient à l'irréductible possibilité théorique de la surprise. L'irréductibilité *pratique* de cette possibilité est affaire de puissance calculatoire, mais l'irréductibilité *théorique* est affaire de point de vue : en vertu de la singularité du nôtre, nous ne pouvons pas évaluer la probabilité de *toutes* les circonstances possibles[28]. Cette impossibilité laisse toujours une place, aussi infime soit-elle, pour un ensemble de circonstances dont nous ne pouvons pas évaluer la probabilité. Cette infime place justifie l'affirmation de Turing remarquant que le fait d'être surpris « requiert de toutes façons un "acte de création mentale". »[29] : la marge d'indétermination de nos évaluations des probabilités associées à chaque circonstance correspond à la marge de « liberté » que nous prenons vis-à-vis des événements, liberté qui peut passer, de notre point de vue, pour une « création ».

Cette explication de l'acte de création mentale impliquée dans la surprise comporte néanmoins un risque : si toute surprise implique un acte de création mentale et s'il n'y a pas de pensée sans cet acte, ne pourra-t-on pas refuser d'accorder aux machines la capacité de pensée non en raison du fait qu'elles ne sont pas surprenantes mais de ce qu'elles ne peuvent pas elles-mêmes *être* surprises ? Quel équivalent pourrait avoir la marge d'erreur que nous appelons chez nous « acte de création mentale » si la machine assigne aux circonstances d'un événement des probabilités qu'elle n'oublie jamais ? Pour répondre, nous pourrions d'abord rappeler que l'argument montrant l'irréductibilité théorique de la surprise vaut aussi pour la machine : même si elle n'oublie aucune des probabilités qu'elle assigne à un ensemble donné de circonstances, elle ne peut pas déterminer absolument tous les ensembles de circonstances pertinents pour l'événement qu'elle cherche à circonscrire. La deuxième réponse consiste à intégrer un élément aléatoire dans la machine.

2.3.3 Le rôle du hasard

L'idée que les machines intégrent un élément de hasard est mentionnée deux fois par Turing dans son article de 1950 (Anderson, 1964, p. 45 et 66), et plus longuement dans l'article *Intelligent Machinery*, (Meltzer et Michie, 1969, 3–23). Ces trois mentions interviennent dans trois problématiques différentes : la première rapproche le hasard de l'hypothèse d'un « libre arbitre »[30] de la machine, la seconde invoque le hasard comme méthode heuristique pour l'apprentissage, la troisième évoque le hasard dans le cadre de la définition des « machines inorganisées ».

Les deux dernières mentions sont liées. Comme le montre (Teuscher et Sanchez, 2000), le concept de machine inorganisée anticipait de loin l'approche connexionniste. S'inspirant certainement des recherches de McCulloch et Pitts, Turing imaginait un arrangement aléatoire d'unités du même type, toutes synchronisées entre elles. Teuscher montre que la conception proposée par Turing peut être définie comme un réseau aléatoire booléen. En envisageant des interférences en temps réel avec l'évolution aléatoire des calculs effectués par ces unités, on peut doter ce réseau d'une capacité d'apprentissage. Obtenir cette capacité pourrait être très long, mais au moyen d'une « *mimicking education* » (Meltzer et Michie, 1969, p. 15) appropriée, la machine pourrait acquérir le comportement qu'on attend d'elle.

Ces spéculations sont à rapprocher de l'idée d'une heuristique aléatoire, présentée rapidement en (Anderson, 1964, p. 67). Si le problème de la machine est d'apprendre à adopter un comportement qui satisfera le « maître », alors la machine aura avantage à tester les solutions de manière aléatoire, plutôt que de manière systématique. Une telle heuristique se justifie techniquement par l'économie qu'elle permet de faire au niveau des ressources de mémoire employées, et pratiquement par le fait que l'ensemble des comportements satisfaisant le maître est

(probablement) indéterminé. Cette indétermination de ce que le maître jugera « intelligent » permet d'utiliser efficacement l'indétermination de la machine inorganisée.

Le concept de hasard impliqué dans la conception des machines inorganisées est donc proche de celui d'indétermination. Comme le remarque Turing, cette indétermination n'a pas besoin d'être absolue : savoir si une machine est organisée ou non peut dépendre du point de vue que l'on adopte sur elle (cf. (Meltzer et Michie, 1969, p. 9) et (Teuscher et Sanchez, 2000, p. 3).) C'est seulement dans la discussion entre hasard et « libre arbitre » que l'exigence d'un hasard *absolu* devient nécessaire. Nous pensons (i) qu'il faut mettre cette exigence en rapport avec l'objection selon laquelle une machine ne pourrait jamais être « surprise » et (ii) que le rôle joué par ce hasard absolu relativement au comportement de la machine est comparable au rôle joué par les clefs aléatoires relativement au comportement d'un cryptosystème.

Rappelons d'abord que ce n'est qu'avec la théorie de la calculabilité et des fonctions récursives qu'une définition rigoureuse du hasard est devenue possible[31], bien qu'il n'y ait pas, en 1950, d'accord général sur la définition exacte de l'aléatoire (cf. (Delahaye, 1999, p. 29–64) pour une discussion des différentes définitions). Mais ce qui intéresse Turing dans la discussion du rapport entre hasard et libre arbitre, ce n'est pas tant la définition formelle de la notion d'aléatoire que la difficulté, pour un observateur quelconque, de faire la différence entre hasard et pseudo-hasard. L'exemple qu'il prend est particulièrement significatif[32] : il est impossible de faire la différence entre un comportement parfaitement aléatoire et un comportement dépendant d'un nombre même fini de décimales de π. Faire dépendre les calculs effectués par une machine des décimales de π, c'est augmenter l'imprédictibilité de ce calcul à un tel point qu'elle coïncide presque parfaitement avec une imprédicitibilité absolue. Comme nous l'avons justifié plus haut, cette imprédictibilité parfaite produite par l'intégration d'un mécanisme aléatoire serait nécessaire pour réserver à la machine la possibilité de surprendre et d'être surprise.

Nous avons donc un jeu entre l'imprédictibilité de la machine du point de vue d'un observateur et l'imprédictibilité propre à la machine. Cette dernière peut être soit de l'ordre du pseudo-hasard (comme π), soit de l'ordre du hasard absolu (en intégrant, par exemple, un amplificateur quantique). Nous voudrions maintenant préciser un point de l'analogie entre lois du comportement et lois d'un cryptosystème : dans les deux cas, le rapport entre ces deux types d'imprédictibilité est le même.

Le but d'un cryptosystème est de produire en sortie des messages qui paraissent parfaitement aléatoires alors qu'ils sont déterminés (sans quoi le déchiffrement ne pourrait pas avoir lieu.) Sur quoi repose l'aspect aléatoire du message chiffré ? Il repose sur les lois de manipulation symbolique définissant le cryptosystème en question. Si nous sommes ignorants de ces lois, le message ne peut que nous paraître aléatoire. Mais si nous connaissons parfaitement l'ensemble

des lois de manipulation symbolique du cryptosystème, le message clair nous est-il pour autant accessible ? Dans le cas d'*Enigma*, non, car l'ensemble des transformations subies par un message clair repose ici sur la combinaison entre les lois du cryptosystème et la valeur de la clef. Turing était dans une position où il connaissait les lois de chiffrement d'*Enigma*, mais où toute la difficulté était de déterminer, d'après la connaissance de ces lois, quelle était la clef du jour. Aussi, la pseudo-aléatoirité du message dépendait, en dernière instance, de l'aléatoirité réelle de la clef : plus cette clef était choisie de manière aléatoire, plus le message avait de chances de brouiller les probabilités attachées aux lettres du messages clair ; plus cette clef était prévisible, moins le système était sûr.

Si nous suivons l'analogie que nous avons exposée, nous observons qu'un générateur de hasard au sein de la machine jouerait le même rôle qu'une clef au sein d'un cryptosystème. Sans générateur de hasard, l'imprédictibilité du comportement de la machine ne vient que de notre ignorance de la totalité des lois de son comportement. Avec un générateur de hasard, cette imprédictibilité devient théoriquement irréductible. Plus le générateur de hasard est fiable, moins cette anticipation du comportement de la machine sera certaine. Ce n'est pas que cette anticipation ne pourra plus s'appuyer sur ce que nous savons du comportement déterminé de la machine, mais c'est qu'elle ne saura pas dans quelle mesure ce comportement déterminé est guidé par des processus aléatoires sous-jacents.

2.4 Le jeu de l'imitation comme problème de projection

Nous avons vu les liens entre la cryptanalyse et la dimension probabiliste du jeu de l'imitation. Cette dimension permet de concevoir l'efficacité de l'imitation et la possibilité de l'authentification comme des problèmes statistiques. Maintenant, nous exploitons cette dimension pour tenter de reformuler le test de Turing comme un problème de confirmation et de projection d'hypothèses. Nous devrons pour cela exposer la théorie de la projection esquissée par Goodman en (Goodman, 1984, p. 96–127). A partir de cette théorie, nous tenterons de justifier l'idée d'une fiabilité *évolutive* du test de Turing comme test permettant d'affirmer qu'une machine pense.

2.4.1 Théorie de la projection

Par « projection », nous ne désignons pas les faits psychologiques consistant à se mettre à la place de quelqu'un ou à ne voir dans un objet qu'un reflet de soi-même. Nous employons ce terme dans son acception épistémologique, celle élucidée notamment par Goodman. Nous dirons d'un prédicat qu'il est manifeste s'il est directement observable. Un prédicat dispositionnel est la *projection* d'un prédicat manifeste : une barre est dite « flexible » par projection du

prédicat manifeste « fléchit ». La légitimité de cette projection constitue un premier problème. Goodman a montré comment l'énigme de l'induction exigeait que l'on résolve un problème d'ordre supérieur : celui du rapport entre l'ensemble des projections réelles d'une hypothèse et sa projectibilité, soit la question de savoir ce qui fait la projectibilité du prédicat « projeté ».

Avant même qu'une hypothèse soit projetée, les instanciations de celle-ci peuvent être vraies, fausses ou indéterminées. Les instanciations vraies sont les *cas positifs* de l'hypothèse, les instanciations fausses sont ses *cas négatifs*. L'ensemble des cas positifs à un instant donné constitue la classe des *preuves empiriques* de l'hypothèse. L'ensemble des cas indéterminés constitue la classe de *projection*. Une hypothèse est supportée (ou soutenue) dans la mesure où elle a des cas positifs, violée dans la mesure où elle a des cas négatifs. Une hypothèse violée est fausse, mais une hypothèse fausse peut ne pas avoir été violée à un moment donné. Si tous les cas d'une hypothèse ont été exhaustivement examinés, l'hypothèse est dite exhaustivement parcourue.

Une hypothèse est dite *réellement projetée* si, lors de son adoption, certaines de ses instanciations ont été examinées et tenues pour vraies, et si d'autres cas restent à examiner. Plus une hypothèse ou un prédicat ont été réellement projetés, plus ils sont *implantés*. La condition nécessaire pour qu'une hypothèse soit projectible est de n'avoir que des cas positifs et des cas indéterminés. Une hypothèse violée et une hypothèse exhaustivement parcourues ne peuvent donc pas être projetées. Mais cette condition n'est pas suffisante : une hypothèse est projectible pour autant qu'elle n'est pas supplantée par une hypothèse conflictuelle elle-même soutenue, inviolée et mieux implantée[33]. Par « pratique projective », nous désignerons l'ensemble des projections bien implantées pour un individu, c.-à-d. celles qu'il a coutume de faire sans avoir jusque-là subi de démenti et sans craindre de se tromper.

2.4.2 La fiabilité évolutive du test de Turing

Notre pratique projective nous a habitués à supposer qu'un être humain pense et qu'une machine ne pense pas. Le jeu de l'imitation est une situation dans laquelle la force de cette pratique projective sera détournée de son cours habituel : à l'insu de l'interrogateur, la confiance qu'il a dans le fait qu'un homme pense se transformera en confirmation de l'hypothèse selon laquelle la machine est capable de penser, *via* le fait que cette machine est capable de se faire passer pour un être humain. Dans ce qui suit, nous détaillons cet argument en reformulant le jeu de l'imitation comme un problème de projection d'hypothèses. Nous tentons de justifier le fait que les situations semblables à celle du jeu de l'imitation pourront faire évoluer notre pratique projective, et nous montrons comment cette évolution rend de plus en plus cohérent le test de Turing.

Trois problèmes de projection

Le premier problème se situe au niveau de l'interrogateur : celui-ci doit tenter de confirmer l'une ou l'autre des deux hypothèses d'identification, soit X est A et Y est B (hypothèse correcte que nous nommerons h_1), soit X est B et Y est A (hypothèse erronnée h_2). Le deuxième problème se situe au niveau d'un observateur extérieur au test de Turing, capable d'observer les trois joueurs : cet observateur doit évaluer l'hypothèse selon laquelle l'interrogateur est capable ou non d'authentifier correctement l'identité des deux joueurs (H_1 ou H_2). Un troisième problème surgit lorsque nous adoptons une perspective diachronique sur le test de Turing : que devient, d'un test à l'autre, la projectibilité de l'hypothèse selon laquelle ce test est cohérent[34] ?

Ce découpage en trois problèmes distincts nous permet de détailler ce qui est impliqué dans le pronostic de Turing : « dans une cinquantaine d'années il sera possible de programmer des ordinateurs, avec une capacité de mémoire d'à peu près 10^9, pour les faire si bien jouer au jeu de l'imitation qu'un interrogateur moyen n'aura pas plus de 70 pour cent de chances de procéder à l'identification exacte après cinq minutes d'interrogation. » Dans cette phrase, l'expression « procéder à l'identification exacte » renvoie à notre premier problème de projection. Les 70 pour cent renvoient au deuxième problème de projection. Le syntagme circonstanciel « dans une cinquantaine d'années » évoque l'évolution possible des machines et des résultats qu'elles obtiendront au cours des différents tests, renvoyant ainsi à notre troisième problème de projection.

Cette question de la distinction des points de vue à l'intérieur du problème de la confirmation d'une hypothèse n'était pas indifférente à Turing. Dans (Good, 1950, p. 72), Good rapporte un paradoxe que lui a indiqué Turing en 1940. Soit $P(E|H)$ la *vraisemblance* d'une hypothèse H étant donnée l'expérience E ; soit $\frac{P(E|H)}{P(E|\overline{H})}$ le *facteur* en faveur de H en vertu de l'expérience E (le logarithme de ce facteur en base dix définit le *poids d'évidence*. Voir annexe, section 4.3 page 66.). Le paradoxe souligné par Turing est que le facteur attendu d'une hypothèse fausse en vertu de n'importe quelle expérience est de un. Pour A sachant que l'hypothèse H faite par B est fausse, la probabilité à ses yeux pour qu'une expérience confirme H aux yeux de B est de un[35].

Dans ce qui suit, nous soulignons la connexion qui existe entre ces trois problèmes de projection. Nous voudrions montrer que si l'interrogateur se trompe souvent dans la détection d'une différence entre la machine et l'être humain, alors un observateur extérieur au test de Turing se trompera moins souvent en supposant que l'être humain est incapable d'identifier correctement les deux joueurs. Si nous considérons ensuite que la classe des observateurs potentiels est identique à la classe des interrogateurs potentiels, alors cela signifie qu'un observateur jouant le rôle d'un interrogateur sera dans le droit de dire, *en parlant de lui-même*, qu'il est incapable de distinguer l'homme de la machine. Cet aveu suffit à laisser ouverte la possibilité d'une évo-

lution de notre pratique projective, d'une meilleure implantation de l'hypothèse selon laquelle l'homme est incapable de détecter la machine dans le jeu de l'imitation, et donc d'une plus grande cohérence du test de Turing comme test indiquant le fait que les machines « pensent ».

Comme pour notre réponse aux objections de la boîte noire et de la simulation, notre argument repose sur deux faits : d'une part celui de la nécessité de distinguer les points de vue depuis lesquels sont projetées les hypothèses, d'autre part le fait qu'il n'y a pas d'observateur absolument extérieur au test de Turing, pas d'observateur qui ne soit susceptible de jouer un jour le rôle de l'interrogateur. Notre argument pour défendre la fiabilité évolutive du test de Turing a donc la forme suivante : si h_2 n'est jamais projetée, alors le test de Turing ne prouvera jamais que la machine pense ; si h_2 est projetée, alors soit nous sommes sûrs que h_2 est fausse, et notre pratique projective ne pourra pas évoluer, soit nous n'en sommes pas sûr ; or il y a une incertitude quand à la fausseté de h_2, et elle nous est garantie par l'impossibilité de déterminer à l'avance la probabilité de H_2. Reconnaître cette impossibilité, c'est reconnaître la possibilité d'une évolution de notre pratique projective, évolution qui aura pour effet indirect de faire de h_2 une hypothèse mieux implantée.

La connexion entre les trois problèmes de projection

Dans le cas où la tâche d'identification entreprise par l'interrogateur repose sur un prédicat manifeste (par exemple « être blonde »), le fait que l'interrogateur projette réellement l'hypothèse selon laquelle la machine possède ce prédicat ne rend pas cette hypothèse projectible. Cela prouve seulement que l'interrogateur peut se tromper. Puisqu'il existe nécessairement un point de vue à partir duquel l'hypothèse fausse de l'interrogateur est perçue comme fausse (i.e. le point de vue d'un observateur extérieur), la projection réelle de l'hypothèse que fait l'interrogateur à son insu n'entraîne pas sa projectibilité, car il y a moyen de s'assurer que cette hypothèse est violée. Dans cette situation, le deuxième problème de projection consiste à savoir si un observateur peut projeter l'hypothèse selon laquelle la pratique projective d'un interrogateur pourra être mise en défaut par une machine, mais cette mise en défaut ne pourra pas avoir pour effet de faire évoluer cette pratique projective, car l'observateur sera toujours là pour rappeler son erreur à l'interrogateur.

Dans le cas où la tâche d'identification ne repose pas seulement sur des prédicats manifestes mais sur n'importe quel prédicat possible (et notamment sur l'authentification d'une disposition), les données du problèmes sont différentes. Ici aussi, plus h_2 sera réellement projetée (c.-à-d. plus l'interrogateur se trompera), plus H_2 sera plausible (c.-à-d. plus un observateur extérieur aura de chances d'avoir raison en supposant que l'interrogateur est incapable d'identifier correctement les deux joueurs.) Mais le fait que l'interrogateur puisse tester la présence de n'importe quel prédicat possible donne un statut particulier à la plausibilité de H_2. Si celle-

ci pouvait être déterminée à l'avance et une bonne fois pour toutes, nous nous retrouverions dans le cas des Martiens vis-à-vis des hommes, c.-à-d. dans une situation où des observateurs connaissent *a priori* les probabilités qu'un ordinateur donné a de tromper un homme dans le jeu de l'imitation, parce qu'ils connaissent à l'avance ce que cet être humain juge « intelligent ». Dans ce cas, le fait de jouer au jeu de l'imitation n'introduirait aucune différence entre les probablités *a priori* et *a posteriori* de H_2, car les Martiens seraient toujours observateurs et les humains interrogateurs. Mais, à supposer que les interrogateurs et les observateurs soient seulement des hommes, alors la probabilité de H_2 ne peut pas être déterminée une fois pour toute, elle évolue en fonction de la probabilité de h_2.

Le fait que la pratique projective d'un interrogateur puisse le mener à réellement projeter h_2 amène un observateur extérieur à réellement projeter H_2. Le fait que l'interrogateur et l'observateur soient tous les deux des hommes implique que l'observateur ne peut pas de lui-même s'assurer *dans tous les cas* que H_2 puisse être réellement projetée *et* violée. Aussi, l'incertitude sur la possibilité de la violation H_2 se répercute en incertitude sur la violation possible de h_2. Donc, de ce que h_2 sera réellement projetée, il s'ensuivra aussi qu'elle sera, du point de vue d'un ensemble fini d'êtres humains, projectible. Projeter réellement l'hypothèse h_2 ne la rend pas directement projectible (car il existe effectivement des cas où un observateur extérieur dirait tout simplement que l'être humain s'est trompé, ce qui rendrait h_2 violée, et donc improjectible) ; mais h_2 est indirectement projectible du fait que H_2 peut être réellement projetée. La pratique projective des êtres humains pourra alors évoluer de manière à attribuer des états mentaux aux machines avec la même confiance inductive qu'ils en attribuent aujourd'hui aux autres êtres humains, cette évolution rendant le test de Turing de plus en plus cohérent.

Partie 3

Logiques de l'authentification

3.1 Le modèle inférentiel de la communication

Jusqu'ici, nous nous sommes intéressés aux notions d'imitation et d'authentification dans le cadre du modèle communicationnel du code. Nous avons souligné que ce cadre avait pu susciter l'ambition d'une étude quantitative des aspects sémantiques et pragmatiques de la communication, et montré les rapports que l'on peut établir entre ce modèle communicationnel du code et les recherches effectuées par Turing en cryptanalyse.

Le modèle communicationnel du code est largement remis en cause. A sa place a été développé le modèle dit « inférentiel » : dans celui-ci, les propositions ne sont plus la partie linguistiquement chiffrée d'une contrepartie sémantique ou pragmatique claire, mais des *indices* à partir desquels peuvent être inférés les états mentaux d'autrui. Si l'on s'intéresse aux états mentaux comme à des états épistémiques, ce modèle se prête au formalisme de la logique épistémique. Si l'on suppose que les états mentaux sont des *choses* sur lesquelles peuvent agir les participants d'une communication, alors ce modèle communicationnel a d'importantes conséquences dans le domaine de la pragmatique – conséquences d'abord soulignées par Grice puis développées par Sperber et Wilson. Alors que le modèle communicationnel du code permettait d'expliquer le lien *arbitraire* existant entre la forme des propositions et leur signification, Sperber et Wilson prennent appui sur le modèle inférentiel pour défendre l'idée qu'il existe un lien de *ressemblance* entre ce qui est communiqué et le véhicule de la communication (cf. (Andler, 1992, p. 219–238)). Nous mentionnons ce renversement de perspective non pour le critiquer, mais pour montrer que la remise en cause du modèle communicationnel du code appelle une nouvelle approche des notions d'imitation et d'authentification.

Cependant, nous n'aborderons pas ce problème de front, car l'implication des notions de simulation et d'authentification dans l'activité du *mindreading*[36] est une question dont les données sont très encore très confuses (cf. la tentative de clarification récemment proposée par

Alvin Goldman, amorcée dans (Sperber, 2000, p. 171–196, « *The mentalizing folk* »).) Nous entreprendrons plus modestement d'analyser les protocoles cryptographiques d'authentification. Voici pourquoi et à quelle fin.

(i) Ces protocoles manipulent une définition simple de l'*identité* : une entité est reconnue par sa clef privée. Ceci évite que l'on se perde en cherchant à définir la nature de ce qui est imité ou authentifié. (ii) Les logiques utilisées pour l'analyse des protocoles d'authentification sont cousines des logiques épistémiques, ce qui permet d'expliciter en quoi les notions d'authentification et d'imitation impliquées dans ces protocoles sont relatives au modèle *inférentiel* de la communication. (iii) L'analyse des protocoles d'authentification n'opère pas une réduction de nos problèmes définitionnels, mais précise les moyens dont nous disposons pour les aborder. Ainsi, nous serons déjà heureux si l'étude des logiques de l'authentification pouvait nous aider à y voir plus clair.

3.2 Présentation des protocoles d'authentification

Si je communique avec quelqu'un à l'aide d'un code privé, si je sais que ce code n'a jamais été diffusé et si je sais que ce code ne peut pas être brisé, alors je suis assuré de l'identité de mon correspondant. Mais ces conditions sont rares. La plupart du temps, le code est doublement vulnérable. Un intrus peut (i) le calculer à partir des messages codés qu'il a interceptés, (ii) l'intercepter au moment où les agents se le communiquent. Il doit donc exister un moyen pour que deux agents s'échangent une clef privée en toute sécurité, de manière à être assurés de l'identité de leur correspondant.

Un protocole d'authentification est un échange de messages entre agents[37] utilisant les procédés cryptographiques à des fins d'authentification. La cryptographie à clef symétrique (aussi appelée cryptographie à clef *secrète*) se sert de la même clef pour le chiffrement et le déchiffrement. Les exemples classiques sont le standard de chiffrement de données (DES) et son successeur récent, le standard avancé de chiffrement (AES). La cryptographie asymétrique (aussi appelée cryptographie à clef *publique*) utilise des clefs différentes pour le chiffrement et le déchiffrement. L'exemple le plus connu est le RSA. Une clef publique est ainsi nommée car elle est accessible à tous ; elle sert à chiffrer le message. A chaque clef publique correspond une unique clef privée, connue du seul intéressé ; cette clef sert au déchiffrement des messages chiffrés avec la clef publique correspondante. Parce qu'elle est connue d'un seul individu, la clef privée peut aussi servir de signature digitale ; la clef publique sert alors à vérifier la conformité du lien entre la signature et son auteur. Dans la pratique, on utilise la cryptographie à clef publique pour l'établissement d'une *clef de session*, laquelle sert ensuite comme clef classique (i.e. symétrique) pour l'échange de messages. Un serveur d'authentification a autorité sur (et peut garantir) la

conformité du lien entre une clef publique et sa contrepartie privée. Dans ce qui suit nous nous intéresserons exclusivement aux protocoles permettant l'établissement sécurisé d'une clef de session.

Authentifier un agent, c'est s'assurer de son identité. Cette assurance peut prendre différentes formes : on peut s'assurer que ceux qui reçoivent une clef de session sont bien ceux qu'ils prétendent être, ou bien s'assurer que l'un des agents a bien la clef qu'il devrait avoir, ou bien que celui avec qui vous croyez partager une clef privée croit aussi que vous possédez cette clef privée, etc. Un protocole est utilisé à des fins de sécurité parce que la communication a lieu dans un environnement hostile ; on suppose donc l'existence d'un adversaire mal intentionné, classiquement nommé *Eve*. Plus cet adversaire aura de pouvoir, plus le protocole devra être sécurisé . L'adversaire standard, défini par (Dolev et Yao, 1983), est un adversaire très puissant : tous les messages de l'échange passent par lui, il peut les lire, les altérer et les détourner. Toutefois, la partie algorithmique du chiffrement est considérée comme infaillible[38] : l'adversaire ne peut déchiffrer un message que si et seulement s'il possède la bonne clef, et il ne peut chiffrer des messages qu'à partir des clefs qu'il possède.

3.2.1 Exemple de protocole

Dans cette section nous présentons le protocole Needham-Schrœder à clef privée (NSSK). Ce protocole est le plus connu de la littérature non seulement à cause de sa relative simplicité, mais surtout parce qu'il permet d'illustrer le type d'analyse que la logique BAN est capable d'effectuer et le type de faille qu'elle permet de déceler. Les contreparties honnêtes d'Eve (E) sont connues sous les prénoms d'Alice (A) et Bernard (B). Le dernier agent du protocole est le serveur (S).

Alice et Bernard partagent des clefs secrètes avec le serveur ; ces clefs sont appelées clefs à long-terme pour les distinguer des clefs de session et sont notées k_{AS}, k_{BS}, k_{AB}, etc. Le protocole nécessite l'usage de *nonces* : ce terme désigne tout nombre aléatoire inclus dans un message pour en indiquer la « fraîcheur ». Lorsqu'un agent inclut un *nonce* dans un message, il est assuré que tout message contenant ce *nonce* (ou un nombre fonction de ce *nonce*) aura été généré à partir du déchiffrement de son message. Un *nonce* généré par Alice est noté n_A. La clef de session que le serveur génère pour Alice et Bernard est notée k_{AB}. Un message chiffré avec la clef k est noté $\{M\}_k$.

Message 1	$A \rightarrow S$:	A, B, n_A
Message 2	$S \rightarrow A$:	$\{n_A, B, k_{AB}, \{k_{AB}, A\}_{k_{BS}}\}_{k_{AS}}$
Message 3	$A \rightarrow B$:	$\{k_{AB}, A\}_{k_{BS}}$
Message 4	$B \rightarrow A$:	$\{n_B\}_{k_{AB}}$
Message 5	$A \rightarrow B$:	$\{n_B - 1\}_{k_{AB}}$

FIG. 3.1 *Protocole Needham-Schræder à clef privée*

Protocole Needham-Schrœder à clef privée

Dans ce protocole, Alice indique au serveur qu'elle voudrait communiquer avec Bernard, et inclut le *nonce* n_A. Le serveur S lui renvoie un message chiffré avec sa clef à long terme (k_{AS}). Ce message contient le *nonce* n_A (pour qu'elle puisse en vérifier la fraîcheur), l'identifiant de Bernard (pour qu'elle sache qu'il s'agit bien d'une session entre elle et Bernard), la clef de session k_{AB}, et un sous-message chiffré $\{k_{AB}, A\}_{k_{BS}}$ à faire suivre à Bernard. Celui-ci déchiffre ce sous-message et envoie à Alice un *nonce* n_B chiffré avec la clef de session, de manière à montrer qu'il possède la clef de session et afin de s'assurer qu'elle la possède aussi. Alice déchiffre le message, soustrait un du *nonce* ($n_B - 1$), chiffre le résultat et l'envoie à Bernard, ce qui met fin au protocole.

3.3 La logique BAN

Dans cette section, nous présentons succintement la logique BAN[39] : ses concepts, sa notation et ses règles. Nous montrons comment cette logique permet d'analyser le protocole Needham-Schrœder à clef secrète et de mettre à jour une grosse faille de ce protocole.

La logique BAN est une logique de la croyance. Elle permet de dire explicitement quelles sont les propositions que les agents tiennent pour vraies à l'issu d'un protocole. Par exemple, Alice pourra parvenir à croire qu'une clef qu'elle a reçue du protocole est une « bonne » clef de session pour communiquer avec Bernard – ce que signifie « bonne » sera précisé par la suite. Si Bernard atteint ce même état de croyance, alors on pourra affirmer que l'authentification est réussie - au moins selon une certaine formulation des requisits de l'authentification.

3.3.1 La notation BAN

Dans toutes ces expressions, X est soit un message soit une formule. Voici les expressions du langage BAN[40] :

P believes X: P peut agir comme si X est vrai.

P *received* X: P a reçu un message contenant X, et P peut avoir X depuis ce message ; ceci peut nécessiter un déchiffrement.

P *said* X: P a envoyé un message contenant X à un instant passé ; de plus P croit X et comprend qu'il a envoyé X à ce moment précédent.

P *controls* X: P a juridiction sur X, c.-à-d. qu'on peut lui faire confiance au sujet de X.

$fresh(X)$: (Lu « X est frais. ») X n'a été envoyé dans aucun message précédent l'échange actuel.

$P \overset{k}{\longleftrightarrow} Q$: (Lu « k est une bonne clef pour P et Q. ») k ne sera jamais découvert par d'autres agents que P, Q et les agents auxquels P et Q font confiance. Ces agents de confiance sont nécessaires, car le serveur voit souvent - et même génère - k.

$PK(P,k)$: (Lu « k est une clef publique de P. ») La clef secrète k^{-1} correspondant à k ne sera jamais découverte par un autre agent que P ou un agent auquel P fait confiance.

$\{X\}_k$: Abrégé de « $\{X\}_k$ *from* P » (Lu « X chiffré avec k de P. ») C'est la notation pour le chiffrement. Les agents peuvent reconnaître leurs propres messages. Les messages chiffrés sont lisibles et vérifiables seulement par ceux qui possèdent les bonnes clefs.

3.3.2 Les règles BAN

Dans une analyse, le protocole est d'abord *idéalisé* en messages contenant des assertions ; on fait ensuite état des assomptions du protocole, puis les conclusions sont inférées à partir des assertions du protocole idéalisé et des assomptions. Nous donnons maintenant les principales règles d'inférence et nous renvoyons à l'appendice pour l'exposé exhaustif (section 4.4 page 68).

Signification du message

$$\frac{P \text{ believes } P \overset{k}{\longleftrightarrow} Q \qquad P \text{ received } \{X\}_k}{P \text{ believes } Q \text{ said } X}$$

« Si P reçoit X chiffré avec k et si P croit que k est une clef valide pour communiquer avec Q, alors P croit que Q a dit X. »

Cette règle ne nous dit rien des sous-messages que P peut extraire d'un message chiffré. Ce sera le rôle des règles de réception. Cette règle dit plutôt que P peut savoir *qui* a envoyé le message[41].

Vérification de *nonce*

$$\frac{P \text{ believes } fresh(X) \qquad P \text{ believe } Q \text{ said } X}{P \text{ believes } Q \text{ believes } X}$$

Cette règle autorise le passage d'un événement passé à une croyance présente : quelque chose qui a été dit dans le passé donne lieu à une croyance actuelle. Pour pouvoir être appliquée, cette règle requiert que X ne contienne aucun texte chiffré. Elle requiert aussi que tout ce qu'un agent honnête et compétent vient de dire doit être quelque chose qu'il croit[42].

Juridiction

$$\frac{P \textit{ believes } Q \textit{ controls } X \qquad P \textit{ believes } Q \textit{ believes } X}{P \textit{ believes } X}$$

La règle de juridiction est celle qui autorise les inférences à partir desquelles un agent est amené à croire qu'une clef est valide, même s'il s'agit d'une chaîne aléatoire qu'il n'a jamais vue avant. Notons que cette règle ne permet pas d'inférer la validité même de la clef du point de vue d'un agent, mais seulement d'inférer la *croyance* de l'agent en la validité de la clef.

3.3.3 L'analyse BAN des protocoles

L'analyse BAN procède en quatre étapes :

1. Idéalisation du protocole.
2. Explicitation des assomptions faites à l'état initial.
3. Annotation du protocole : cela revient, pour chaque transmission de message « $P \rightarrow Q$: M » dans le protocole, à asserter Q *received* M *from* P.
4. Utiliser la logique pour dériver les croyances de chaque agent.

Dans le cas du protocole NSSK, nous pouvons dériver les croyances suivantes (voir section 4.4 page 69 pour la dérivation détaillée) :

- A *believes* $fresh(A \overset{K_{AB}}{\longleftrightarrow} B)$
- A *believes* $(A \overset{K_{AB}}{\longleftrightarrow} B)$
- A *believes* B *believes* $(A \overset{K_{AB}}{\longleftrightarrow} B)$
- B *believes* $(A \overset{K_{AB}}{\longleftrightarrow} B)$
- B *believes* A *believes* $(A \overset{K_{AB}}{\longleftrightarrow} B)$

Mais la dérivation de ces énoncés a rendu nécessaire l'assomption suivante :

$$B \textit{ believes } fresh(A \overset{K_{AB}}{\longleftrightarrow} B)$$

Or le fait que le protocole oblige à faire cette assomption pour achever son but d'authentification est dangereux. On peut se servir de cette assomption cachée pour monter l'attaque suivante, présentée dans la figure 3.2 page 46.

Message 3	$E_A \rightarrow B : \{k_{AB}, A\}_{k_{BS}}$
Message 4	$B \rightarrow A : \{n'_B\}_{k_{AB}}$
Message 3	$E_A \rightarrow B : \{n'_B - 1\}_{k_{AB}}$

FIG. 3.2 *Attaque du protocole Needham-Schræder à clef privée*

E_A désigne l'attaquant E se faisant passer pour A. L'attaque repose sur le fait que Bernard n'a aucun moyen de s'assurer de la fraîcheur du message 3. Etant donné qu'un attaquant peut dépenser autant de temps qu'il le souhaite pour briser la clef k_{AB}, s'il parvient du moins à le faire avant l'expiration de k_{BS} (qui est une clef à long terme), il peut alors lancer cette attaque. Bernard pensera avoir confirmé le partage de k_{AB} avec Alice, alors qu'Alice n'existe pas. Ici, le rôle de l'analyse BAN n'a pas été de découvrir directement cette attaque, mais de mettre à jour l'assomption implicite qu'elle exploite.

3.3.4 Critiques et extensions de la logique BAN

La création de la logique BAN répondait à un problème précis : formaliser la vérification de la validité des protocoles d'authentification. Aussitôt publiée, cette logique a subi de nombreuses critiques, les unes justifiées, les autres non. Au fur et à mesure que ces critiques se firent entendre, on commença à entrevoir que la difficulté était peut-être d'ordre conceptuel : qu'entend-on par *authentification* ? Dans ce qui suit, nous présentons succinctement les problèmes que les logiques de l'authentification doivent affronter, puis nous nous attardons sur la réflexion qui a émergé autour de la définition de l'authentification.

Les premières critiques de la logique BAN se sont concentrées sur l'étape de l'idéalisation. Pour parer aux assomptions implicites, il a été tenté de donner à la logique BAN un langage permettant d'exprimer des requisits plus nombreux et plus précis sur l'état épistémique de chaque instance. En 1990, Nessett proposa le protocole suivant afin de montrer les faibles de l'analyse BAN.

Message 1	$A \rightarrow B : \quad \{n_A, k_{AB}\}_{k_A^{-1}}$
Message 2	$B \rightarrow A : \quad \{n_B\}_{k_{AB}}$

FIG. 3.3 *Le Protocole Nessett (1990)*

Dans le premier message, Alice chiffre une clef de session en utilisant sa clef privée (dont la contrepartie publique est supposée accessible à tous). Bernard lui envoie en retour la valeur n_B chiffrée avec cette clef de session. Bien sûr, cette clef de session n'est pas bonne du tout pour les communications entre Alice et Bob, car tout le monde peut l'extraire à partir du message d'Alice. Mais la dérivation des croyances de chaque entité à l'aide de la logique BAN mène à

la conclusion que ce protocole est un bon protocole d'authentification. Nessett en conclut que la logique BAN a été mise en défaut.

D'où vient le problème ? Est-ce un défaut de la logique elle-même ou un défaut de son usage ? Nessett pointe vers le fait que la logique BAN permet de dire qui reçoit et reconnaît une clef, mais ne permet pas de dire qui ne *devrait* pas en recevoir. Elle traite des croyances liées à l'authenticité d'un message sans traiter le problème de la confidentialité des clefs. Dans (Burrows et al., 1990b), les auteurs ont répondu en affirmant qu'ils avaient explicitement limité leurs analyses aux cas d'authentification entre entités *honnêtes*, ce qui exclut que leur logique puisse servir à détecter les problèmes de confidentialité. C'est bien grâce à l'assomption initiale A *believes* $A \overset{k_{AB}}{\longleftrightarrow} B$ que le protocole Nessett peut être validé par une analyse BAN, mais cette assomption n'est pas permise à l'intérieur des limites explicites de la logique BAN. L'erreur ne vient donc pas de BAN elle-même, mais de l'usage qu'on en a fait. Cette discussion a eu trois effets : (i) montrer les limites et la portée réelle des outils logiques qui commençaient à se développer ; (ii) encourager l'invention de logiques plus expressives, prenant en compte non seulement les requisits de l'authentification mais aussi ceux de la confidentialité ; (iii) ouvrir un débat sur la définition de ce qu'il faut entendre par « authentification ». Nous donnons maintenant un aperçu des discussions à ce sujet.

3.4 De la difficulté de définir l'authentification

La majeure partie des problèmes relatifs aux logiques de l'authentification sont nés de malentendus conceptuels sur ce qu'il faut entendre par « authentification ». Voici la définition officielle donnée par le standard international ISO/IEC 9798-1 (International Organization for Standardization, 1991) :

> « Les mécanismes d'authentification d'entités permettent la vérification, par une entité, de l'identité déclarée d'une autre entité. L'authenticité de l'entité peut être seulement assurée pour une instance d'échange d'authentification. »

Toute l'obscurité de cette définition se concentre dans l'expression « identité déclarée » : qu'est-ce qui définit l'identité d'une entité et par quels moyens cette identité est-elle déclarée ? D'une manière générale, c'est la formulation des requisits de l'authenticité en termes anthropomorphiques qui, si elle aide à abréger les descriptions d'échanges protocolaires, ne permet pas d'expliciter toutes les assomptions faites sur ces échanges. Cette remarque apparaît dans (Gollmann, 1996), dont l'article a été le premier à poser explicitement le problème d'une définition conceptuelle de l'authentification. Dans ce qui suit, nous nous intéresserons surtout à la distinction faite par (Abadi, 2000) entre attribution de l'*origine* et attribution de la *responsabilité*.

3.4.1 Attribution de la responsabilité et attribution de l'origine

La contribution qui nous semble conceptuellement la plus importante vient de (Abadi, 2000). L'auteur distingue deux sens de l'authentification. M est *authentifié* par B comme étant un message de A peut vouloir dire que :

- B attribue à A l'autorité sur le message M (lui assigne la responsabilité de M) ;
- B identifie M comme étant en provenance de A (lui assigne l'origine de M).

Prouver qu'une entité a envoyé un message est différent de prouver qu'elle en est responsable et encore différent de prouver qu'on peut lui en attribuer l'origine. La différence entre l'attribution de la responsabilité et l'attribution de l'orgine peut s'énoncer comme suit : l'assignation de la responsabilité établit un lien entre émetteur et auteur d'un message, celle de l'origine un lien entre émetteur et initiateur. Selon que l'authentification sera définie par le premier ou le deuxième de ces sens, la validité d'un protocole d'authentification sera déterminée différemment. Nous illustrons cette différence en donnant maintenant un exemple pour un protocole simple.

Soit un protocole où une entité A crée une paire de clefs à court terme – une clef publique et sa contrepartie secrète. A envoie la clef publique à court terme à une entité B, signant ce message avec sa clef secrète à long terme. Ensuite, A utilise la clef secrète à court terme qu'il vient d'envoyer à B pour signer les messages suivants.

Message 1 $\quad A \rightarrow B : \ A, B, \{K, A, B, T\}_{K_A^{-1}}$

Message 2 $\quad A \rightarrow B : \ A, B, \{\{M\}_{K^{-1}}\}_{K_B}$

Ici, M correspond à un message quelconque, T est un marqueur temporel, K_A est la clef publique à long terme de A, K_A^{-1} est la clef secrète correspondante, K est la clef publique à court terme de A, K^{-1} est la clef secrète correspondante. Les accolades sont traditionnellement utilisées pour la signature (comme dans $\{M\}_{K^{-1}}$) et pour le chiffrement (comme dans $\{\{M\}_{K^{-1}}\}_{K_B}$). Le message 2 n'est qu'un exemple illustrant l'utilisation par A de la clef à court terme K.

D'après une première interprétation de ce protocole, le message 1 contient l'information que A prend la responsabilité de la clef K, de telle sorte que B peut tenir A pour responsable de tous les messages signés avec K^{-1}. Par exemple, ce protocole semble adéquat pour une situation où B est un serveur de fichiers, A un client et M un message demandant l'effacement d'un fichier particulier. Si M est signé avec K^{-1}, B peut assigner à A la responsabilité de la demande. Ceci est aussi valable quand A a délégué sa responsabilité à un tiers : si A donne K a un tiers, ce tiers pourra faire la demande d'effacement de fichier au nom de A. En acceptant le premier sens d'authentification, cette demande ne sera pas considérée comme une entorse faite au protocole.

D'après une deuxième interprétation de ce protocole, B attribuera l'origine de tous les messages signés avec K^{-1} à A. Mais cette interprétation n'est pas justifiée. Si elle l'était, l'échange suivant passerait pour une violation des buts d'authentification du protocole.

Message 1	$A \to B:\ A, B, \{K, A, B, T\}_{K_A^{-1}}$	(intercepté par C)
Message 1 bis	$C \to B:\ C, B, \{K, C, B, T\}_{K_C^{-1}}$	
Message 2	$A \to B:\ A, B, \{\{M\}_{K^{-1}}\}_{K_B}$	(intercepté par C)
Message 2 bis	$C \to B:\ C, B, \{\{M\}_{K^{-1}}\}_{K_B}$	

Ici, C est un attaquant qui signe la clef publique générée par A. A la réception de M, B attribuerait l'origine du message à C plutôt qu'à A. Cette séquence de messages constitue donc une attaque si l'on choisit la deuxième définition de l'authentification, car la détermination de C comme étant à l'origine du message M est erronée. D'un autre côté, avec la première interprétation, C pourrait être faussement tenu pour responsable du message M.

Pour renforcer ce protocole, on peut (entre autres) exiger que A signe son nom avec la clef secrète qu'il vient de générer (K^{-1}). Soit :

Message 1	$A \to B:\ A, B, \{K, A, B, T\}_{K_A^{-1}}, \{A\}_{K^{-1}}$
Message 2	$A \to B:\ A, B, \{\{M\}_{K^{-1}}\}_{K_B}$

ou

Message 1	$A \to B:\ A, B, \{K, A, B, T\}_{K_A^{-1}}$
Message 2	$A \to B:\ A, B, \{\{A, M\}_{K^{-1}}\}_{K_B}$

Ces protocoles modifiés n'empêchent pas qu'un attaquant signe la clef K avec sa clef secrète K_C^{-1}, mais elles empêchent qu'il puisse signer son nom avec la clef K^{-1}, car C n'a pas accès à K^{-1}.

Nous nous sommes attardés sur cet exemple car il montre bien que ce qui est requis pour l'attribution de la responsabilité diffère grandement de ce qui est requis pour l'attribution de l'origine[43].

Conclusion

Nous avons proposé une définition générique de l'imitation comme référence expressive et dénotative. A partir de cette définition nous nous sommes penché sur le cas particulier des dispositions, tentant par là de montrer comment l'impossibilité d'authentifier une disposition permettait de défendre la validité du test de Turing. Nous avons ensuite mentionné les travaux de Turing en cryptographie, insistant sur la manière dont ils pouvaient éclairer la construction du jeu de l'imitation. Enfin, nous avons exploré les logiques de l'authentification pour voir ce qu'elles pouvaient nous apprendre de la notion d'authentification. C'est sur ce point que nous aimerions conclure.

En cherchant à circonscrire la notion d'authentification telle qu'elle est impliquée dans les protocoles cryptographiques, nous avons distingué l'attribution de la responsabilité et l'attribution de l'origine. Cette distinction est l'équivalent d'une distinction dépassant le champ des protocoles d'authentification et qu'une analyse plus poussée de l'authentification devrait éclairer davantage : celle entre *déférence* et *référence*. Par « référence » nous désignons le fait que le jugement d'authentification soit toujours contextuel, qu'il requiert ce contexte pour remplir sa fonction. Par « déférence » nous désignons la référence à l'autorité d'un tiers non impliqué directement dans le contexte - que celui-ci soit un serveur ou une norme sociale. L'enchevêtrement de la référence et de la déférence est ce qui rend l'analyse des protocoles d'authentification si délicate. Nous pressentons que c'est aussi ce qui rend l'analyse de l'imitation si passionnante.

Partie 4

Annexes

4.1 Eléments de cryptologie

On peut aborder la cryptologie, historiquement et techniquement, de deux points de vue différents : en s'intéressant soit aux différentes techniques de chiffrement, soit aux différentes techniques de déchiffrement des messages. Le premier point de vue est celui de la cryptographie, le second point de vue est celui de la cryptanalyse. Historiquement, on assiste tantôt à la domination de la cryptographie (dans les périodes où les systèmes cryptographiques résistent aux assauts des cryptanalystes), tantôt à celle de la cryptanalyse. Dans cette section, nous présentons quelques techniques classiques de cryptographie et de cryptanalyse.

L'idée fondamentale de la cryptographie est de permettre la communication entre deux personnes (Alice et Bernard) à travers un canal peu sûr, de telle sorte qu'un opposant (Eve) ne puisse pas comprendre ce qui est échangé. L'information qu'Alice souhaite transmettre à Bernard, le texte clair, peut être n'importe quel type de données. Alice transforme ce texte clair en texte chiffré par un procédé de chiffrement utilisant une clef prédéterminée, puis envoie le texte chiffré à Bernard au travers du canal. Eve peut espionner le canal et intercepter des textes chiffrés, mais seul Bernard peut déchiffrer les messages, car il possède la clef de déchiffrement. Par convention, nous notons en minuscule les éléments de texte clair et en majuscule les éléments de texte chiffré.

Techniques de cryptographie

Formellement, un système cryptographique est un quintuplet $(\mathcal{P}, \mathcal{C}, \mathcal{K}, \mathcal{E}, \mathcal{D})$ satisfaisant :

1. $\mathcal{P}$ est un ensemble fini de blocs de textes clairs possibles.
2. $\mathcal{C}$ est un ensemble fini de blocs de textes chiffrés possibles.
3. $\mathcal{K}$ est un ensemble fini de clefs possibles.
4. Pour tout $K \in \mathcal{K}$, il y a une règle de chiffrement $e_K \in \mathcal{E}$ et une règle de déchiffrement correspondante $d_K \in \mathcal{D}$. Chaque $e_K : \mathcal{P} \to \mathcal{P}$ et $d_K : \mathcal{C} \to \mathcal{P}$ sont des fonctions telles que $d_K(e_K(x)) = x$ pour tout texte clair $x \in \mathcal{P}$.

Chiffrement par décalage

Le chiffrement par décalage est basé sur l'arithmétique modulaire. ZZ_m symbolise l'ensemble $\{0, \ldots, m-1\}$ munis des deux opérations $+$ et $-$. L'addition et la multiplication dans ZZ_m fonctionnent exactement comme l'addition et la multiplication usuelles, excepté le fait que tous les résultats sont réduits modulo m.

Soit $\mathcal{P} = \mathcal{C} = \mathcal{K} = \mathrm{ZZ}_{26}$. Pour $0 \leq K \leq 25$, on définit

$$e_K(x) = x + K \bmod 26$$

et

$$d_K(y) = y - K \bmod 26$$

(x,$y \in \mathrm{ZZ}_{26}$).

FIG. 4.1 *Définition du chiffrement par décalage*

Les définitions de la figure 4.1 sont bien telles que $d_K(e_K(x)) = x$.

Le chiffrement par décalage se définit dans ZZ_{26} car il y a vingt-six lettres dans l'alphabet. Mais on peut théoriquement le définir pour n'importe quel m. Pour $K = 3$, le chiffrement par décalage est connu sous le nom de chiffrement de César : celui-ci chiffrait ses messages en décalant toutes les lettres de trois places dans l'alphabet.

Chiffrement par substitution

Le chiffrement par décalage est un cas particulier du chiffrement par substitution. Dans le chiffrement par substitution, les opérations de chiffrement et de déchiffrement sont des permutations sur l'ensemble des caractères alphabétiques.

Soit π une permutation aléatoire sur l'ensemble de l'alphabet. Dans ces tableaux, nous notons par convention en minuscule les éléments de $\mathcal{P}$ et en majuscule les éléments de $\mathcal{C}$.

a	b	c	d	e	f	g	h	i	j	k	l	m
X	N	Y	A	H	P	O	G	Z	Q	W	B	T

n	o	p	q	r	s	t	u	v	w	x	y	z
S	F	L	R	C	V	M	U	E	K	J	D	I

Ainsi, $e_\pi(\mathtt{h}) = \mathtt{G}$, $e_\pi(\mathtt{z}) = \mathtt{I}$, etc. A l'inverse, $d_\pi(\mathtt{G}) = \mathtt{h}$, $d_\pi(\mathtt{C}) = \mathtt{r}$, etc.

La clef est simplement la permutation des vingt-six caractères alphabétiques. Le nombre de clefs est donc 26!, ce qui rend computationnellement impossible un parcours exhaustif de l'espace des clefs.

Chiffrement de Vigenère

Dans le chiffrement par substitution (et donc *a fortiori* dans le chiffrement par décalage), dès qu'une clef est fixée, chaque caractère alphabétique est transformé en un caractère alphabétique

unique. Pour cette raison, le procédé est appelé *monoalphabétique*. On présente maintenant un système qui n'est pas monoalphabétique, le chiffrement de Vigenère (du nom de Blaise Vigenère, qui vécut au seizième siècle : voir (Singh, 1999)).

Soit m un entier strictement positif. Soit $\mathcal{P} = \mathcal{C} = \mathcal{K} = (\mathrm{ZZ}_{26})^m$. Pour toute clef $K = (k_1, k_2, \ldots, k_m)$ on définit

$$e_K(x_1, x_2, \ldots, x_m) = (x_1 + k_1, x_2 + k_2, \ldots, x_m + k_m)$$

et

$$d_K(y_1, y_2, \ldots, y_m) = (y_1 - k_1, y_2 - k_2, \ldots, y_m - k_m)$$

où les opérations sont effectuées dans ZZ_{26}.

FIG. 4.2 *Définition du chiffrement de Vigenère*

En utilisant la correspondance biunivoque $A \leftrightarrow 0$, $B \leftrightarrow 1, \ldots, Z \leftrightarrow 25$, on décrit chaque clef K par une chaîne de caractères de longueur m appelée *mot-clef*. Le chiffrement de Vigenère traite m caractères alphabétiques à la fois : chaque bloc de texte clair de longueur m est chiffré en fonction du mot-clef de même longueur. Prenons un exemple.

Supposons que le mot-clef est `CIPHER`. Converti en nombre, cela nous donne : $K = (2, 8, 15, 7, 4, 17)$. Supposons le texte clair suivant :

`thiscryptosystemisnotsecure`

Le chiffrement se fait par bloc de six (en fonction de la longueur du mot-clef) :

`thiscr-yptosy-stemis-notsec-ure`.

La première lettre de chaque bloc est chiffrée en fonction de la première lettre du mot-clef, et ainsi de suite. L'addition des valeurs se fait toujours modulo 26. Soit numériquement :

Clair	19	7	8	18	2	17	24	15	19	14	18	24
Mot-clef	2	8	15	7	4	17	2	8	15	7	4	17
Chiffré	21	15	23	25	6	8	0	23	8	21	22	15

Clair	18	19	4	12	8	18	13	14	19	18	4	2
Mot-clef	2	8	15	7	4	17	2	8	15	7	4	17
Chiffré	20	1	19	19	12	9	15	22	8	25	8	19

Le texte chiffré est donc :

`VPXZGIAXIVWPUBTTMJPWIZITWZT`

Pour déchiffrer, on utilise le même mot-clef, mais on fait une soustraction modulo vingt-six au lieu d'une addition. On remarque que le nombre de clefs possibles dans le chiffrement de Vigenère est 26^m, ce qui signifie que même pour des petites valeurs de m, une recherche exhaustive de clef demande beaucoup de temps. D'autre part, avec un mot-clef de longueur m, un caractère alphabétique peut être transformé en m caractères distincts (si tous les caractères

du mot-clef sont distincts). Un tel procédé est donc nommé polyalphabétique.

L'un des inconvénients du chiffrement de Vigenère est son incommodité. Pour le pallier, on a recours au carré de Vigenère, lequel accélère et rend sensiblement plus intuitive la procédure de chiffrement.

clair	A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
1	B C D E F G H I J K L M N O P Q R S T U V W X Y Z A
2	C D E F G H I J K L M N O P Q R S T U V W X Y Z A B
3	D E F G H I J K L M N O P Q R S T U V W X Y Z A B C
4	E F G H I J K L M N O P Q R S T U V W X Y Z A B C D
5	F G H I J K L M N O P Q R S T U V W X Y Z A B C D E
6	G H I J K L M N O P Q R S T U V W X Y Z A B C D E F
7	H I J K L M N O P Q R S T U V W X Y Z A B C D E F G
8	I J K L M N O P Q R S T U V W X Y Z A B C D E F G H
9	J K L M N O P Q R S T U V W X Y Z A B C D E F G H I
10	K L M N O P Q R S T U V W X Y Z A B C D E F G H I J
11	L M N O P Q R S T U V W X Y Z A B C D E F G H I J K
12	M N O P Q R S T U V W X Y Z A B C D E F G H I J K L
13	N O P Q R S T U V W X Y Z A B C D E F G H I J K L M
14	O P Q R S T U V W X Y Z A B C D E F G H I J K L M N
15	P Q R S T U V W X Y Z A B C D E F G H I J K L M N O
16	Q R S T U V W X Y Z A B C D E F G H I J K L M N O P
17	R S T U V W X Y Z A B C D E F G H I J K L M N O P Q
18	S T U V W X Y Z A B C D E F G H I J K L M N O P Q R
19	T U V W X Y Z A B C D E F G H I J K L M N O P Q R S
20	U V W X Y Z A B C D E F G H I J K L M N O P Q R S T
21	V W X Y Z A B C D E F G H I J K L M N O P Q R S T U
22	W X Y Z A B C D E F G H I J K L M N O P Q R S T U V
23	X Y Z A B C D E F G H I J K L M N O P Q R S T U V W
24	Y Z A B C D E F G H I J K L M N O P Q R S T U V W X
25	Z A B C D E F G H I J K L M N O P Q R S T U V W X Y
26	A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

TAB. 4.1 *Carré de Vigenère*

Après avoir choisi un mot-clef, on repère les rangées dont la première lettre coïncide avec l'une des lettres du mot-clef. Ici, nous avons choisi le mot-clef `KILO` : les lignes correspondantes à chacune des lettres de `KILO` sont grisées dans le carré de Vigenère. Pour chiffrer un texte, on met d'abord en correspondance chaque lettre du texte clair avec la lettre du mot-clef, puis on lit dans le tableau le résultat de la substitution. Dans la rangée commençant par un `K` (la dixième) et dans la colonne `t`, nous avons bien la lettre `D` ; dans la rangée commençant par un `I` (la huitième) et dans la colonne `h`, nous avons bien un `P` ; etc.

Clef	`KILOKILOKILOKILOKILOKILOK`
Clair	`thérussethéjasminthéchine`
Chiffré	`DPPFEADSDPPXKAXWXBSSMPTBO`

Cryptographie à clef publique

Dans la cryptographie à clef privée (parfois aussi improprement nommée « cryptographie à clef secrète »), la même clef est utilisée pour le chiffrement et le déchiffrement, e_K et d_K sont identiques : d'où aussi le nom de cryptographie *symétrique*. La sécurité de tels systèmes repose sur le fait que la clef reste privée. Leur inconvénient est évident : le secret du message repose sur le secret de la clef, alors même qu'Alice et Bernard doivent pouvoir se mettre d'accord sur cette clef.

L'objectif des systèmes à clef publique est de rendre la règle d_K impossible à découvrir à partir de la règle e_K. Ainsi, pour communiquer avec Alice, Bernard disposera de deux clefs : une clef publique personnelle grâce à laquelle Alice chiffrera son message, et une clef secrète grâce à laquelle Bernard pourra le déchiffrer. L'avantage pratique est immédiat : Alice et Bernard n'ont pas à se mettre d'accord au préalable sur le choix de la clef privée. Tout se passe comme si Bernard mettait son message dans un coffre dont seule Alice possède la combinaison. Remarquons tout de suite que ce système n'assure pas une sécurité absolue, mais seulement une sécurité calculatoire : comme la clef de chiffrement est publique, un individu disposant du texte chiffré y peut chiffrer chaque texte clair possible x jusqu'à ce qu'il trouve l'unique x tel que $y = e_K(x)$.

L'idée de système à clef publique est due à Diffie et Hellman (1976). La première réalisation d'un système à clef publique fut publiée en 1978 par Rivest, Shamir et Edleman : c'est le chiffrement RSA. Ce système, comme la plupart des systèmes à clef publique, repose sur la notion de fonction à sens unique avec trappe. Une fonction de chiffrement est à sens unique si on ne peut l'inverser en un temps raisonnable (i.e. si son inversion repose sur la résolution d'un problème NP-complet). Elle est dite « à trappe » s'il existe néanmoins une petite information à partir de laquelle on peut facilement l'inverser. Le fait qu'une fonction de chiffrement soit à sens unique interdit à quiconque de découvrir la fonction de déchiffrement, alors même que de nombreux textes chiffrés peuvent être connus. Le fait qu'elle soit à trappe donne la possibilité au destinataire autorisé, connaissant la « trappe » de déchiffrer facilement le message qui lui est adressé.

Techniques de cryptanalyse

Dans cette section, nous présentons quelques techniques de cryptanalyse. L'hypothèse généralement faite est que l'opposant, Eve, connaît le système cryptographique utilisé par Alice et Bernard. Il serait en effet douteux d'assurer la sécurité du système cryptographique sur la protection de sa description. Il y a plusieurs niveaux d'attaque possible, selon ce qu'Eve sait.

Texte chiffré connu Eve ne connaît que la chaîne du message chiffré y.

Texte clair connu Eve dispose d'un texte clair x et du texte chiffré y correspondant.

Texte clair choisi Eve a accès à une machine chiffrante ; elle peut ainsi choisir son texte clair x et obtenir son texte chiffré y.

Texte chiffré choisi Eve a temporairement accès à une machine déchiffrante ; ainsi, elle peut choisir un texte chiffré y et obtenir son texte déchiffré x.

Ces niveaux sont énumérés par ordre de difficulté décroissante. Dans tous les cas, Eve a pour but de déterminer la clef utilisée.

La plupart des techniques de cryptanalyse s'appuient sur les propriétés statistiques de la langue écrite. A partir d'une grande masse de données, on a calculé la fréquence d'apparition de chaque lettre de l'alphabet, ainsi que celle de chaque couple de lettres (digramme). En faisant de même avec un texte chiffré, on peut établir des corrélations probables. En commençant par celles qui sont les plus certaines et en éliminant au fur et à mesure les résultats incompatibles, on déchiffre progressivement le texte.

lettre	probabilité	lettre	probabilité
A	0,082	N	0,067
B	0,015	O	0,075
C	0,028	P	0,019
D	0,043	Q	0,001
E	0,127	R	0,060
F	0,022	S	0,063
G	0,020	T	0,091
H	0,061	U	0,028
I	0,070	V	0,010
J	0,002	W	0,023
K	0,008	X	0,001
L	0,040	Y	0,020
M	0,024	Z	0,001

TAB. 4.2 *Probabilité d'occurrence des lettres de l'alphabet en anglais*

La table de probabilité des lettres permet de classer les lettres en plusieurs groupes[44].

1. `E`, ayant pour probabilité 0,120 ;
2. `T`, `A`, `O`, `I`, `N`, `S`, `H` et `R`, ayant une probabilité entre 0,06 et 0,09 ;
3. `D` et `L` ayant une probabilité d'environ 0,04 ;
4. `C`, `U`, `M`, `W`, `F`, `G`, `Y`, `P` et `B`, ayant une probabilité entre 0,015 et 0,023 ;
5. `V`, `K`, `J`, `X`, `Q` et `Z`, ayant une probabilité inférieure à 0,01.

A ce classement peut s'ajouter celui des digrammes et des trigramme. Voici la probabilité d'apparition des digrammes en ordre décroissant : `TH`, `HE`, `IN`, `ER`, `AN`, `RE`, `ED`, `ON`, `ES`, `ST`, `EN`, `AT`, `TO`, `NT`, `HA`, `ND`, `OU`, `EA`, `NG`, `AS`, `OR`, `TI`, `IS`, `ET`, `IT`, `AR`, `TE`, `SE`, `HI` et `OF`. Et voici celle d'apparition des douze premiers trigrammes, par ordre décroissant : `THE`, `ING`, `AND`, `HER`, `ERE`, `ENT`, `THA`, `NTH`, `WAS`, `ETH`, `FOR` et `DTH`.

Cryptanalyse du chiffrement par substitution

Le chiffrement par décalage (modulo 26) peut être cryptanalysé par la méthode de recherche exhaustive des clefs. Puisque ce chiffrement conserve l'ordre des lettres de l'alphabet, il suffit de tenter les 26 clefs possibles et de tomber sur le texte clair.

Le chiffrement par substitution ne fait pas intervenir d'opération proprement algébrique, aussi sa cryptanalyse est-elle plus délicate. Elle repose néanmoins sur la méthode simple d'analyse des fréquences d'apparition des lettres. Nous illustrons cette méthode dans un exemple.

Soit le texte chiffré :

```
YIFQFMZRWQFYVECFMDZPCVMRZWNMDZVEJBTXCDDUMJ
NDIFEFMDZCDMQZKCEYFCJMYRNCWJCSZREXCHZUNMXZ
NZUCDRJXYYSMRTMEYIFZWDYVZVYFZUMRZCRWNZDZJJ
XZWGCHSMRNMDHNCMFQCHZJMXJZWIEJYUCFWDJNZDIR
```

L'analyse des fréquences pour ce texte chiffré est donnée dans le tableau 4.3.

lettre	fréquence	lettre	fréquence
A	0	N	9
B	1	O	0
C	15	P	1
D	13	Q	4
E	7	R	10
F	11	S	3
G	1	T	2
H	4	U	5
I	5	V	5
J	11	W	8
K	1	X	6
L	0	Y	10
M	16	Z	20

TAB. 4.3 *Fréquence des vingt-six lettres dans l'exemple de texte chiffré*

Comme `Z` apparaît plus souvent que toutes les autres lettres, on peut conjecturer que $d_K(\texttt{Z}) = \texttt{e}$. Les autres caractères qui apparaissent au moins dix fois chacun sont `C`, `D`, `F`, `J`, `M`, `R`, `Y`. On peut supposer que ces lettres sont chiffrées à partir de caractères parmi `t`, `a`, `o`, `i`, `n`, `s`, `h`, `r`, mais leur fréquence ne varie pas suffisamment pour que l'on puisse établir une correspondance plus probable.

On peut étudier les digrammes, surtout ceux de la forme `-Z` ou `Z-`, puisque l'on a déjà reconnu la lettre `Z`. On trouve que les digrammes de ce type les plus fréquents sont `DZ` et `ZW` (quatre occurrences chacun), et `RZ`, `HZ`, `XZ`, `FZ`, `ZR`, `ZV`, `ZC`, `ZD`, `ZJ` (deux occurrences chacun). Comme `ZW` apparaît quatre fois et `WZ` aucune, et que `W` apparaît moins que plusieurs autres caractères, on peut conjecturer que $d_K(\texttt{W}) = \texttt{d}$. En effet, toujours avec l'hypothèse $d_K(\texttt{Z}) = \texttt{e}$, cela nous donne $d_K(\texttt{ZW}) = \texttt{ed}$ et $d_K(\texttt{WZ}) = \texttt{de}$; on a une bonne corrélation entre la fréquence de `ED` et celle de `ZW`, ainsi qu'entre la rareté de `DE` et l'absence de `WZ`. D'autre part, comme `DZ` apparaît quatre fois et `ZD` deux, on imagine que $d_K(\texttt{D}) \in \{\texttt{r}, \texttt{s}, \texttt{t}\}$, car `RE`, `SE` et `TE` figurent parmi les digrammes les plus fréquents ; mais nous ne pouvons pas encore décider laquelle de ces trois lettres est la bonne.

Sous l'hypothèse $d_K(\texttt{Z}) = \texttt{e}$ et $d_K(\texttt{W}) = \texttt{d}$, en regardant une nouvelle fois le texte chiffré, on constate que `ZRW` et `ZWR` apparaissent au début du texte, et que `RW` apparaît plus loin. Comme `R` apparaît souvent et que `nd` est un digramme fréquent, on peut supposer que $d_K(\texttt{R}) = \texttt{n}$.

Voici où nous en sommes de la cryptanalyse :

```
------end---------e----ned---e------------
```

```
YIFQFMZRWQFYVECFMDZPCVMRZWNMDZVEJBTXCDDUMJ
--------e----e---------n--d---en----e----e
NDIFEFMDZCDMQZKCEYFCJMYRNCWJCSZREXCHZUNMXZ
-e---n------n------ed---e---e--ne-nd-e-e--
NZUCDRJXYYSMRTMEYIFZWDYVZVYFZUMRZCRWNZDZJJ
-ed-----n-----------e----ed-------d---e--n
XZWGCHSMRNMDHNCMFQCHZJMXJZWIEJYUCFWDJNZDIR
```

La suite procède de la même manière, les hypothèses s'appuyant les unes sur les autres, et se confirmant au fur et à mesure que le texte est déchiffré. L'aspect artisanal de cette technique n'est pas contingent : les corrélations de fréquence guident le tâtonnement de manière rigoureuse, mais le cryptanalyste est toujours en train d'évaluer la plausibilité des hypothèses à sa disposition. Cette évaluation ne peut jamais être intégralement mécanisée.

Cryptanalyse du chiffrement de Vigenère

Présentation informelle Le défi que représentait la cryptanalyse du chiffre de Vigenère fut surmonté pour la première fois par Charles Babbage en 1854, suivi de près par Friedrich W. Kasiski (Singh, 1999, p. 93). Le principe de cette cryptanalyse est simple, même si la mise en œuvre peut parfois devenir complexe. Nous présentons d'abord la méthode de manière informelle, avant d'expliciter la procédure de manière rigoureuse.

La force du chiffrement de Vigenère est de chiffrer une lettre ou un mot de plusieurs manières différentes à l'intérieur d'un même texte. Mais la faiblesse de ce chiffrement tient au fait que ce nombre de manières distinctes de chiffrer une lettre ou un mot est strictement égal à la longueur du mot-clef. Ceci signifie notamment que pour un mot-clef de quatre lettres (comme `KILO`), un mot ne pourra être chiffré que de quatre manières distinctes. Si un mot apparaît plus de quatre fois, alors il sera chiffré au moins deux fois de la même manière ; cette simple répétition suffit à donner prise à la cryptanalyse.

La marche à suivre consistera d'abord à déterminer la longueur de la clef (l). Pour cela, on recherche les séquences de lettres qui se répètent dans le texte chiffré. Deux explications sont envisageables pour cette répétition : soit la même séquence de lettres du texte clair a été cryptée avec la même partie de la clef, soit différentes parties du texte clair ont été chiffrées avec des parties différentes de la clef, se trouvant alors juxtaposées par pure coïncidence. Plus les séquences répétées sont longues, plus ce genre de coïncidences est improbable. Dans l'exemple suivant, nous rechercherons donc les séquences répétées d'au moins quatre lettres.

```
XAUNMEESYIEDTLLFGSNBWQ
UFXPQTYORUTYIINUMQIEUL
SMFAFXGUTYBXXAGBHMIFII
MUMQIDEKRIFRIRZQUHIENO
OOIGRMLYETYOVQRYSIXEOK
IYPYOIGRFBWPIYRBQURJIY
EMJIGRYKXYACPPQSPBVESI
RZQRUFREDYJIGRYKXBLOPJ
ARNPUGEFBWMILXMZSMZYXP
NBPUMYZMEEFBUGENLRDEPB
JXONQEZTMBWOEFIIPAHPPQ
```

BFLGDEMFWFAHQ

Nous trouvons trois séquences : U-M-Q-I, O-I-J-R et J-I-G-R-Y. Si nous prenons la première, nous observons qu'elle se répète après trente lettres. Si cette répétition est due au fait que cette séquence est le chiffre d'une même séquence de texte clair, cela permet d'éliminer les longueurs de clef qui ne sont pas des facteurs entiers de trente. En faisant le même raisonnement avec les deuxième et troisième séquences répétées, nous obtenons le tableau suivant.

Séquence répétée	Espace de répétition	Longueurs de clef possibles
U-M-Q-I	30	2, 3, **5**, 6, 10, 15
O-I-J-R	25	**5**
J-I-G-R-Y	30	2, 3, **5**, 6, 10, 15

TAB. 4.4 *Espaces de répétitions de séquences dans l'exemple de texte chiffré*

Ce tableau indique clairement que la seule longueur de clef possible est $l = 5$: pour tout autre longueur de clef, la répétition des séquences chiffrées mentionnées ci-dessus ne pourrait pas correspondre à une répétition de séquences identiques du texte clair, mais devrait relever de la simple coïncidence.

Une fois cette longueur déterminée, le chiffrement polyalphabétique n'est plus qu'une combinaison de l chiffrements monoalphabétiques (par décalage). Pour achever la cryptanalyse, il suffit de morceler le texte chiffré en l textes chiffrés : le premier de ces blocs de texte sera le résultat du chiffrement obtenu par la première lettre du mot-clef, le deuxième sera le résultat du chiffrement obtenu par la deuxième lettre, etc. Dans le cas qui nous intéresse, nous formons un premier bloc de texte par concaténation de la 1ère lettre, de la 6ème, de la 11ème lettre, etc (soit X-E-E-F-...). L'analyse statistique de ce premier bloc de texte révèle que la lettre I est la plus fréquente. On peut donc raisonnablement supposer que ce I correspond à la lettre e du texte clair. Ainsi, le décalage provoqué par la première lettre du mot-clef (appelons-la L_1) est tel que $e_{L_1}(\texttt{e}) = \texttt{I}$; en consultant le carré de Vigenère, nous concluons immédiatement $L_1 = \texttt{E}$.

En répétant cette procédure pour les quatre autres blocs de textes, nous obtenons aisément les quatre dernières lettres du mot-clef : EMAUX. En utilisant le carré de Vigenère, nous déchiffrons alors le texte :

Clef	EMAUXEMAUXEMAUXEMAUXEMAUXEMAUXEMAUX...
Chiffré	XAUNMEESYIEDTLLFGSNBWQUFXPQTYORUTYI...
Clair	toutpasselartrobusteseulaleternitel...

Présentation formelle Pour déterminer la longueur de la clef, nous utilisons l'indice de coïncidence[45]. Pour déterminer la clef elle-même, nous utilisons l'indice de coïncidence mutuel.

Soit $x = x_1 x_2 \dots x_n$ une chaîne de n caractères alphabétiques. L'indice de coïncidence de x, noté $I_c(x)$, est la probabilité que deux caractères aléatoires de x soient identiques. On note respectivement $f_0, f_1, \dots, f_{25}$ les fréquences de A, B, C, ..., Z. Il y a $\binom{n}{2}$ choix possibles de deux caractères de x[46]. Pour chaque i, $0 \leq i \leq 25$, il y a $\binom{f_i}{2}$ façons de choisir deux caractères i. Donc on a la formule

$$I_c(x) = \frac{\sum_{i=0}^{25} f_i(f_i - 1)}{n(n-1)}$$

Supposons que x soit un texte écrit en anglais. En notant $p_0, p_1, p_2, \ldots, p_{25}$ les probabilités d'apparition des lettres `A`, `B`, ..., `Z` de la table des fréquences (voir le tableau 4.2), on peut s'attendre à ce que

$$I_c(x) \approx \sum_{i=0}^{25} p_i^2 = 0{,}065$$

car la probabilité que deux caractères soient identiques à `A` est ${p_0}^2$, la probabilité que deux caractères soient identiques à `B` est ${p_1}^2$, etc.

Supposons maintenant que l'on ait un texte obtenu par le chiffrement de Vigenère. On définit m sous-chaînes $y_1, y_2, \ldots, y_m$ de y en écrivant le texte chiffré colonne par colonne, dans un tableau de dimensions $m \times (n/m)$. Les lignes de ce tableaux sont les sous-chaînes y_i, $1 \leq i \leq m$. Si m est bien la longueur du mot-clef, chaque $I_c(y_i)$ doit être proche de 0,065 (puisque chaque sous-chaîne y_i est le résultat du chiffrement de x par simple décalage). Inversement, si m n'est pas la longueur du mot-clef, les sous-chaînes y_i devront apparaître comme plus aléatoires, leur indice de coïncidence devra être plus petit. On montre qu'une chaîne complètement aléatoire est telle que

$$I_c \approx 26(1/26)^2 = 1/26 = 0{,}038$$

La différence entre cette valeur et la valeur de l'indice de coïncidence d'un texte clair (0,065) est suffisamment significative pour que l'on puisse déterminer sans problème la longueur du mot clef (m).

Définissons maintenant l'indice de coïncidence mutuel de deux chaînes. Soit $x = x_1 x_2 \ldots x_n$ et $y = y_1 y_2 \ldots y_{n'}$ deux chaînes de longueur respective n et n'. L'indice de coïncidence mutuel de x et y, noté $MI_c(x, y)$ est la probabilité pour qu'un caractère aléatoire de x soit égal à un caractère aléatoire de y. En notant $f_0, f_1, \ldots, f_{25}$ et ${f'}_0, {f'}_1, \ldots, {f'}_{25}$ les fréquences respectives de `A`, `B`, `C`, ..., `Z` dans x et y, on a

$$MI_c(x, y) = \frac{\sum_{i=0}^{25} f_i {f'}_i}{nn'}$$

Si l'on a trouvé la bonne valeur de m, les sous-chaînes y_i sont obtenues par un chiffrement par décalage. Supposons que la clef soit $K = (k_1, k_2, \ldots, k_m)$ et regardons si l'on peut estimer $MI_c(y_i, y_j)$. Si l'on prend un caractère aléatoire de y_i et un caractère aléatoire de y_j, la probabilité pour qu'ils correspondent tous les deux à `A` est $p_{-k_i} p_{-k_j}$, car les sous-chaînes y_i et y_j sont respectivement produites à partir d'un décalage de x_i par k_i et de x_j par k_j. De même, la probabilité pour qu'ils correspondent à `B` est $p_{1-k_i} p_{1-k_j}$, etc. On a donc l'estimation suivante

$$MI_c(y_i, y_j) \approx \sum_{i=0}^{25} p_{h-k_i} p_{h-k_j} = \sum_{i=0}^{25} p_h p_{h+k_i-k_j}$$

Autrement dit, l'indice de coïncidence mutuel de deux sous-chaînes de y dépend seulement du décalage existant entre les lettres du mot-clefs qui servent au chiffrement de ces chaînes : par exemple, l'indice de coïncidence mutuel des deux premières sous-chaînes y_1 et y_2, l'une étant le résultat du chiffrement par décalage à partir de k_1 et l'autre à partir de k_2, dépend seulement du décalage entre k_1 et k_2 modulo 26. Plus généralement, $MI_c(y_i, y_j)$ ne dépend que

de la différence $k_i - k_j \bmod 26$, appelée *décalage relatif* de y_i et y_j. Si le décalage relatif est nul, alors l'indice de coïncidence mutuel des deux sous-chaînes est environ de 0,065.

L'idée est ensuite de rechercher les valeurs probables des décalages relatifs de y_i et y_j : pour cela, on fixe y_i et on considère l'effet obtenu en chiffrant y_j à l'aide du chiffrement par décalage avec les clefs 0, 1, 2,..., 25. Cela revient à chercher la différence de décalage qu'il y a entre le chiffrement à l'aide de la clef k_i et celui réalisé à partir de k_j. Notons respectivement $y_j^0, y_j^1, \ldots, y_j^{25}$ la valeur de ce chiffrement. Ce que l'on veut, c'est trouver la valeur g telle que le décalage relatif entre y_i et y_j^g soit nul. Cette valeur de g nous donnera le décalage entre k_i et k_j.

On fait les calculs en parcourant l'ensemble des valeurs de i, j, g, $0 \leq i \leq j \leq g \leq 25$, on relève pour chaque couple $\{i, j\}$ la valeur de g telle que le décalage relatif de y_i et y_j^g soit nul, et l'on en conclut $k_i - k_j = g$. Une fois que l'on connaît toutes les relations que les lettres du mot-clef ont entre elles, il ne reste plus qu'à essayer les vingt-six possibilités restantes pour le mot clef.

4.2 *Enigma*

Historique

La machine à chiffrer *Enigma* fut élaborée au lendemain de la première guerre mondiale par l'allemand Arthur Scherbius. Il breveta son invention en 1918 et tenta immédiatement de la vendre à l'armée allemande, mais celle-ci attendit 1923 avant d'être convaincue de son efficacité et de l'importance que pouvait jouer la cryptographie dans un conflit. L'année 1923 est en effet celle de la parution de *The World Crisis* de Winston Churchill, ouvrage dans lequel celui-ci relate notamment les succès de la cryptographie alliée. La production en série de machines *Enigma* commence donc en 1925 : l'armée en achètera plus de 30 000 durant les dix années qui suivent. Cette machine devenue mythique a joué un double rôle pendant la seconde guerre mondiale : d'abord, elle dérouta longtemps les Alliés par la complexité de son système de chiffrement ; puis, peu à peu, les cryptanalystes de Bletchley Park en Grande Bretagne parvinrent à déchiffrer des messages de plus en plus nombreux. C'est la dissymétrie entre l'inébranlable confiance allemande en son système de chiffrement et les percées discrètes des alliées qui permit de retourner cette arme contre son possesseur. Si l'importance de la cryptanalyse dans la victoire des Alliés reste sujette à controverse, le rôle d'*Enigma* dans la cryptanalyse est clair : c'est cette machine qui donna sa réelle impulsion à la mécanisation du décodage, laquelle donna à entrevoir la manière dont une machine pouvait traiter de grandes masses d'informations.

Le fonctionnement d'*Enigma*

Enigma se compose de trois éléments : un clavier, un système de brouillage et un écran lumineux. Une lettre du texte clair est tapée au clavier, associée à une autre à l'intérieur d'un brouilleur, puis affichée sur l'écran lumineux. Le brouilleur est associé à un réflecteur, lequel rend chiffrement et déchiffrement symétriques.

L'idée principale de Scherbius fut de permettre au brouilleur de tourner sur lui-même. Après qu'une lettre est tapée, le brouilleur (ou rotor) se décale d'un vingt-sixième de tour. Ceci permet un chiffrement polyalphabétique dont les vingt-six clefs possibles sont les vingt-six

positions possibles du rotor. La faiblesse de ce système vient du fait qu'une même lettre répétée vingt-sept fois se verra chiffrée au moins deux fois de la même manière : on retrouve ici la faille du chiffrement de Vigenère, dont la cryptanalyse était déjà bien connue à l'époque où Scherbius conçoit sa machine.

La force principale d'*Enigma* est d'être mécanique. Ceci permet deux choses : compliquer à loisir le système mécanique de brouillage et utiliser des mot-clefs aléatoires. Ces deux avantages sont l'envers des inconvénients du chiffrement de Vigenère « manuel » : ce chiffrement était fastidieux à mettre en œuvre, et l'usage de mot-clefs de la langue courante en facilitait la cryptanalyse.

Pour ce qui est de la complication mécanique, Scherbius ne se priva pas. Premièrement, il ajouta deux autres rotors au premier, ce qui porte à $26 \times 26 \times 26 = 17\,576$ le nombre de positions-clefs différentes. Un chiffrement de Vigenère classique correspondrait à la rotation d'un seul brouilleur sur lequel serait criculairement inscrit le mot-clef : inversement, le chiffrement d'*Enigma* correspond à un chiffrement de Vigenère dont le mot-clef est d'une longueur de 17 576. De plus, les trois rotors sont amovibles, ce qui permet de les disposer selon $3! = 6$ combinaisons possibles. Deuxièmement, il introduisit un tableau de connexions à fiches entre le clavier et le premier rotor. Ce tableau de connexions apparie six lettres du clavier avec six lettres différentes du premier rotor. Les possibilités d'un tel appariement sont énormes : 100 391 791 500. Le nombre de clefs possibles est obtenu par multiplication de ces trois nombres, soit environ 10 000 000 000 000 000.

Pour se servir d'*Enigma*, les allemands disposaient de carnets de codes, lesquels indiquaient quel code utiliser pour chaque jour. Voir la figure 4.3 pour la forme que prenait ce code.

- Branchements de la table de connexions : `A/L-P/R-T/D-B/W-K/F-O/Y`
- Réglage des rotors : `2-3-1`
- Orientations des rotors : `Q-C-W`

FIG. 4.3 *Exemple de code pour Enigma*

La clef du jour ne servait pas à chiffrer directement les messages, mais à chiffrer la clef de chaque message. Par exemple, si un message était chiffré à partir de l'orientation `B-Z-G` des rotors, ces trois lettres étaient doublées en `B-Z-G-B-Z-G` (par précaution), ces six lettres étaient ensuite chiffrées à l'aide de la clef du jour, puis le résultat était accolé en tête du message. Cette opération visait à réduire le nombre de messages chiffrés avec la même clef, afin notamment de limiter les ressources d'une attaque statistique.

La cryptanalyse d'*Enigma*

La cryptanalyse d'*Enigma* mêle événements historiques et éléments techniques. Parmi les événements historiques, le plus important est certainement la trahison de l'allemand Hans-Thilo Schmidt, lequel livre les plans d'*Enigma* aux Alliés dès 1931. Il continua ensuite à leur vendre des carnets de code jusqu'en 1938, mais cessa toute activité au moment même où les Allemands renforçaient *Enigma* en portant à cinq le nombre de rotors. La cryptanalyse d'*Enigma* n'aurait pas été possible sans ces carnets de code ni tous ceux que les Alliés purent se procurer, par un moyen ou par un autre.

Pour ce qui est des techniques de cryptanalyse proprement dites, deux noms sont à retenir : Martin Rejewski, travaillant pour les services secrets polonais, et Alan Turing, travaillant pour les services secrets anglais à Bletchley Park.

Rejewski et les bombes

Rejewski chercha un moyen de distinguer dans un message chiffré ce qui relevait seulement de l'orientation des rotors. Pour cela, il s'appuya sur le fait que la première et la quatrième lettre d'un message chiffré était liées à une lettre identique du message clair. Par exemple, si les six premières lettres d'un message étaient `L-O-K-R-G-M`, il en concluait que le mot-clef du jour imposait un lien entre `L` et `R`. Disposant de plusieurs messages par jour, Rejewski put mettre en correspondance biunivoque l'ensemble des premières lettres d'un message et celui des quatrièmes lettres. Soit par exemple :

Première lettre	`ABCDEFGHIJKLMNOPQRSTUVWXYZ`
Quatrième lettre	`FQHPLWOGBMVRXUYCZITNJEASDK`

Il chercha ensuite le lien entre cette correspondance et l'orientation des rotors. Il observa que la correspondance entre les deux alphabets peut être décrite en terme de chaînes de caractères : le `A` est chiffré en `F`, lui-même chiffré en `W`, lui-même chiffré en `A`. Soit sur l'ensemble de l'alphabet :

`A` → `F` → `W` → `A`	3 liens
`B` → `Q` → `Z` → `T` → `V` → `E` → `L` → `R` → `I` → `B`	9 liens
`C` → `H` → `S` → `O` → `Y` → `D` → `P` → `C`	7 liens
`J` → `M` → `X` → `G` → `K` → `N` → `U` → `J`	7 liens

Répétant ces relevés de chaînes avec les deuxièmes et cinquièmes lettres, puis les troisièmes et sixièmes lettres, il obtenait ainsi un indice au sujet du mot-clef quotidien : celui était lié à un certain nombre de chaînes, ces chaînes étant elles-mêmes composées d'un certain nombre de liens. Le tout était de démêler, dans cet indice, ce qui relevait de l'orientation des rotors de ce qui était dû au tableau de connexions. Mais Rejewski fit l'observation décisive suivante : quelle que soit la permutation des lettres, le nombre de liens à l'intérieur des chaînes reste fixe, ainsi que (*a fortiori*) le nombre de chaînes. Autrement dit, le relevé du nombre de chaînes et du nombre de liens donne bien plus qu'un simple indice sur le mot-clef : ce relevé en est une véritable «empreinte».

A partir de cette observation, Rejewski prit un an pour constituer un répertoire associant chaque empreinte aux 105 456 orientations possibles des trois rotors. Ceci lui permit de s'attaquer au rôle que jouait le tableau de connexions. Puisqu'il connaissait désormais quel mot-clef était associé à un texte chiffré, il pouvait taper celui-ci sur une *Enigma* réglée avec le mot-clef approprié : le résultat obtenu était chiffré à partir du seul tableau de connexions. A l'intérieur de ces messages apparaissaient quelques bribes lisibles, dues au fait que le tableau de connexions ne permute pas toutes les lettres. Tombant par exemple sur « alliveaberlin », il en concluait que le `l` et le `r` devaient être permutés pour donner « arriveraberlin ». Ainsi, d'hypothèses en confirmation, le texte clair faisait son apparition.

La seconde contribution importante de Rejewski est d'ordre mécanique. N'importe quelle modification dans la méthode de transmission des messages allemands rendait obsolète le répertoire qu'il avait mis un an à dresser. Aussi était-il urgent de mécaniser la procédure de création des répertoires. En adaptant *Enigma*, Rejewski construisit les fameuses bombes : ces machines testaient les 17546 orientations possibles des rotors et indiquaient en sortie leur empreinte. Aidé de ces bombres, Rejewski pouvait trouver la clef du jour en moins de deux heures.

Alan Turing et les cribs

Les travaux de Rejewski furent décisifs. Mais ils reposaient sur deux faits contingents : d'une part sur le fait qu'*Enigma* ne fonctionnait qu'avec trois rotors ; d'autre part sur le fait que les Allemands répétaient l'énoncé du mot-clef en début de message. Dès que les Allemands employèrent cinq rotors, Rejewski fut techniquement pris de cours. Mais il était aussi à prévoir que les Allemands cesseraient un jour de répéter le mot-clef en début de message : le problème ne serait plus seulement technique mais aussi théorique, car la méthode mise au point par Rejewski exploite principalement cette répétition.

L'objectif d'Alan Turing était de permettre la cryptanalyse d'*Enigma* même en l'absence de cette répétition du mot-clef. On disposait déjà d'une astuce reposant sur le fait que les mot-clefs n'étaient pas toujours aléatoires. Souvent, par facilité, les allemands choisissaient des séquences de trois lettres se suivant sur le clavier, ou bien des séquences linguistiquement aisées à repérer. Ces mot-clefs furent baptisés des « *cillies* ». Dans la recherche des mot-clefs, les *cillies* étaient prioritaires, et les Anglais exploitaient souvent cette faiblesse strictement humaine de la cryptographie ennemie.

C'est aussi en s'intéressant au fait que les messages chiffrés sont produits par des êtres humains que Turing élabore la méthode des mots probables (nommés *cribs* en anglais). Cette méthode s'appuie sur la stéréotypie des messages chiffrés : on savait par exemple qu'un bulletin météo était souvent émis à six heures du matin. Ayant intercepté un tel message, il suffisait ensuite d'estimer la position du mot clair `wetter`, d'extraire les six lettres chiffrées censées lui correspondre (disons `ETJWPX`), puis d'identifier les réglages d'*Enigma* qui transformaient `wetter` en `ETJWPX`.

Turing exploita les *cribs* de la même manière que Rejewski exploita la répétition du mot-clef, le but étant aussi d'isoler les effets des rotors de ceux du tableau de connexions. Pour cela, il repéra les *cribs* dont l'équivalent chiffré permettait qu'une boucle se forme entre les lettres claires et les lettres chiffrées, de même qu'une boucle se formait entre le mot-clef répété et son chiffrement en six lettres. Dans l'exemple que nous avons choisi, une boucle se forme ainsi : `w` $\rightarrow$ `E` $\rightarrow$ `t` $\rightarrow$ `W`. Cette boucle signifie qu'il existe une disposition des rotors d chiffrant `w` et `E`, une autre $d+1$ chiffrant `e` en `T`, et une autre $d+3$ chiffrant `t` en `W`.

L'idée de Turing fut de faire marcher trois bombes en parallèle : la première chercherait l'orientation des rotors chiffrant `w` en `E`, la deuxième chercherait l'orientation pour `e` $\rightarrow$ `T`, et la troisième pour `t` $\rightarrow$ `W`. A première vue, cette division du travail n'apporte rien : chaque bombe doit toujours tester toutes les orientations possibles pour trouver la bonne. Mais Turing synchronisa les trois bombes en initialisant la première à l'état e, la deuxième à l'état $e+1$ et la troisième à l'état $e+3$; de cette manière, les trois bombes devaient tomber sur la bonne orientation des rotors au même instant. Ensuite, il les brancha entre elles, reliant la sortie de la première à l'entrée de la seconde, la sortie de celle-ci à l'entrée de la troisième et la sortie de la troisième à l'entrée de la première : la boucle allant du texte clair au texte chiffré

à l'intérieur d'un *crib* s'incarnait donc dans un circuit. Ce circuit avait deux avantages. Le premier, purement pragmatique, était de pouvoir installer une ampoule de contrôle : lorsque les trois bombes tombaient en même temps sur la bonne orientation de leurs rotors, le courant passait et l'ampoule s'allumait. Le second, de loin le plus important, était d'annuler le brouillage induit par le tableau de connexions. En effet, quand bien même l'effet du tableau de connexion viendrait s'ajouter à celui des rotors pour chiffrer `w` en `E`, le fait de brancher la sortie de la première bombe à l'entrée de la deuxième vient inverser cet effet : la permutation qui intervient entre `w` et `E` s'ajoute à la permutation inverse qui s'intercale entre `e` et `T`. Etant donné que la boucle entre les trois bombes est telle qu'il y a exactement autant de permutations dans un sens que dans l'autre, alors ces permutations n'ont aucune incidence sur le passage du courant dans le circuit. Ainsi Turing évitait les milliards de possibilités de clefs générées par le tableau de connexion.

4.3 *Banburismus* et théorie de l'information

Dans cette section nous présentons la formalisation proposée par Turing de la notion de « poids d'évidence » ainsi que l'unité de mesure « ban ». Nous introduisons les idées de Turing en présentant d'abord le théorème de Bayes, sur lequel elles s'appuient, et en montrant ses liens avec la théorie de l'information de C. Shannon. Pour Turing aussi bien que pour Shannon, ces réflexions ont émergé comme réponses à des problèmes de cryptologie.

Théorème de Bayes

Soit X et Y des distributions sur des ensembles d'événements élémentaires différents. On note $P(x)$ la probabilité d'un événement x suivant X et $P(y)$ la probabilité d'un événement y suivant Y. La probabilité mutuelle $P(x, y)$ est la probabilité que x et y soient réalisés simultanément. La probabilité conditionnelle $P(x|y)$ représente la probabilité de x sachant que y est réalisé. Les distributions X et Y sont dites indépendantes si $P(x, y) = P(x)P(y)$ pour tout x et y possibles.

La probabilité mutuelle est liée à la probabilité conditionnelle par la formule

$$P(x, y) = P(x|y)P(y).$$

En échangeant x et y on a

$$P(y, x) = P(y|x)P(x).$$

De ces deux formules, on obtient immédiatement le théorème de Bayes :

$$P(x|y) = \frac{P(x)P(y|x)}{P(y)}$$

Un corollaire immédiat de ce théorème est que X et Y sont des distributions indépendantes si et seulement si $P(x|y) = P(x)$ pour tout x et pour tout y.

Ban et poids d'évidence

L'intérêt du théorème de Bayes en épistémologie est de servir de base à la formalisation du gain en plausibilité (ou en « poids d'évidence ») d'une théorie ou d'une hypothèse scientifique. Le terme d'« évidence » est ici à prendre dans son sens anglais d'indice empirique.

Soit H une proposition variable (désignant ici une hypothèse) et E une proposition fixe (désignant un résultat expérimental), on a

$$\frac{P(H|E)}{P(H)} \textit{ proportionnel à } P(E|H)$$

Ici, $P(H)$ désigne la probabilité *a priori* de l'hypothèse et $P(H|E)$ désigne sa probabilité *a posteriori*, après enregistrement du résultat expérimental E. Appelons $P(E|H)$ la *vraisemblance* de l'hypothèse H. Dans cette version, le théorème de Bayes énonce que le rapport entre les probabilités *a posteriori* et *a priori* d'une hypothèse est proportionnel à sa vraisemblance – c.-à-d. à la probabilité qu'un résultat expérimental la confirme.

Prenons le cas le plus simple du choix exclusif entre deux hypothèses, H et $\overline{H}$. Définissons les chances de H étant donné E, notées $O(H|E)$: $O(H|E) = P(H|E)/\{1 - P(H|E)\}$. On a

$$\frac{O(H|E)}{O(H)} = \frac{P(E|H)}{P(E|\overline{H})}$$

Le rapport $O(H|E)/O(H)$ représente le rapport entre les vraisemblances de H et de $\overline{H}$ étant donné E. Comme nous n'avons que deux hypothèses couvrant l'ensemble des possibilités, ce rapport est donc le facteur par lequel multiplier les chances initiales de l'hypothèse H (c.-à-d. $O(H)$) pour en obtenir les chances finales (c.-à-d. $O(H|E)$). Ce facteur représente le gain en plausibilité de l'hypothèse H après l'enregistrement du résultat expérimental E. Turing s'inspira de l'unité de mesure acoustique *bel* pour proposer une unité de mesure de ce facteur : le *ban*. Le bel est le logarithme en base 10 du rapport entre deux intensités sonores. De même, si f désigne le rapport entre les chances *a priori* et *a posteriori* d'une hypothèse, alors on dira que cette hypothèse a gagné $\log_{10} f$ bans, soit $10 \log_{10} f$ décibans. Cette unité de mesure permet de quantifier le gain en « poids d'évidence » ou en plausibilité d'une hypothèse après une expérience.

La théorie de l'information

Les deux notions formellement définies par Claude Shannon sont celles d'information et d'entropie (ou information moyenne). Contrairement à ce que laisse entendre le sens ordinaire du mot « information », le concept défini par Shannon est moins en rapport avec ce qui est dit qu'avec ce qui *pourrait* être dit. Dans une situation élémentaire où le choix est binaire, l'information est d'une unité. Dans le cas où le choix s'effectue entre quatre possibilités et en supposant que nous adoptons un système d'écriture binaire, l'information est de deux unités, car il suffit d'un nombre de deux chiffres pour décrire en binaire le choix effectué. Dans le cas où le choix s'effectue entre n possibilités, la quantité d'information est donnée par le logarithme en base deux de n. Cette évaluation de la quantité d'information en fonction du logarithme du nombre de possibilités parmi lesquelles un message est choisi peut d'abord sembler étrange. Mais supposons que nous ayons quatre relais électriques, chacun étant dans seulement deux positions possibles. Avec un seul relais, nous ne pouvons obtenir qu'un seul message : la quantité

d'information est donc un. Avec quatre relais, nous voudrions naturellement que la quantité d'information puisse être quadruplée. Et c'est bien ce qui arrive avec la définition logarithmique (en base deux) de l'information : pour 4 relais, le nombre de combinaisons possibles entre les relais est de 16, et $\log_2 16 = \log_2 2^4 = 4$. Remarquons qu'il est ambigü de parler de « contenu » en information d'un message, car l'information désigne plus la situation d'où est extrait le message que quelque chose que nous pourrions trouver à l'intérieur.

L'entropie introduite par Shannon est une mesure de l'incertitude de l'information. Soit une distribution X sur un ensemble fini défini par les probabilités élémentaires $p_1, p_2, \ldots, p_n$. L'entropie de cette distribution est

$$H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

Si les événements possibles de X sont $x_i, 1 \leq i \leq n$, on a

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

Le logarithme n'est pas défini si $p_i = 0$; l'entropie est donc définie par la somme sur toute les probabilités non nulles. D'autre part, on vérifie facilement que l'entropie est maximale lorsque X est équidistribuée, c.-à-d. lorsque $P(x_i) = 1/n$ pour tout i. Dans ce cas, l'entropie vaut

$$H(X) = -n \times \frac{1}{n} \log \frac{1}{n} = \log n$$

La notion que Shannon introduisit et qui intéressait spécifiquement ses recherches de cryptologie est celle d'entropie conditionnelle. C'est aussi là que ses réflexions rejoignent partiellement celles de Turing sur les poids d'évidence. L'entropie conditionnelle mesure l'incertitude de l'information associée à une distribution d'événements X lorsqu'on connaît la distribution d'événements Y. Pour tout événement élémentaire y de Y, les valeurs $P(x|y)$ définissent une distribution $X|y$ sur le domaine de X. L'entropie de cette distribution est

$$H(X|y) = -\sum_{x} P(x|y) \log_2 P(x|y)$$

L'entropie conditionnelle $H(X|Y)$ est la moyenne pondérée (avec les cœfficients $P(y)$) des entropies $H(X|y)$ sur tous les choix possibles de y. On a donc

$$H(X|Y) = -\sum_{y} \sum_{x} P(y) P(x|y) \log_2 P(x|y)$$

L'entropie conditionnelle mesure l'information sur X obtenue en moyenne lorsque l'événement suivant Y est réalisé. Cette entropie conditionnelle est ce qui permet à Shannon de définir la perfection d'un système cryptographique. Soit $\mathcal{P}$ un ensemble de textes clairs, $\mathcal{C}$ un ensemble de textes chiffrés et $\mathcal{K}$ un ensemble de clefs telles que $K : P \to C$. Un tel système cryptographique est parfait si l'entropie conditionnelle d'un message clair connaissant son message chiffré est égale à l'entropie du message clair toute seule : $H(P|C) = H(P)$. Autrement dit, il faut que la mesure de l'incertitude restante au sujet du texte clair soit égale à cette même incertitude quand on connaît le texte chiffré correspondant.

Lien entre la théorie de Turing et la théorie de Shannon

Marquons tout de suite les différences. Turing ne s'intéresse pas à la définition de la quantité d'information d'un message individuel. Il va directement au problème de la confirmation d'une hypothèse et à la mesure de l'information acquise entre ses probabilités initiales et finales, après enregistrement d'un résultat expérimental. Shannon propose une définition de cette quantité d'information, à partir de laquelle il définit l'entropie. Les autres différences sont historiques et contingentes. D'une part Turing a travaillé seul, il n'a pas mis sa théorie en rapport avec celles qui existaient déjà au sujet du théorème de Bayes, et sa démarche théorique est essentiellement destinée à rendre possible la mécanisation de sa technique des mots probables. A l'inverse, les travaux de Shannon sont plus académiques et moins informés de la réalité mécanique de la cryptographie qui lui est contemporaine.

Mais les deux théories ont de nombreux points en commun. D'abord, comme le fait déjà remarquer Good dans (Good, 1950, p. 75), les notions de poids d'évidence ont leur équivalent dans la théorie de l'information. Ensuite, elles sont toutes deux motivées par des besoins cryptologiques : les avancées de Turing sont surtout orientées vers la cryptanalyse tandis que celle de Shannon sont dirigées vers la cryptographie, mais il s'agit dans les deux cas de mettre en rapport des distributions apparemment aléatoires d'information avec la distribution ordinaire des lettres du langage écrit.

4.4 L'analyse des protocoles avec la logique BAN

Règles d'inférence (suite)

Conjoncaténation de croyances

$$\frac{P\ believes\ X \qquad P\ believes\ Y}{P\ believes\ (X,Y)}$$

$$\frac{P\ believes\ Q\ believes\ (X,Y)}{P\ believes\ Q\ believes\ X} \qquad \frac{P\ believes\ Q\ said\ (X,Y)}{P\ believes\ Q\ said\ X}$$

Nous reprenons ici le néologisme « conjoncaténation » introduit par Syverson dans (Syverson et Cervesato, 2000) : ce néologisme est justifié par le fait que les concaténations de messages sont traitées par cette règle comme des conjonctions de formules : les deux sont représentées par (X,Y) dans les règles ci-dessus.

Conjoncaténation de fraîcheur

$$\frac{P\ believes\ fresh(X)}{P\ believes\ fresh\ (X,Y)}$$

Cette règle stipule simplement que si P croit que X est frais, alors P croit que n'importe quel message contenant X est frais. Attention : la fraîcheur de (X,Y) ne nous dit en revanche rien de la fraîcheur de X ou de celle de Y.

Règles de réception : voir c'est recevoir

$$\frac{P \textit{ believes } P \overset{k}{\longleftrightarrow} Q \qquad P \textit{ received } \{X\}_{k^{-1}}}{P \textit{ received } X} \qquad \frac{P \textit{ received } (X,Y)}{P \textit{ received } X}$$

Cette règle dit qu'un agent recevant un message reçoit aussi les sous-messages qu'il peut déchiffrer. Elle ne permet pas de faire la distinction entre les messages que les agents *reçoivent* et ceux qu'ils *possèdent* ; la plupart des successeurs de BAN rendront cette distinction exprimable.

Analyse du protocole NSSK à clef secrète

Message 2 $\quad S \rightarrow A : \quad \{n_A, A \overset{k_{AB}}{\longleftrightarrow} B, fresh(k_{AB}), \{A \overset{k_{AB}}{\longleftrightarrow} B\}_{k_{BS}}\}_{k_{AS}} \; from \; S$

Message 3 $\quad A \rightarrow B : \quad \{ A \overset{k_{AB}}{\longleftrightarrow} B\}_{k_{BS}} \; from \; S$

Message 4 $\quad B \rightarrow A : \quad \{n_B, A \overset{k_{AB}}{\longleftrightarrow} B\}_{k_{AB}} \; from \; B$

Message 5 $\quad A \rightarrow B : \quad \{n_B, A \overset{k_{AB}}{\longleftrightarrow} B\}_{k_{AB}} \; from \; A$

Le premier message du protocole NSSK n'est pas pris en compte dans l'idéalisation : la logique BAN ne tient pas compte des parties non chiffrées des messages car ces parties n'ont aucune incidence sur les inférences dont seront capables les agents. La partie $from$ des messages s'explique par le fait que chaque agent est supposé reconnaître ses propres messages : si un agent reçoit un message chiffré avec une clef privée, il peut immédiatement dire de qui est le message. Dans cette version idéalisée du protocole, l'intérieur des messages est aussi reformulé : en particulier, la clef k_{AB} est remplacée par des assertions à son sujet. Ainsi, le message 2, idéalisé, est un message chiffré par le serveur pour Alice, message qui contient un *nonce*, l'affirmation de la fraîcheur de k_{AB}, l'affirmation de la validité de k_{AB} pour communiquer avec Bernard, et un message chiffré à faire suivre à Bernard. Notons que la forme idéalisée du message 5 contient seulement n_B et non $n_B - 1$: c'est que la soustraction a seulement pour rôle de différencier ce message du précédent, différence représentée dans la forme idéalisée du protocole par la différence entre les deux $from$ (BAN n'a de règle directe pour inférer la fraîcheur de $n_B - 1$ à partir de celle de n_B.)

Assomptions initiales dans le protocole NSSK

P1. $A \textit{ believes } A \overset{k_{AS}}{\longleftrightarrow} S$
P2. $B \textit{ believes } B \overset{k_{BS}}{\longleftrightarrow} S$
P3. $A \textit{ believes } S \textit{ controls } A \overset{k}{\longleftrightarrow} B$
P4. $B \textit{ believes } S \textit{ controls } A \overset{k}{\longleftrightarrow} B$
P5. $A \textit{ believes } S \textit{ controls } fresh(A \overset{k}{\longleftrightarrow} B)$
P6. $A \textit{ believes } fresh(n_A)$
P7. $B \textit{ believes } fresh(n_B)$

La plupart de ces assomptions s'expliquent d'elles-mêmes. P1 et P2 expriment la croyance en la qualité des clefs à long terme (on pourrait exprimer le même genre d'assomptions pour S, mais elles ne sont pas nécessaire dans l'exemple d'analyse que nous proposons.) P3, P4 et P5 expriment les croyances d'Alice et Bernard au sujet de ce que le serveur contrôle. P6 et P7 expriment la croyance de chaque agent en la fraîcheur des *nonces* qu'il génère.

Annotation du protocole NSSK

L'annotation fait état des assomptions directement issues du protocole dans sa forme idéalisée : les quatre propositions suivantes sont les traductions des quatre messages du protocole idéalisé.

P8. A *received* $\{n_A, A \overset{k_{AB}}{\longleftrightarrow} B, fresh(k_{AB}),\{ A \overset{k_{AB}}{\longleftrightarrow} B\}_{k_{BS}}\}_{k_{AS}}$

P9. B *received* $\{A \overset{k_{AB}}{\longleftrightarrow} B\}_{k_{BS}}$ *from* S

P10. A *received* $\{n_B, A \overset{k_{AB}}{\longleftrightarrow} B\}_{k_{AB}}$ *from* B

P11. B *received* $\{n_B - 1, A \overset{k_{AB}}{\longleftrightarrow} B\}_{k_{AB}}$ *from* A

Ici s'achève la formulation des assomptions nécessaires à l'analyse du protocole. Dans les dérivations qui suivent, chaque ligne est suivie d'une justification, c.-à-d. des règles qui ont servi à la dérivation et des prémisses et/ou formules dérivées qui y ont été utilisées.

1. A *believes* S *said* $(n_A, A \overset{k_{AB}}{\longleftrightarrow} B, fresh(A \overset{k_{AB}}{\longleftrightarrow} B), \{A \overset{k_{AB}}{\longleftrightarrow} B\}_{k_{BS}})$
   Par la règle de signification du message, en utilisant P1, P8.
2. A *believes* $fresh(n_A, A \overset{k_{AB}}{\longleftrightarrow} B, fresh(A \overset{k_{AB}}{\longleftrightarrow} B), \{A \overset{k_{AB}}{\longleftrightarrow} B\}_{k_{BS}})$
   Par la règle de conjoncaténation de fraîcheur, en utilisant 1, P6.
3. A *believes* S *believes* $(n_A, A \overset{k_{AB}}{\longleftrightarrow} B, fresh(A \overset{k_{AB}}{\longleftrightarrow} B), \{A \overset{k_{AB}}{\longleftrightarrow} B\}_{k_{BS}})$
   Par la règle de vérification de *nonce*, en utilisant 2, 1.
4. A *believes* S *believes* $(A \overset{k_{AB}}{\longleftrightarrow} B)$
   Par la règle de conjoncaténation de croyances, en utilisant 3.
5. A *believes* S *believes* $(fresh(A \overset{k_{AB}}{\longleftrightarrow} B))$
   Par la règle de conjoncaténation de croyances, en utilisant 3.
6. A *believes* $(A \overset{k_{AB}}{\longleftrightarrow} B)$
   Par la règle de juridiction, en utilisant 4, P3.
7. A *believes* $fresh(A \overset{k_{AB}}{\longleftrightarrow} B)$
   Par la règle de juridiction, en utilisant 4, P5.

Nous avons dérivé la croyance d'Alice en la validité et la fraîcheur de k_{AB}. Passons maintenant à Bernard.

8. B *believes* S *said* $A \overset{k_{AB}}{\longleftrightarrow} B$
   Par la règle de signification du message, en utilisant P2, P9.

Avec les assomptions que nous avons exprimées jusqu'à présent, c'est tout ce que nous pouvons dériver pour la croyance de Bernard concernant la validité de k_{AB}. A l'inverse d'Alice, Bernard n'a envoyé aucun *nonce* à ce moment du protocole; la seule façon d'aller plus loin est d'ajouter l'assomption selon laquelle Bernard croit en l'assertion « $fresh(A \overset{k_{AB}}{\longleftrightarrow} B)$ ». Ce que nous faisons ainsi :

P12. B *believes* $fresh(A \overset{k_{AB}}{\longleftrightarrow} B)$

Cette assomption est différente des assomptions précédentes faites sur la fraîcheur des messages, car celles-ci s'appuyaient sur des valeurs que les agents avaient eux-mêmes générées. Cette assomption exprime le fait que Bernard croit à la fraîcheur d'une valeur aléatoire qu'il n'a pas lui-même générée. Nous reviendrons à cette assomption étrange après avoir complété la dérivation.

9. *B believes S believes* $A \overset{k_{AB}}{\longleftrightarrow} B$
   Par la règle de vérification de *nonce*, en utilisant P12, 8.
10. *B believes* $A \overset{k_{AB}}{\longleftrightarrow} B$
    Par la règle de juridiction, en utilisant P4, 9.

Nous avons maintenant dérivé, pour Alice et Bernard, les croyances de premier ordre en la validité et la fraîcheur de k_{AB}. Nous dérivons maintenant leurs croyances de second ordre.

11. *A believes B said* $(n_B, A \overset{k_{AB}}{\longleftrightarrow} B)$
    Par la règle de signification du message, en utilisant 6, P10.
12. *A believes* $fress(n_B, A \overset{k_{AB}}{\longleftrightarrow} B)$
    Par la règle de conjoncaténation de fraîcheur, en utilisant 7.
13. *A believes B believes* $(n_B, A \overset{k_{AB}}{\longleftrightarrow} B)$
    Par la règle de vérification de *nonce*, en utilisant 12, 11.
14. *A believes B believes* $A \overset{k_{AB}}{\longleftrightarrow} B$
    Par la règle de conjoncaténation de croyances, en utilisant 13.

Par un raisonnement semblable, nous pourrions obtenir *A believes B believes* $A \overset{k_{AB}}{\longleftrightarrow} B$, mais avec une différence importante. Comme Bernard croit que $A \overset{k_{AB}}{\longleftrightarrow} B$ est frais (par P12), il n'y a pas besoin de recourir à la vérification de *nonces* : n_B et $n_B - 1$ n'ont pour rôle que de différencier les deux derniers messages, non de garantir la fraîcheur de la clef aux yeux de Bernard.

Analyse du protocole Nessett 1990

Protocole idéalisé

Message 1 $\quad A \rightarrow B : \{n_A, A \overset{k_{AB}}{\longleftrightarrow} B\}_{k_A^{-1}}$

Message 2 $\quad B \rightarrow A : \{A \overset{k_{AB}}{\longleftrightarrow} B\}_{k_{AB}}$

Annotation des prémisses

P1. *B received* $\{n_A, A \overset{k_{AB}}{\longleftrightarrow} B\}_{k_A^{-1}}$

P2. *A received* $\{A \overset{k_{AB}}{\longleftrightarrow} B\}_{k_{AB}}$

Assomptions initiales

P3. *B believes* $PK(k_A, A)$

P4. *A believes* $A \overset{k_{AB}}{\longleftrightarrow} B$

P5. *A believes* $fresh(A \overset{k_{AB}}{\longleftrightarrow} B)$

P6. *B believes* $fresh(n_A)$

P7. *B believes A controls* $(A \overset{k_{AB}}{\longleftrightarrow} B)$

Dérivations

1. *B believes A said* $(n_A, A \overset{k_{AB}}{\longleftrightarrow} B)$
   Par la règle de signification de message utilisant P3, P1.
2. *B believes* $fresh(n_A, A \overset{k_{AB}}{\longleftrightarrow} B)$
   Par la règle de conjoncaténation de fraîcheur utilisant P6.
3. *B believes A believes* $(n_A, A \overset{k_{AB}}{\longleftrightarrow} B)$
   Par la règle de vérification de *nonce* utilisant 2, 1.
4. *B believes A believes* $A \overset{k_{AB}}{\longleftrightarrow} B$
   Par la règle de conjoncaténation de croyances utilisant 3.
5. *B believes* $A \overset{k_{AB}}{\longleftrightarrow} B$
   Par la règle de juridiction utilisant P7, 4.
6. *A believes B said* $A \overset{k_{AB}}{\longleftrightarrow} B$
   Par la règle de signification de message utilisant P4, P2.
7. *A believes B believes* $A \overset{k_{AB}}{\longleftrightarrow} B$
   Par la règle de vérification de *nonce* utilisant P5, 6.

Ainsi, nous avons pu dériver tous les buts d'authentification. Ceci montre que, si l'on se fie à l'analyse BAN, k_{AB} est une bonne clef de session pour Alice et Bernard, ce qu'elle n'est évidemment pas du fait que tout le monde y a accès.

4.5 Le prix Lœbner

A la fin du siècle dernier est né un tournoi destiné à récompenser le programme informatique qui sera le plus apte à tromper un interrogateur humain. Le programme victorieux et son créateur se voient décerner le prix Lœbner, du nom de la personne qui finance cet événement. Inaugurée par Daniel Dennett en 1991, la compétition tient autant du folklore que de la recherche scientifique, mais les événements dont elle a été le théâtre témoignent des espoirs et déceptions liés aux différentes interprétations du test de Turing.

Précisons tout de suite que le test subi par les programmes lors de cet événement est un test de Turing triplement modifié. Premièrement, ce test ne comporte qu'un moment, alors que le jeu de l'imitation original en comporte deux : l'un où l'interrogateur est face à deux êtres humains, l'autre où l'un de ces deux êtres humains est remplacé par une machine. Deuxièmement, l'interrogateur ne fait face qu'à une entité – homme ou machine; il ne doit pas simultanément comparer entre elles les réponses de deux entités pour savoir laquelle est la plus crédible

lorsqu'elle revendique la possession de tel ou tel prédicat, mais s'intéresse seulement à la ressemblance globale d'une entité comme personne humaine. Troisièmement, les sujets de discussion auxquels les joueurs peuvent prendre part ont été restreints. Alors que le concept de machine universelle permettait à Turing d'envisager qu'on fasse d'une machine une simulatrice universelle, les limitations techniques auxquelles on fait face à la fin du vingtième siècle obligent à ne confronter l'interrogateur qu'à des « experts ». Ces trois modifications nous font perdre trois dimensions du jeu de l'imitation : nous n'avons plus de comparaison entre la capacité humaine de simulation et son équivalent dans la machine, plus de comparaison en temps réel, plus de comparaison potentiellement universelle.

Quels ont été les résultats de cette compétition[47] ? En 1991, six machines et deux êtres humains (dont une femme) faisaient face aux huit interrogateurs. Ceux-ci disposaient d'une échelle de 1 à 8 pour évaluer l'« humanité » de leur interlocuteur, une note au-dessus de cinq signifiant que l'interlocuteur a eu l'impression de s'être adressé à un être humain. La femme fut jugée la plus humaine, mais fut aussi classée comme « machine » par l'un des interrogateurs. L'autre humain fut classé deux fois comme « machine » par certains juges qui avaient du mal à admettre qu'un être humain puisse en savoir autant sur Shakespeare.

Nous sommes à la fois loin des pronostics optimistes de Turing et loin des conditions originales de son test. L'intérêt de ce genre d'événements est non seulement de susciter de nouvelles recherches quant aux moyens de produire la ressemblance entre une machine et un être humain, mais aussi de mettre en évidence la possibilité, pour un interrogateur, de se tromper. Le problème est que, comme les programmes les plus efficaces à ce jour sont des programmes reposant pour la plupart sur des astuces linguistiques, nous ne pouvons raisonnablement attribuer ces erreurs de l'interrogateur qu'à sa promptitude à projeter des états mentaux sur un agent quelconque. C'est alors l'image du test de Turing qui évolue négativement : nous ne le voyons plus comme un moyen de prouver qu'une machine est intelligente, mais comme un moyen de trouver des emballements douteux dans notre machine à attribuer des états mentaux.

4.6 L'erreur de la décontextualisation

Dans ce qui suit, nous tentons de voir ce qu'il y a de commun entre deux erreurs : l'erreur de catégorie au sujet des états mentaux et l'erreur que Perry nomme dans (Perry, 2001) le « *subject matter fallacy* ». Nous exposons l'un des arguments de Perry contre le dualisme et nous résumons l'argument de Ryle (déjà exposé 1.4.1 section page 15). Dans cette exposition des arguments de Ryle et de Perry, nous mettons en évidence le fait que les deux pièges dénoncés fonctionnent sur le même principe : la confusion entre une description contextuelle et une description décontextualisée, la tentation de se placer depuis deux points de vue à la fois. Cette erreur concerne particulièrement l'analyse de l'imitation et de l'authentification si nous considérons que tout jugement relatif à une imitation et tout jugement d'authentification est contextuel.

Subject matter fallacy

Le but de Perry est de soutenir une position physicaliste affirmant qu'il existe un rapport d'identité entre les états mentaux et les états cérébraux. Pour cela, il tente de montrer que la plupart des arguments dualistes commettent le même type d'erreur, celle qu'il nomme « *sub-*

ject matter fallacy ». Nous prendrons l'exemple de son analyse de l'expérience de pensée de Jackson, censée mettre en évidence le fait que les qualités phénoménales de la conscience sont irréductibles à une description physicaliste.

Voici l'expérience de pensée. Imaginons Mary, une jeune scientifique enfermée dans un monde en noir et blanc, ayant accès à toute la connaissance scientifique possible. Elle sait à quels états cérébraux correspond la perception de la couleur rouge[48]. Jackson nous demande ensuite d'imaginer que Mary sort de sa chambre et perçoit, pour la première fois, la couleur rouge. Il paraît évident qu'elle apprendra quelque chose de nouveau sur la perception de la couleur : plus précisément, elle saura l'*effet que ça fait* de percevoir du rouge. De ce qu'elle apprend quelque chose de nouveau qui n'était pas dans la connaissance parfaite que Mary avait des états cérébraux correspondant à la perception de la couleur rouge, Jackson en conclut – et les dualistes avec lui – que certains des aspects de la conscience phénoménale sont irréductibles à une description physicaliste.

Comme point de départ de son contre-argument, Perry remarque que l'expérience de pensée de Jackson repose sur une vieille énigme philosophique : comment un jugement d'identité peut-il être informatif? Comment la reconnaissance de ce que A est identique à B peut-elle m'apprendre quelque chose de plus que la reconnaissance du fait que A est identique à lui-même? Comment le fait de reconnaître la couleur rouge en la voyant pour la première fois peut-il apprendre quelque chose de plus à Mary que ce qu'elle sait déjà sur la couleur rouge (attendu qu'elle sait tout ce qu'il y a à savoir) et grâce à quoi elle a pu la reconnaître?

L'information acquise dans un jugement d'identité ne peut pas être dans le rapport d'identité, car ce rapport est le même quels que soient les termes de l'identité. Frege répondait à cette question en soutenant que l'informativité des jugements d'identité n'était pas à chercher dans l'objectivité du contenu jugeable, mais dans la manière dont le contenu jugeable était évalué (Benmakhlouf, 1997, p. 105–106) . Nous montrons maintenant comment Perry s'inspire de cette solution frégéenne pour répondre à Jackson.

Voici ce que sait Mary quand elle est encore dans la chambre :

(1) La stimulation des fibres C est la perception de la couleur rouge.

Voici ce qu'elle constate en sortant de la chambre :

(2) *Cette* perception actuelle est la perception de la couleur rouge.

D'après le contenu objectif exprimé dans ces deux propositions, nous devrions en conclure que :

(3) *Cette* perception actuelle est la stimulation des fibres C. rouge.

Les conditions de vérité du contenu objectif de ces trois propositions sont les mêmes. Où peut donc se cacher ce que Mary a appris?

La suite de l'argumentation de Perry consiste à montrer que pour (2), l'évaluation des conditions de vérités ne coïncide pas avec l'évaluation des conditions de vérité du seul contenu objectif. La présence de la référence contextuelle « *Cette* perception » oblige à considérer, en plus du contenu objectif de (2), ce que Perry nomme son contenu « réflexif ». Le contenu réflexif d'un énoncé est le contenu dont l'évaluation dépend des conditions d'énonciation.

Que nous dit le fait que Mary apprenne quelque chose de nouveau? De ce que Mary apprend quelque chose de nouveau, les dualistes concluent qu'il y a *quelque chose* de nouveau à

apprendre, que le contenu objectif de (2) est différent du contenu objectif de (1). Si le contenu objectif de (2) est différent de (1), alors nous ne pouvons pas inférer (3) à partir de (1) et de (2). Si cette inférence n'est pas valide, c'est qu'il y a dans la perception de la couleur rouge quelque chose à connaître que Mary ne pouvait pas connaître seulement grâce à (1). La supposition exprimée par « c'est qu'il y avait quelque chose [...] à connaître » est précisément ce que Perry nomme *Subject matter fallacy*. Pour Perry, le fait que Mary apprenne quelque chose en voyant pour la première fois la couleur rouge n'est pas à mettre au compte de ce qu'il y aurait *quelque chose* de nouveau à apprendre (seulement accessible à la conscience phénoménale), mais dans le fait qu'une même connaissance est exprimée à l'intérieur d'un énoncé dont l'évaluation exige que l'on tienne compte des conditions d'énonciation. Autrement dit, (1) et (2) diffèrent par leur contenu réflexif, non par leur contenu objectif, aussi avons-nous le droit de conclure (3). Ce qui nous donne l'illusion de l'illégitimité de cette conclusion, c'est l'erreur consistant à penser que le contexte influence les conditions de vérité d'un contenu objectif quand il ne fait que conditionner l'évaluation subjective de ce contenu.

Réification des dispositions

Nous avons déjà exposé l'erreur de catégorie : elle consiste à attendre d'un concept qu'il fasse le même travail qu'un concept appartenant à une autre catégorie. L'erreur que nous nommons la « réification des dispositions » est un cas particulier de l'erreur de catégorie. Elle consiste à attendre la même chose de l'évaluation d'une disposition et de l'évaluation d'une propriété observable ou non observable. Par exemple, lorsque nous voyons qu'une personne applique toute son attention à une tâche, croire que nous pouvons demander « où est son attention ? », c'est poser le genre de question que nous poserions à une propriété observable; et c'est en posant ce genre de questions que nous sommes amenés à croire que le fait d'être attentif est adéquatement décrit comme une propriété *cachée*.

Or les propriétés dispositionnelles ne sont pas de la même catégorie que les propriétés observables *et* non observables. Comme l'explique longuement Ryle, les jugements dispositionnels sont des jugements hypothétiques : juger que quelqu'un est attentif, ce n'est pas juger qu'il possède une propriété, observable ou non, c'est juger qu'il est susceptible d'agir d'une manière déterminée dans des conditions déterminées. Les jugements dispositionnels sont donc des jugements hypothétiques, et c'est en oubliant cet aspect que nous croyons pouvoir leur faire faire le même travail que des jugements sur des propriétés observables ou non observables, à savoir des jugements catégoriques. Ryle analyse longuement les manières dont cette erreur engendre la plupart de nos égarements quand nous tentons de définir les états mentaux comme événements secrets, ayant lieu sur une scène privée.

Ce qui fait la difficulté des dispositions, c'est que nous ne pouvons pas définir à l'avance l'ensemble des conditions dans lesquelles le fait d'être disposé d'une certaine manière entraîne une action déterminée. Le contexte intervient deux fois : une fois comme l'ensemble des conditions dans lesquelles la disposition donnera lieu à un comportement manifeste (c'est ce qui fait que nos évaluations des propriétés dispositionnelles restent hypothétiques); une autre fois comme le point de vue à partir duquel nous évaluons le contexte à prendre en compte pour notre évaluation des conditions hypothétiques liées à la manifestation de la disposition. Autrement dit, l'ensemble des hypothèses faites sur le contexte pertinent à prendre en compte pour la manifestation possible d'une disposition est lui-même choisi relativement au point de vue depuis lequel l'évaluation est faite. Ce qui se passe dans l'erreur de la réification des dispositions, c'est

non seulement que nous oublions ce que l'ensemble des conditions dans lesquelles la disposition doit se manifester est hypothétique, mais en plus nous faisons comme si nous pouvions un instant négliger la relativité de notre point de vue. En laissant de côté ce que l'évaluation des propriétés dispositionnelles doit aux conditions d'évaluation, à la relativité de notre point de vue, nous nous préparons la possibilité de connaître toutes les conditions dans lesquelles la disposition mène à un comportement particulier, ensemble de conditions qui, une fois oublié son caractère hypothétique, nous mène à remplacer les propriétés seulement dispositionnelles par des propriétés non obervables. En faisant comme si nous étions virtuellement des observateurs omniscients, nous postulons des *états* là où notre statut d'observateur limité ne nous donne à inférer que des dispositions.

4.7 Problèmes formels pour la réduction des dispositions à des propriétés observables

Les deux problèmes formels relatifs aux propriétés dispositionnelles concernent respectivement l'impossibilité et la difficulté de les réduire (i) à des conjonctions (finies ou infinies) de propriétés observables ou (ii) à des disjonctions (finies ou infinies) de propriétés observables.

Intuitivement, nous relions sans cesse des dispositions à des propriétés observables : la fragilité du verre nous semble directement liée à sa structure cristalline, la flexibilité d'une barre de métal nous semble directement liée au matériau dans lequel elle est faite. Mais il se peut parfaitement qu'un verre ait la même structure cristalline qu'un autre sans pour autant être fragile, ou qu'une barre soit faite dans le même métal qu'une autre sans pour autant être flexible. Aussi une conjonction de propriétés observables ne suffit-elle jamais à épuiser les situations dans lesquelles la fragilité du verre peut se manifester. Comme l'a montré Goodman dans (Goodman, 1984), les difficultés concernant la réduction des prédicats dispositionnels à des prédicats manifestes sont de même nature que les difficultés relatives à l'induction. Si le fait de fléchir une barre de fer confirme bien l'hypothèse de sa flexibilité, nous ne pouvons jamais épuiser par induction l'ensemble des confirmations possibles d'une hypothèse faite sur la possession d'un prédicat dispositionnel. Cela signifie qu'étant données une disposition et une conjonction de propriétés observables, la disposition ne pourra jamais être prouvée comme cœxtensive à cette conjonction de propriétés observables, ni *a fortiori* identique à celle-ci.

Peut-on se contenter, pour définir une propriété dispositionnelle, d'une disjonction de propriétés observables ? Soit ϕ une propriété dispositionnelle et P_i^α une propriété observable du monde actuel α. Peut-on poser l'égalité : $\phi = P_1^\alpha \vee P_2^\alpha \vee \ldots$? La fragilité est-elle cœxtensive avec l'ensemble des objets ayant telle structure cristalline *ou* ayant tel degré de transparence *ou* ... ? La réponse est non. Imaginons en effet qu'une certaine espèce d'entités K_1 dont certains individus sont ϕ en vertu de P_1^α, une autre espèce K_2 dont certains individus sont ϕ en vertu de P_2^α. Considérons maintenant qu'il peut tout à fait exister certains individus de K_1 ne possédant pas P_1^α mais possédant P_2^α ; comme ils ne possèdent pas la propriété P_1^α qui est à la base de la ϕ-ité des individus de K_1, ces individus de K_1 possédant seulement P_2^α ne seront pas ϕ. Or ils seraient ϕ si ϕ était définie comme la disjonction $P_1^\alpha \vee P_2^\alpha \vee \ldots$ Donc cette définition de ϕ est fausse. Soit un groupe d'hommes dont l'irritabilité de certains tient à la présence d'un dérèglement hormonal et soit un groupe de femmes dont l'irritabilité de certaines tient à leur position politique. Il se peut qu'un homme soit du même parti politique qu'une femme irritable

sans être irritable. Or il devrait être irritable si l'irritabilité était cœxtensive à la disjonction avoir-tel-dérèglement-hormonal $\vee$ être-de-tel-parti-politique.

Cet argument est convaincant, mais non définitif. On peut dire par exemple que l'irritabilité est cœxtensive à la propriété définie par (avoir-tel-dérèglement-hormonal & être-un-homme) $\vee$ (être-de-tel-parti-politique & être-une-femme). En définissant la propriété dispositionnelle comme une disjonction de conjonctions entre propriétés manifestes et espèce d'entités possédant cette propriété manifeste, on a : $\phi = K_1 \& P_1^\alpha \vee K_2 \& P_2^\alpha \vee \ldots$ Dans (Cohen, 2002), l'auteur montre de manière convaincante que cette approche permet de définir les termes dispositionnels dans le monde actuel comme dans l'ensemble des mondes possibles.

Notes

[1]Jean Lassègue situe l'apparition de l'expression « test de Turing » au milieu des années 70' : voir (Lassègue, 1996b).

[2]Nous pouvons aussi formuler la fiabilité d'un test en nous servant de la définition que propose Goodman pour la projectibilité : un test sera fiable si et seulement si l'hypothèse de son adéquation est projectible. Cf. section 2.4.1 page 35.

[3]Nous reprenons ici les quatre objections telles qu'elles sont présentées en (Copeland, 1993, p. 44–50).

[4]Nous n'assumons pas les engagements douteux de Copeland quand à la pensée des animaux, mais la portée de l'objection reste la même.

[5]« Partition » doit bien sûr être entendu au sens large : tout support recevant l'inscription des signes dénotant une exécution déterminée est une partition.

[6]La définition de l'authenticité dans le cas des arts allographiques peut-être doublement nuancée : d'une part, une seule fausse note ne suffit pas à dire que l'orchestre joue une *autre* œuvre ; d'autre part, une fugue de Bach jouée sur un clavecin d'origine sera souvent jugée plus authentique. Mais ces problèmes internes à la théorie de Goodman n'ont pas d'incidence sur l'usage que nous faisons de sa terminologie.

[7]De même que toutes les notations ne se valent pas lorsqu'il s'agit de définir un art allographique, toutes les descriptions de protéine ne se valent pas lorsqu'il faut définir leurs propriétés nécessaires et suffisantes ; nous pouvons néanmoins supposer que le système descriptif scientifique est un système assez adéquat pour garantir l'égale conformité de la protéine naturelle et de la protéine artificielle à un même concept scientifique de protéine. Pour un aperçu de l'importance de ces problèmes d'étiquetage, voir (Atlan, 2002, p. 83).

[8]Nous devons à Alexandre Viros de nous avoir mis sur la voie de la distinction entre traits prototypiques et traits constitutifs.

[9]Notons que les difficultés (i) et (iii) sont liées sans se confondre. En effet, si nous parvenons à dresser une liste des propriétés constitutives d'un état mental donné, il y a des chances pour que nous puissions dire quel est le rapport de cet état mental à ses conditions de production. Si nous savons, par exemple, qu'un état mental est identique à une disjonction de propriétés cérébrales, alors l'élucidation de la manière dont un individu accède à cet état mental n'est plus que l'élucidation de la manière dont l'un des états cérébraux corrélés est produit. Inversement, si nous montrons qu'une intention donnée n'est produite qu'en présence de tel ou tel objet de l'environnement, alors cette spécification des conditions de production de l'intention pourra nous guider dans la définition de ses propriétés constitutives. Mais les deux problèmes ne sont pas confondus. D'une part il est tout à fait concevable que certaines propriétés d'un état mental soient *indépendantes* de ses conditions de production : dans ce cas, la référence à ses conditions de production n'est pas pertinente pour son authentification. D'autre part, il est concevable qu'une propriété constitutive d'un état mental soit absolument *dépendante* de ses conditions de production, que l'ensemble de ces conditions soit descriptible de manière finie ou non. Ce serait le cas des *qualia* qui, en tant que propriétés « intrinsèques » de la conscience phénoménale ne sont pas, par définition, déterminables hors de leur conditions subjectives de production.

[10]La tristesse est bien une propriété exprimée par le tableau : celui-ci peut exemplifier des propriétés de manière plus ou moins littérale, mais ce seront toujours *ses* propriétés.

[11]Nous ferons la distinction entre imitation-processus et imitation-état pour éviter certaines ambiguités dans l'usage du terme « imitation ». Dans la mesure du possible, nous parlerons de « simulation » pour désigner l'imitation-processus et nous réserverons le mot « imitation » pour désigner l'imitation-état.

[12]Remarquons à nouveau que toute imitation n'a pas pour fonction de tromper : l'assignation ou non de cette fonction à l'imitation est une conséquence de la manière dont on s'en sert, non de sa définition. L'intention à laquelle nous nous référons lorsque nous jugeons qu'une reproduction est une imitation ne coincide pas avec

l'intention de se servir de cette imitation comme leurre.

[13]Le mot « disposition » récouvre ici un ensemble de phénomènes aussi distincts que l'habitude, l'instinct, les penchants, les inclinations, la propension, etc.

[14]Nous reprenons l'exemple donné par P. Jacob dans (Andler, 1992, p. 324).

[15]Loin de nous l'idée de suggérer qu'il existe des comportements typiquement féminin ; nous ne faisons que nous placer dans l'imagination de l'interrogateur, lequel aura besoin de recourir à ce genre de schématisation pour faire des hypothèses sur l'identité sexuelle de son interlocuteur.

[16]Dans le cas d'un tableau, on peut difficilement douter de l'existence passée d'un tel expert : le peintre lui-même peut jouer ce rôle.

[17]Putnam discute successivement ces trois points pour combattre les positions de Fodor et de Searle et pour insister sur la nature nécessairement *sociale* de la référence. Il refuse donc l'hypothèse fodorienne du mentalais et donne des arguments contre les présupposés que Searle accepte, c.-à-d. (1) et (3).

[18]Peirce employait de manière très libérale la notion de « machine », justifiant souvent l'analogie entre l'homme et la machine pour ce qui était du raisonnement déductif. Mais, portant de plus en plus d'intérêt aux procédés inductifs et abductifs, et reconnaissant de plus en plus l'importance des processus iconiques dans la pensée, il en vînt à souvent critiquer cette analogie. Voir (Tiercelin, 1993).

[19]Pour la définition des notions d'information et de *ban*, et pour le rapport entre ces deux définitions, voir la section 4.3 page 65 de l'annexe.

[20]Ce passage est aussi cité et commenté utilement dans (Hodges, 1983, p. 383).

[21]Ce que Turing entend ici par « *cryptography* » correspond en fait à ce que nous appelons plus strictement « cryptanalyse », comme le montre la citation suivante.

[22]Nous traduisons. La version originale : « The fiel of cryptography will perhaps be the most rewarding. There is a remarkably close parallel between the problems of the physicist and those of the cryptographer. The system on which a message is enciphered corresponds to the laws of the universe, the intercepted messages to the evidence available, the keys for a day or a message to important constants which have to be determined. »

[23]Nous ne prétendons pas qu'en construisant son jeu de l'imitation, Turing avait en tête un schéma d'un béhaviorisme aussi rudimentaire , mais nous prétendons que ce schéma permet de mieux comprendre l'usage que fait Turing de la notion de « surprise » lorsqu'il répond à l'objection de Lady Lovelace ainsi que l'idée d'introduire un élément de hasard dans les machines participant au jeu de l'imitation.

[24]Cette incertitude est mesurée par l'entropie conditionnelle de M sachant C. Voir l'annexe, section 4.3 page 66.

[25]Cette définition probabiliste de la *confidentialité* n'interfère en rien avec une définition phénoménologique du caractère *intrinsèque* d'un état mental.

[26]L'adage que Turing mentionne : « Il n'y a rien de nouveau sous le soleil. » (Anderson, 1964, p. 57) sonne plutôt comme une boutade que comme un argument.

[27]Soit C l'ensemble des circonstances dont Turing suppose réalisées (« Je suppose que le voltage ici devrait être le même que là. ») et K l'ensemble des circonstances effectivement réalisées par la machine pendant le calcul ; soit E_t l'énoncé du résultat d'après Turing et E_m l'énoncé du résultat par la machine. Le fait que E_m soit surprenant ne correspond pas seulement au fait que, pour Turing, E_m était moins probable que E_t, mais correspond plus précisément au fait que $P(C|E_m) \ll P(C)$. Pour Turing, E_m était très peu probable au regard de l'ensemble de circonstances C (qu'il tient pour très probable), mais oubliant que cet ensemble de circonstances est *seulement* probable, il est incapable de prendre en considération l'ensemble de circonstances K au regard duquel E_m est très probable (et E_t peu probable). On a donc bien $P(K)P(E_m|K) \gg P(C)P(E_m|C)$.

[28]Le seul point de vue à partir duquel cette évaluation est possible peut difficilement être encore appelé un *point* de vue.

[29]Les guillemets sont laissés volontairement par Turing pour montrer qu'il ne souscrit à aucune interprétation dualiste de ces actes de création mentale.

[30]Turing se garde ici aussi de souscrire à l'usage de cette expression.

[31]Il est intéressant de noter que Peirce avait proposé plusieurs approches pour définir la hasard. Sa définition la plus satisfaisante est la définition récursive de « distribution fortuite », qui est remarquable si l'on en juge par le peu d'outils formels qu'il avait pour l'établir. Putnam discute de ce point dans (Peirce, 1995, p. 87–94).

[32]La singularité de π est que la suite de ses décimales résiste parfaitement à tous les tests d'aléatoirité, quand bien même π est finiment définissable et parfaitement déterminé. La calculabilité d'un nombre fini de décimales de π était déjà bien connue de Turing, lequel avait trouvé à l'âge de quinze ans une suite permettant d'approximer π. Dans le passage que nous commentons, Turing se sert justement de cette particularité de π pour opposer ce qui nous paraît aléatoire et ce qui l'est effectivement. A noter que cette singularité est aussi exploitée dans de nombreux cryptosystèmes.

[33]De plus, une hypothèse est dite *improjectible* si elle est supplanté, et *non projectible* s'il est impossible de savoir si elle est supplantée par une autre hypothèse mieux implantée. Mais nous n'utiliserons pas cette distinction.

[34]L'idée d'aborder le test selon une perspective diachronique est déjà présente dans l'article de Putnam en (Anderson, 1964, p. 110–134).

[35]Imaginons par exemple que l'hypothèse fausse H soit : ce dé a deux faces « pile », et que l'expérience E soit le lancé du dé. Au yeux de A (qui sait que le dé n'est *pas* pipé), il y a bien une chance sur deux pour que pile tombe. Si pile tombe, l'hypothèse fausse de B est confirmée aux yeux de B avec un facteur de 2 ; en effet, $P(E|H) = 1$, $P(E|\overline{H}) = 1/2$ donc le facteur $\frac{P(E|H)}{P(E|\overline{H})}$ vaut deux. Si face tombe, l'hypothèse fausse est infirmée (ou confirmée avec un facteur de zéro). Donc, aux yeux de A, il y a une probabilité de $2 \times 1/2$ (soit 1) que A voit son hypothèse fausse confirmée.

[36]Terme intraduisible pour désigner la « lecture » des états mentaux d'autrui.

[37]Les « agents » peuvent être aussi bien des ordinateurs, que des services ou des processus sur un même ordinateur.

[38]Le fait que les aspects proprement cryptographiques du protocole soient considérés comme une « boîte noire » peut se révéler dangereux quant à l'implémentation du protocole : voir (Clark, 1996).

[39]« BAN » vient du nom des trois auteurs à l'origine de cette logique : Burrow, Abadi, Needham. Pourquoi le choix d'exposer *cette* logique ? Pour deux raisons : historiquement, cette logique est à l'orgine de très nombreuses recherches dans le champ de l'analyse logique des protocoles d'authentification : toutes les réflexions actuelles sur les logiques de l'authentification se réfèrent à cette logique ; conceptuellement, cette logique pose clairement les bases de toutes les réflexions possibles sur le concept d'authentification dans le cadre d'une analyse inférentielle de la communication.

[40]Nous n'emploierons pas la notation originaire de BAN (Burrows et al., 1989), mais celle introduite par Abadi et Tuttle dans (Abadi et Tuttle, 1991).

[41]Dans le cas des clefs symétriques, il n'y a pas de distinction explicite entre signer et chiffrer. Dans la logique BAN, il n'y a pas non plus de distinction dans le cas des clefs publiques : la signature et le chiffrement sont tous les deux représentés par $\{X\}_k$. La distinction est implicite dans la notation pour la clef utilisée : k ou k^{-1}. Voici la version clef-publique de la règle de signification du message :

$$\frac{P \textit{ believes } PK(Q,k) \qquad P \textit{ received } \{X\}_{k^{-1}}}{P \textit{ believes } Q \textit{ said } X}$$

[42]Cette dernière n'a pas tellement de sens dans le cas où un agent vient de « dire » un *nonce*, mais le but de cette logique est de clarifier l'analyse des protocoles d'authentification, non de formaliser la croyance ordinaire.

[43]Dans la littérature sur les protocoles d'authentification, il semble y avoir un large consensus pour admettre que les protocoles doivent au moins assurer l'attribution de la responsabilité. On n'en demande souvent pas plus lorsqu'il s'agit de connexions à distance. La propriété fondamentale d'un canal sécurisé ($\mathcal{C}$) est de « parler pour » une entité. Si $\mathcal{C}$ « parle pour » A et si $\mathcal{C}$ dit S, alors A dit S. Ainsi, « A dit S » énonce seulement la propriété qu'a A de pouvoir être tenu pour responsable d'un message, non d'en être à l'origine. En revanche, il ne semble pas y avoir de consensus pour savoir si le but d'un protocole d'authentification doit aussi être d'assurer une attribution correcte de l'origine des messages. Avec le problème de l'attribution de l'origine (et la question de savoir si cette attribution doit faire partie des objectifs de l'authentification), nous retrouvons ce qui fait la singularité du but n° 4 décrit par Gollmann dans (Gollmann, 1996).

[44]Cinq groupes, d'après la classification de Becker et Piper.

[45]Notion introduite par Wolfe Friedman en 1920.

[46]Le cœfficient binomial $\binom{n}{k} = n!/(k!(n-k)!)$ représente le nombre de sous-ensembles de k éléments dans un ensemble n.

[47]Pour un rapport plus détaillé et une discussion rapide de la manière dont ces programmes ont été écrits, voir (Anceau, 1999, p. 67–73).

[48]Il n'y a aucun problème à ce niveau : on peut supposer que la perception de la couleur rouge correspond à *un* état cérébral, à une conjonction finie d'états cérébraux, à une disjonction finie d'états cérébraux, etc. Dans tous les cas, Mary sait *tout* ce qu'il y a à savoir sur la perception de la couleur rouge.

Remerciements

Nous tenons ici à remercier Daniel Andler, Ali Benmakhlouf pour ses encouragements, Pierre Bieber pour l'envoi de sa thèse, l'association Cognivence pour son environnement stimulant, Jean-Gabriel Ganascia pour sa disponibilité, Jean Lassègue pour ses désaccords, Catherine Le Forestier et Nicole Morain pour leur accueil, Jérôme Ségal pour ses conseils et Alexandre Viros pour ses corrections. Merci aussi à Emilie pour sa patience et à Julie Neveux pour tout le reste.

Bibliographie

Abadi, M. (2000). « Security protocols and their properties ».

Abadi, M. et Tuttle, M. R. (1991). « A Semantics for a Logic of Authentication ». Dans Logrippo, L., éditeur, *10th Annual ACM Symposium on Principles of Distributed Computing*, pages 201–216, Montréal, Québec, Canada. ACM Press.

Anceau, F. (1999). *Vers une étude objective de la conscience.* Hermès Science Publications, Paris.

Anderson, A. R., éditeur (1964). *Pensée et machine.* Editions du Champ Vallon, 1983, Seyssel. Tr. fr. Patrice Blanchard.

Andler, D., éditeur (1992). *Introduction aux sciences cognitives.* Editions Gallimard.

Andler, D. (1999). *Science et philosophie.* Bibliothèque du CREA. CREA.

Atlan, H. (2002). *La science est-elle inhumaine?* Bayard Editions, Paris.

Bacharach, M. et Gambetta, D. (1997). « Trust in Signs ». http:// www. economics. ox.ac.uk/ Research/ Breb/ TSI/ trust.pdf.

Bacharach, M. et Gambetta, D. (1999). « Trust as Type Detection ». http:// www. economics. ox.ac.uk/ Research/ Breb/ TSI/ trustype.pdf.

Belna, J.-P. (2000). *Cantor.* Les Belles Lettres, Paris.

Benmakhlouf, A. (1997). *Gottlob Frege: logicien philosophe.* Presses Universitaires de France, Paris.

Bieber, P. (1989). « *Aspects épistémiques des protocoles cryptographiques* ». Informatique, Université Paul Sabatier, Toulouse.

Brassard, G. (1993). *Cryptologie contemporaine.* Masson, Paris.

Burrows, M., Abadi, M., et Needham, R. (1989). « A Logic of Authentication ». Rapport Technique 39, Digital Equipment Corporation, Systems Research Centre.

Burrows, M., Abadi, M., et Needham, R. (1990a). « A Logic of Authentication ». *ACM Transactions on Computer Systems*, 8(1):18–36.

Burrows, M., Abadi, M., et Needham, R. (1990b). « The scope of a logic of authentication ». Dans Feigenbaum, J. et Merritt, M., éditeurs, *Distributed Computing and Cryptography*, volume 2 de *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 119–126. AMS and ACM. Proceedings of a DIMACS workshop, October 1989.

Cassou-Noguès, P. (2001). *Hilbert.* Les Belles Lettres, Paris.

Casti, J. L. (1998). *The Cambridge Quintet.* Abacus, London.

Clark, J. et Jacob, J. (1997). « A Survey of Authentication Protocol Literature ». http:// www-users.cs. york. ac.uk/ jac/.

CLARK, J. A. (1996). « Attacking Authentication Protocols ». http://www-users.cs.york.ac.uk/ jac/.

COHEN, J. (2002). « On An Alleged Non-Equivalence Between Dispositions And Disjunctive Properties ». *The British Journal for the Philosophy of Science*, (53):77–81.

COLLINS, A. et SMITH, E. E. (1988). *Readings in Cognitive Science: A Perspective from Psychologie and Artificial Intelligence*. Morgan Kaufmann Publishers, Inc., San Mateo, California.

COPELAND, J. (1993). *Artificial Intelligence: a philosophical introduction*. Blackwell Publishers Inc., Malden, Massachusetts, USA.

DELAHAYE, J.-P. (1997). « *Le fascinant nombre* π », Partie 9, π est-il aléatoire?, pages 169–192. Pour la Science, Paris. Diffusion Belin.

DELAHAYE, J.-P. (1999). *Information, complexité, hasard*. HERMES Science Publications, Paris.

DENNETT, D. (1991a). « Two Black Boxes ». http://www.u.arizona.edu/ chalmers/biblio.html.

DENNETT, D. (1992). « Verbal Language... ». http:// www.u. arizona.edu/ chalmers/ biblio.html.

DENNETT, D. (1994). « Consciousness in Human and Robots Minds ». http:// www.u. arizona.edu/ chalmers/ biblio.html.

DENNETT, D. C. (1990). *La stratégie de l'interprète*. Editions Gallimard, Paris.

DENNETT, D. C. (1991b). *Consciousness Explained*. Little, Brown and Company.

DENNETT, D. C. (1993). *La conscience expliquée*. Editions Odile Jacob, Paris. Tr. fr. Pascal Engel.

DOLEV, D. et YAO, A. C. (1983). « On the security of public-key protocols ». *IEEE Transaction on Information Theory*, 2(29):198–208.

DREYFUS, H. L. (1984). *Intelligence artificielle: mythes et limites*. Editions Flammarion, Paris.

DUPUY, J.-P. « Philosophical Foundations of a new Concept of Equilibrium in the Social Sciences: Projected Equilibrium ». Distribué dans un cours du DEA de sciences cognitives 2002.

DUPUY, J.-P. « Sciences cognitives et sciences sociales: limites de la rationalité et nature du lien social ». Distribué dans un cours du DEA de sciences cognitives 2002.

DUPUY, J.-P. (1999). *Aux origines des sciences cognitives*. Editions La découverte, Paris.

ENGEL, P. (1989). « Sauver la croyance ». *Philosophie*, (24):72–94.

FAGIN, R., HALPERN, J. Y., et VARDI, M. Y. (1986). « What Can Machines Know? ». Dans *Proceedings AAAI-86*, Philadelphia.

FISETTE, D. et POIRIER, P. (2000). *Philosophie de l'esprit*. Librairie philosophique J. VRIN, Paris.

FODOR, J. A. (1986). *La modularité de l'esprit*. Editions de Minuit, Paris.

GABBAY, D., HOGGER, C. J., et ROBINSON, J. A., éditeurs (1995). « *Handbook of Logic in Artificial Intelligence and Logic Programming* », volume 4, Partie Reasoning About Knowledge: A Survey, pages 1–34. Oxford University Press.

GANASCIA, J.-G. (1990). *L'âme-machine, les enjeux de l'intelligence artificielle*. Editions du Seuil, Paris.

GOLLMANN, D. (1996). « What do we mean by an entity authentication? ». Dans *Proceeding the IEEE Computer Society Symposium on Research in Security and Privacy*, pages 46–54. IEEE Computer Society Press.

GONG, L., NEEDHAM, R., et YAHALOM, R. (1990). « Reasoning about Belief in Cryptographic Protocols ». Dans *Proceeding the IEEE Computer Society Symposium on Research in Security and Privacy*, pages 234–248. IEEE Computer Society Press.

GOOD, I. (1950). *Probability and the weight of evidence*. Charles Griffin & Co. Ltd., London.

GOODMAN, N. (1968). *Langages de l'art*. Editions Jacqueline Chambon, 1990, Nîmes. Tr. fr. Jacques Morizot.

GOODMAN, N. (1978). *Maniéres de faire des mondes*. Editions Jacqueline Chambon, 1992. Tr. fr. Marie-Dominique Popelard.

GOODMAN, N. (1984). *Faits, fictions et prédictions*. Editions de Minuit, Paris. Tr. fr. Martin Abran, revue par Pierre Jacob.

HALPERN, J. Y. et MOSES, Y. (1990). « Knowledge and Common Knowledge in a Distributed Environment ». *Journal of the ACM*, 37(3):549–587.

HARDY, G. H. (1985). *Apologie d'un mathématicien*. Editions Belin, Paris.

HEMPEL, C. (1966). *Eléments d'épistémologie*. Editions Armand Colin, 1996, Paris.

HODGES, A. (1983). *Alan Turing: the enigma*. Vintage.

HODGES, A. (1988). *Alan Turing ou l'énigme de l'intelligence*. Editions Payot.

HOFSTADTER, D. et DENNETT, D. C. (1987). *Vues de l'esprit*. InterEditions.

International Organization for Standardization (1991). « Information technology – Security techniques – Entity authentication mechanisms; Part 1: General model. ». ISO/IEC 9798-1, Second Edition.

JACOB, P., éditeur (1980). *De Vienne à Cambridge*. Editions Gallimard, Paris.

JEFFREY, R. (2002). « Subjective probability ». http:// www.princeton.edu/ bayesway/ Book*.pdf. En construction.

KÖHLER, W. (1964). *Psychologie de la forme*. Editions Gallimard, Paris.

LA METTRIE (1981). *L'homme machine*. Editions Denoël, Paris.

LASSÈGUE, J. (1996a). « La méthode expérimentale, la modélisation informatique et l'intelligence artificielle ». *Intellectica*, (22):21–65.

LASSÈGUE, J. (1996b). « What Kind of Turing Test did Turing Have in Mind? ». *Tekhnema*, (3).

LASSÈGUE, J. (1998). *Turing*. Edition Les Belles Lettres, Paris.

LOWE, G. (1997). « A hierarchy of authentication specification ». Dans *Proceedings of the 10th IEEE Computer Security Foundations Workshop (CSFW9)*, pages 31–43. IEEE Computer Society Press.

MACKAY, D. (2002). « *Information Theory, Inference and Learning Algorithms* », Partie 19, Units of information content. http:// www.inference. phy.cam.ac.uk/ mackay/ itprnn/ ps/ 291.317.pdf.

MACKAY, D. M. (1951). « Mindlike Behavior in Artefacts ». *The British Journal for the Philosophy of Science*, (2).

MEADOWS, C. (2000). « Open issues in formal methods for cryptographic protocol analysis ». Dans *DISCEX 2000: Proceedings of the DARPA Information Survivability Conference and Exposition*, volume 1, pages 237–250. IEEE Computer Society Press.

MELTZER, B. et MICHIE, D., éditeurs (1969). *Machine Intelligence*. Edinburgh University Press, Edinburgh. Volume 5 de National Physical Laboratory Report.

NAGEL, E., NEWMAN, J. R., GÖDEL, K., et GIRARD, J.-Y. (1989). *Le théorème de Gödel.* Editions du Seuil, Paris.

NAGEL, T. (1974). « What is it like to be a bat? ». *The Philosophical Review*, (83):435–450.

NESSETT, D. (1990). « A critique of the Burrows, Abadi and Needham logic ». *ACM Operating Systems Review*, 24(2):35–38.

NEUMAN, J. V. (1951). *Théorie générale et logique des automates.* Editions Champ Vallon, 1996, Seyssel.

NEUMAN, J. V. (1996). *L'ordinateur et le cerveau.* Flammarion, Paris.

OLSON, E. J. (2002). « Corroborating Testimony, Probability and Surprise ». *The British Journal for the Philosophy of Science*, (53):273–288.

PEIRCE, C. S. (1995). *Le raisonnement et la logique des choses.* Editions du Cerf. Tr. fr. Christiane Chauviré, Pierre Thibaud et Claudine Tiercelin.

PÉLISSIER, A. et TÊTE, A., éditeurs (1995). *Sciences Cognitives: textes fondateurs (1943-1950).* Presses Universitaires de France, Paris.

PENROSE, R. (1989). *L'esprit, l'ordinateur et les lois de la physique.* InterEditions, 1992, Paris. Tr. fr. Françoise Balibar et Claudine Tiercelin.

PERRY, J. (2001). *Knowledge, Possibility, and Consciousness.* The 1999 Jean Nicod Lectures. The MIT Press, Cambridge, Massachusetts; London, England.

PIATELLI-PALMARINI, M., éditeur (1979). *Théories du langage, théories de l'apprentissage.* Editions du Seuil, Paris.

POINCARÉ, H. (1902). *La science et l'hypothèse.* Flammarion, Paris. Réédition de 1968.

POINCARÉ, H. (1907). *La valeur de la science.* Flammarion, Paris. Réédition de 1970.

POUIVET, R., éditeur (1992). *Lire Goodman.* Editions de l'éclat, Paris.

PUTNAM, H. (1984). *Raison Vérité et Histoire.* Editions de Minuit, Paris.

PUTNAM, H. (1988). *Représentation et réalité.* Editions Gallimard, 1990, Paris. Tr. fr. Claudine Engel-Tiercelin.

RYLE, G. (1949). *The Concept of Mind.* Penguin Books, 1963. Introduction de Daniel C. Dennett, 2000.

SEGAL, J. (1998). « *Théorie de l'information: sciences, techniques et société de la seconde guerre mondiale à l'aube du XXIème siècle* ». PhD thesis, Université Lumière Lyon 2.

SHANNON, C. (1948). « A mathematical theory of communication ». *Bell system technical journal*, (27):379–423.

SHANNON, C. (1950). « Programming a computer for playing chess ». *Philosophical Magazine*, (43):256–275.

SHANNON, C. E. et WEAVER, W. (1963). *The Mathematical Theory of Communication.* Illini Books Edition.

SIMON, H. A. (1999). « L'explication en termes de traitement de l'information des phénomènes de Gestalt ». *Intellectica*, 1(28):115–137.

SIMONDON, G. (1958). *Du mode d'existence des objets techniques.* Aubier, Editions Montaigne, Paris.

SINGH, S. (1999). *Histoire des codes secrets.* Editions Jean-Claude Lattès, Paris. Tr. fr. Catherine Coqueret.

SPERBER, D. (1996). *La contagion des idées.* Editions Odile Jacob, Paris.

SPERBER, D., éditeur (2000). *Metarepresentation.* Oxford University Press, New York.

SPERBER, D. et WILSON, D. (1989). *La pertinence: communication et cognition.* Editions de Minuit, Paris.

STAMM, M., éditeur (1998). « *Philosophie in Synthetischer Absicht* », Partie Myself and I, John Perry, pages 83–103. Klett-Cotta.

STERN, J. (1998). *La science du secret.* Editions Odile Jacob, Paris.

STINSON, D. (1996). *Cryptographie - Théorie et Pratique.* Vuibert, 2001, Paris. Tr. fr. Serge Vaudenay.

SYVERSON, P. F. (1990). « Formal Semantics for Logics of Cryptographic Protocols ». Dans *Proc. Computer Security Foundations Workshop*, pages 32–41. IEEE Computer Society Press.

SYVERSON, P. F. (1994). « A taxonomy of replay attacks ». Dans *Proceedings of the Computer Security Foundations Workshop (CSFW7)*, pages 187–191. IEEE Computer Society Press.

SYVERSON, P. F. et CERVESATO, I. (2000). « The Logic of Authentication Protocols ». Dans *FOSAD*, pages 63–136.

SYVERSON, P. F. et VAN OORSCHOT, P. C. (1994). « On Unifying Some Cryptographic Protocol Logics ». Dans *1994 IEEE Computer Society Symposium on Research in Security and Privacy*, pages 14–28.

SYVERSON, P. F. et VAN OORSCHOT, P. C. (1996). « A unified cryptographic protocol logic ». *NRL Publication 5540-227.* Naval Research Lab.

TARDE, G. (2001). *Les lois de l'imitation.* Les Empêcheurs de penser en rond - Editions du Seuil.

TEUSCHER, C. et SANCHEZ, E. (2000). « A Revival of Turing's Forgotten Connectionist Ideas: Exploring Unorganized Machines ». Dans FRENCH, R. M. et SOUGNÉ, J. P., éditeurs, *Connectionist Models of Learning, Development and Evolution. Proceedings of the 6th Neural Computation and Psychology Workshop, NCPW6, Liège, Belgium, 16–18 September 2000*, Perspectives in Neural Computing, pages 153–162. Springer-Verlag, London.

TIERCELIN, C. (1993). « *La pensée-signe: études sur C. S. Peirce* », Partie 4 Un nouveau modèle du mental: pensée-signe et machines logiques, pages 182–257. Editions Jacqueline Chambon, Nîmes.

TURING, A. et GIRARD, J.-Y. (1995). *La machine de Turing.* Editions du Seuil, Paris.

TURING, A. M. (1936). « On Computable Numbers, With an Application to the Entscheidungsproblem ». Dans *Proceedings of the Mathematical Society*, volume 42 de *2*, pages 230–265.

TURING, A. M. (1950). « Computing Machinery and Intelligence ». *Mind*, LIX(236).

TURING, A. M. (1999). « Turing's Treatise on Enigma ». http:// www.turing.org.uk/ turing/ scrapbook/ treatise.html. Reproduit à partir des *National Archives and Records Administration.*

WAGNER, P. (1998). *La machine en logique.* Presses Universitaires de France, Paris.

WIENER, N. (2000). *God & Golem Inc.* Editions de l'éclat.

ZÉMOR, G. (2000). *Cours de cryptographie.* Cassini, Paris.

Glossaire

AES *Avanced Encryption Standard* Standard avancé de chiffrement tendant à remplacer le DES.

Alice Prénom conventionnellement donné à l'agent envoyant un message chiffré ou à celui initiant un protocole d'authentification. Abrégé en A. Voir Bernard et Eve.

BAN Logique portant le nom de ses auteurs: Burrow, Abadi, Needham. La logique BAN est la première logique de l'authentification (1989).

Banburismus Nom donné par Turing au calcul des bans, c'est-à-dire au calcul des «poids d'évidence» accordés à une hypothèse. Le mot «ban» lui-même vient de la ville de Banbury, lieu de provenances des cartes sur lesquels étaient imprimés les calculs. Voir (Hodges, 1988, p.175)

Bernard Prénom conventionnellement donné au destinataire d'un message chiffré ou à l'agent sollicité lors d'un protocole d'authentification. Abrégé en B. Voir Alice et Eve.

Bombe Nom donné aux machines conçues par le polonais Martin Rejewski avant la seconde guerre mondiale et qui servaient à la cryptanalyse d'*Enigma*. Ce nom viendrait du bruit de tic tac de ces machines.

Carré de Vigenère Table où sont mis en correspondance un alphabet en clair et les différents alphabets par lequel il peut être substitué, très utile pour accélérer le chiffrement de Vigenère. Voir tableau 4.2 page 53.

Chiffre Désigne n'importe quel cryptosystème dans lequel chaque lettre est remplacée par une autre.

Chiffre de transposition Cryptosystème dans lequel seule change la position des lettres, non leur valeur.

Chiffre de substitution Cryptosystème dans lequel seule change la valeur des lettres, non leur position.

Substitution monoalphabétique (chiffre de) Cryptosystème dans lequel un caractère est remplacé par un caractère identique tout au long du message.

Substitution polyalphabétique (chiffre de) Cryptosystème dans lequel un même caractère est remplacé par des caractères différents tout au long d'un message.

Cilly Les *cillies* désignaient les mots-clefs d'*Enigma* dont le choix n'était pas purement aléatoire. Pour se faciliter la tâche, les opérateurs utilisaient souvent des séquences de lettres se suivant sur le clavier, des mots allemands de trois lettres ou les initiales de leur femme.

Clef privée Dans un cryptosystème symétrique, désigne la clef que partagent Alice et Bernard. Parfois confondue avec la clef secrète.

Clef publique Dans un cryptosystème à clef publique, la clef publique est la clef de chiffrement: tout le monde y a accès. Sa contrepartie est la clef secrète, connue du seul destinataire.

Clef secrète Dans un cryptosystème à clef publique, la clef secrète est la clef de déchiffrement: seul son utilisateur la connaît. Sa contrepartie est la clef publique, clef de chiffrement publiquement accessible.

Clef à long terme Dans un protocole cryptographique, désigne une clef dont la durée de vie est supérieure à la durée du protocole. Il s'agit généralement des clefs publiques et secrètes de chacune des entités, dont la correspondance est garantie par un serveur de clefs.

Clef à court terme Dans un protocole cryptographique, désigne une clef dont la durée de vie est inférieure à la durée du protocole. Il s'agit généralement des clefs de session.

Clef de session Clef à court terme utilisée pour une unique session de communication. La clef de session est échangée à l'intérieur d'un cryptosystème asymétrique, employée à l'intérieur d'un cryptosystème symétrique et généralement garantie par un serveur de clefs.

Code Désigne n'importe quel cryptosystème dans lequel des unités sémantiques sont remplacées par des symboles.

Crib Les *cribs* désignaient les mots probables, ceux dont on soupçonnait l'existence à l'intérieur d'un message chiffré et à partir desquels on procèdait à la cryptanalyse. La technique des mots probables fut mise au point et exploitée par Turing pendant la seconde guerre mondiale.

Cryptogramme Message chiffré.

Cryptographie asymétrique Cryptosystème dans lequel il est impossible (i.e. aussi difficile qu'un problème NP-complet) de calculer l'algorithme de déchiffrement à partir de l'algorithme de chiffrement.

Cryptographie symétrique Cryptosystème dans lequel chiffrement et déchiffrement ne sont qu'une seule et même opération.

Cryptologie Science et technique du chiffrement et déchiffrement des messages. La cryptologie réunit la cryptographie (chiffrement) et la cryptanalyse (déchiffrement).

Cryptosystème Abréviation usuelle pour «système de cryptographie».

DES *Data Encryption Standard* Standard de chiffrement des données, développé par IBM et adopté en 1976. Tend à être remplacé par AES.

Digramme Couple de lettres. Après avoir analysé la fréquence d'apparition des lettres, on analyse la fréquence d'apparition des digrammes pour obtenir des corrélations plus fines entre texte clair et texte chiffré. Voir aussi trigramme.

Enigma Machine à chiffrer et à déchiffrer mise au point par l'allemand Arthur Scherbius dès la fin de la première guerre mondiale. Il en vendit plus de 30 000 à l'armée allemande, et celle-ci sut longtemps en tirer avantage. C'est à la cryptanalyse d'*Enigma* que se consacrent les chercheurs de Bletchley Park pendant la seconde guerre mondiale. Voir la section 4.2 page 61.

Eve Prénom conventionnellement donné à l'agent qui espionne un canal de communication pour déchiffrer, modifier ou usurper un message. Dans un protocole d'authentification, le but principal d'Eve est de se faire passer pour l'un des agents. Plus les pouvoirs accordés

à Eve sont importants, plus la sécurité du système est mise à l'épreuve. Abrégé en E. Voir Alice et Bernard.

Marqueur temporel Traduction de *timestamp*. Utilisé dans les protocoles d'authentification pour la datation des messages.

Mot-clef Suite de caractères utilisée dans les chiffrements polyalphabétiques et ses variantes. La première lettre du texte clair est chiffrée avec une des lettre du mot-clef, la seconde avec la lettre suivante, etc. On utilise le mot-clef autant de fois que l'exige la longueur du texte clair. La sécurité du cryptosystème augmente avec la longueur du mot-clef.

Nonce Nombre aléatoire généré à l'intérieur d'un protocole d'authentification et utilisé pour indiquer la fraîcheur d'un message

NSSK *Needham-Schroeder Shared Key* Protocole Needham-Schroeder à clef secrète.

Prix Loebner Prix existant depuis 1991 et récompensant le programme obtenant les meilleurs résultats dans un test de Turing. Voir l'annexe, section 4.5 page 72.

Index des noms

Liste des tableaux

Table des figures

Table des matières