

Lê Nguyên Hoang

LA FORMULE DU SAVOIR

**Une philosophie unifiée du savoir
fondée sur le théorème de Bayes**



Lê Nguyêt Hoang

Collaborateur scientifique à l'EPFL, Suisse.
Vidéaste sur la chaîne YouTube Science4All.

Imprimé en France

ISBN (papier) : 978-2-7598-2260-7 – ISBN (ebook) : 978-2-7598-2261-4

Tous droits de traduction, d'adaptation et de reproduction par tous procédés, réservés pour tous pays. La loi du 11 mars 1957 n'autorisant, aux termes des alinéas 2 et 3 de l'article 41, d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinés à une utilisation collective », et d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation intégrale, ou partielle, faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (alinéa 1^{er} de l'article 40). Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles 425 et suivants du code pénal.

© EDP Sciences, 2018

Préface

Arrivant dans une petite ville, avec une lourde valise, vous vous dirigez vers la station de taxi de la gare, où une seule voiture est stationnée. Hélas, le temps que vous vous en approchez, un voyageur plus rapide l'a déjà empruntée et elle disparaît sous vos yeux. Quelle conclusion pouvez-vous tirer de cette mésaventure ? Qu'il semble y avoir des taxis dans cette ville – vu sa taille, c'était loin d'être assuré – et donc, que si vous attendez patiemment, un autre finira par se présenter ? Ou alors, que l'un des rares taxis de la ville vient de vous échapper et que, vu sa taille, une telle aubaine ne se représentera de sitôt ? Ces deux interprétations sont correctes, mais l'une et l'autre dépendent de ce que vous saviez – ou croyiez – avant de descendre du train.

Ce voyageur qui arrive dans une ville inconnue, fait des hypothèses sur le nombre de taxis et les révise en fonction de ses observations, n'est pas très différent d'un bébé qui arrive dans un monde inconnu, ou du chercheur qui, s'étonnant de ce que les autres tiennent pour acquis, se demande pourquoi le soleil se lève tous les matins. Les uns comme les autres explorent le monde, font des hypothèses et les révisent en fonction de leurs observations.

Quel enseignement pouvons-nous tirer de nos expériences ? Que pouvons-nous connaître du monde ? Ce sont ces questions que le magnifique livre de Lê Nguyên Hoang nous invite à nous poser.

Sur ces questions, un point cristallise les controverses depuis plus d'un siècle : est-il possible d'associer, à une hypothèse, une valeur numérique qui mesure sa vraisemblance ? Pour certains, tel Hans Reichenbach, c'est là le but même de la théorie des probabilités. En particulier, toute observation qui confirme une hypothèse, augmente sa probabilité d'être vraie : chaque observation d'un corbeau noir augmente la probabilité que l'hypothèse selon laquelle tous les corbeaux sont noirs soit vraie. Pour d'autres, tel Karl Popper, l'attribution d'une valeur numérique à une telle hypothèse n'est qu'illusion. En observant un corbeau noir, nous ne pouvons que conclure, que notre hypothèse selon laquelle tous les corbeaux sont noirs reste cohérente avec nos observations.

Au cœur de cette controverse, se place une formule d'une simplicité déconcertante, la formule de Bayes, « la formule du savoir », qui donne son titre à ce livre, et qui permet justement de calculer la probabilité que nous devons attribuer à une hypothèse après avoir fait une observation – et donne donc ainsi raison à Reichenbach – mais uniquement à condition que nous ayons su attribuer une

probabilité à cette hypothèse, avant d'effectuer cette observation – et donne donc ainsi raison à Popper.

Si cette question semblait tranchée – en faveur de Popper – au XX^e siècle, l'évolution des techniques de collecte des données, la renouvelle aujourd'hui. Quand nous croyions, au XX^e siècle, qu'il existait des corbeaux blancs, nous pouvons interpréter le fait que les trois corbeaux observés fussent noirs, comme une coïncidence. Quand nous observons, aujourd'hui, mille, un million, un milliard... de corbeaux, et qu'ils sont tous noirs, il faut avoir un certain courage – voire une certaine obstination – pour prétendre que non, tous les corbeaux ne sont pas noirs, et la concordance de nos observations n'est que coïncidence. Au moins sommes-nous contraints de concéder qu'il doit y avoir, parmi les corbeaux, une grande proportion de corbeaux noirs, et même, sans doute, que les corbeaux blancs relèvent de l'exception. Cette objection à la thèse de Reichenbach que constituait le problème des hypothèses *a priori* mis en évidence par la formule de Bayes, se trouve aujourd'hui relativisée par le déluge des données. D'autres problèmes, en revanche, apparaissent : comment ces données ont-elles été collectées ? La méthode de collecte n'introduit-elle pas un biais, voire une discrimination, à l'égard des corbeaux blancs ? Une fois de plus, nous constatons à quel point l'évolution des techniques, notamment des techniques d'instrumentation scientifique, change la manière dont se posent les questions en philosophie des sciences.

C'est cela qui rend le livre de Lê Nguyêñ Hoang passionnant. Il aura été écrit à l'époque d'un basculement. À une époque où l'évolution des techniques changeait le regard que nous portons sur la formule de Bayes et sur sa place dans l'édifice de la connaissance.

Il aura aussi été écrit à une époque où les techniques de communication changeaient notre manière de parler des sciences. Formé à la dure école des vidéos en ligne, Lê Nguyêñ Hoang, a su trouver un ton nouveau pour parler des sciences, un ton à la fois rigoureux et narratif, où les exemples illuminent les questions les plus abstraites.

Gilles Dowek,
Chercheur à l'Inria,
Professeur à l'École normale supérieure de Paris-Saclay.

Table des matières

1 Mon voyage initiatique	13
Collé par un étudiant	13
Sur les traces du bayésianisme	14
Une philosophie unifiée du savoir	16
Une alternative à la méthode scientifique	18
Le mythe de l'objectivité	20
Les objectifs du livre	23
2 Le théorème de Bayes	27
L'énigme des enfants	27
Le problème de Monty Hall	28
Le procès de Sally Clark	30
Le bayésianisme jugé illégal	31
Le théorème de Bayes	32
Les composants de la formule de Bayes	34
Bayes au secours du diagnostic	35
Bayes au secours de Sally Clark	37
L'énigme des enfants enfin résolue !	38
Quelques mots d'encouragement	39
3 Logiquement...	41
Deux modes de raisonnement	41
Les règles de la logique	43
Les dames sont-elles toutes bleues ?	45
Quantificateurs et prédictats	46
Le syllogisme d'Aristote réinterprété	47
L'axiomatisation	48
Platoniciens <i>versus</i> intuitionnistes	48
La logique bayésienne*	50
Au-delà du vrai ou faux	51
Vers une cohabitation de théories incompatibles	53
4 Il faut (bien) généraliser !	57
Le mouton noir d'Écosse	57
Une brève histoire de l'épistémologie	58
Une brève histoire de la planétologie	59

Les sciences contre Popper ?	60
Le fréquentisme*	61
Les statisticiens contre la <i>p-value</i>	63
Le <i>p-hacking</i>	64
Ce qu'en dit un cours de statistique	66
La formule du savoir	67
L'apprentissage cumulatif	68
Revenons-en à Einstein	69
5 Gloire aux préjugés	73
Le problème de Linda	73
Les préjugés au secours de Linda*	74
Les préjugés sont indispensables	76
Le soleil d'xkcd	77
Les préjugés au secours d'xkcd	77
Les préjugés au secours de Sally Clark	78
Les préjugés pour lutter contre les pseudo-sciences	79
Les préjugés au secours des sciences	81
Le bayésien a un préjugé sur <i>tout</i>	83
Les préjugés erronés	86
Les préjugés et la morale	88
6 Les prophètes du bayésianisme	91
Une histoire mouvementée	91
Les origines de la théorie des probabilités	92
Le mystérieux Thomas Bayes	93
Laplace, le père du bayésianisme	94
La loi de succession de Laplace	96
Le grand hiver du bayésianisme	99
Bayes au secours des alliés	100
Des îlots bayésiens dans un océan fréquentiste	102
Bayes secouru par les praticiens	103
Le triomphe de Bayes, enfin !	105
Bayes est partout	106
7 Le démon de Solomonoff	109
Ni homme ni machine	109
Les fondements de l'algorithme	110
Qu'est-ce qu'un <i>pattern</i> ?	112
La complexité de Solomonoff	113
Le mariage de l'algorithme et des probabilités	115
Le préjugé de Solomonoff*	118
Bayes au secours du démon de Solomonoff*	119
La complétude de Solomonoff	120
L'incalculabilité de Solomonoff	121
L'incomplétude de Solomonoff	122

En quête de pragmatisme	123
8 Garder le secret	127
Classé confidentiel	127
La cryptographie d'aujourd'hui	128
Bayes à l'assaut des codes cryptés	129
Le sondage randomisé	131
La confidentialité du sondage randomisé	133
La définition de la confidentialité différentielle*	134
Le mécanisme laplacien	136
Robustesse à la composition	137
L'additivité des pertes de confidentialité	138
En pratique, ça ne va pas !	139
Le chiffrement homomorphe	140
9 Les jeux sont faits	143
La magouilleuse	143
Split or Steal	145
La persuasion bayésienne	146
Les points de Schelling	148
L'équilibre mixte	149
Les jeux bayésiens	152
La conception de mécanismes bayésiens*	153
L'enchère de Myerson	154
Les conséquences sociétales du bayesianisme	155
10 Darwin et Bayes font affaire	159
Le biais du survivant	159
Les lézards colorés de Californie	160
La dynamique de Lotka-Volterra*	161
Les algorithmes génétiques	163
Se faire son propre avis ?	163
Un scientifique n'est pas crédible	165
L'argument d'autorité	167
Le consensus scientifique	169
Le putaclic	170
La puissance prédictive des marchés	171
Les bulles financières	174
11 Exponentiellement contre-intuitif	177
Les nombres archi-méga-super géants	177
Le plafond de verre des calculs	178
L'explosion exponentielle	180
La magie des chiffres indo-arabes	183
La loi de Benford	184
L'échelle logarithmique	185

Le logarithme	187
Bayes rafle un prix Gödel	188
Bayes part en vacances	190
La singularité	191
12 Tranchons avec le rasoir d’Ockham	195
Jeudi dernier	195
Dans le football, rien n'est écrit d'avance	197
Le fléau de la sur-interprétation	198
La complexe quête de simplicité	201
Tout n'est pas simple	202
La validation croisée	204
La régularisation de Tibschirani	206
L'optimisation robuste	207
Bayes au secours de l' <i>overfitting</i> *	208
Seules les inférences bayésiennes sont admissibles*	209
Le rasoir d'Ockham déduit du bayésianisme !	210
13 Les faits sont trompeurs	213
Hôpital ou clinique ?	213
Corrélation n'est pas causalité	215
Cherchez les facteurs de confusion	217
La régression à la moyenne	219
Le paradoxe de Stein	220
L'échec de la stratification endogène	221
Randomisons !	223
Le retour du mouton noir d'Écosse	225
Qu'est-ce qu'un chat ?	226
Le naturalisme poétique	228
14 Vite et (assez) bien	231
Le mystère des nombres premiers	231
Le théorème des nombres premiers	233
Les approximations de τ	234
Les développements limités	235
Les contraintes du pragmatisme	236
Les <i>learning machines</i> de Turing	236
Le bayésianisme pragmatique	239
Les algorithmes sous-linéaires	241
Plusieurs modes de réflexion	243
Devenez post-rigoureux !	244
Les approximations de Bayes	245
15 La faute à pas de chance	249
FiveThirtyEight et l'élection présidentielle de 2016	249
La mécanique quantique est-elle probabiliste ?	250

La théorie du chaos	253
Les automates déterministes imprévisibles	254
La thermodynamique	255
L'entropie de Shannon	256
La compression optimale de Shannon	258
La redondance de Shannon	259
La divergence de Kullback-Leibler	260
La métrique de Wasserstein	262
Les <i>Generative Adversarial Networks</i> (GANs)	263
16 Trou de mémoire	267
La valeur des données	267
Le déluge de données	268
Le problème des toilettes	269
Traiter rapidement un déluge de données	270
Le filtre de Kalman	272
Nos cerveaux confrontés au <i>Big Data</i>	273
Effacer les souvenirs traumatisques	274
Les faux souvenirs	276
Bayes au secours de la mémoire	278
Des mémoires à plus ou moins long terme	279
Les réseaux de neurones récurrents	280
Que faut-il apprendre et enseigner ?	282
17 La nuit porte conseil	285
D'où viennent les idées ?	285
L'art créatif des intelligences artificielles	286
L'allocation de Dirichlet latente (LDA)	287
Le restaurant chinois au secours de LDA	289
Les simulations de Monte-Carlo	290
La descente de gradient stochastique (SGD)	292
Les nombres pseudo-aléatoires	293
L'échantillonnage préférentiel	293
L'échantillonnage préférentiel au secours de LDA	294
Le modèle d'Ising*	296
La machine de Boltzmann	297
MCMC et Google PageRank	298
L'échantillonnage de Metropolis-Hasting	300
L'échantillonnage de Gibbs	301
MCMC et les biais cognitifs	302
La divergence contrastive et les rêves	304
18 La déraisonnable efficacité de l'abstraction	307
Le <i>deep learning</i> , ça marche !	307
L'apprentissage des <i>features</i>	309
La représentation vectorielle des mots	310

L'expressivité exponentielle*	311
L'émergence de la complexité	313
La sophistication de Kolmogorov*	314
La sophistication est un MAP de Solomonoff !*	315
La profondeur logique de Bennett	317
La profondeur des mathématiques	318
La concision des mathématiques	320
La modularité des mathématiques	321
19 Le cerveau bayésien	325
Le cerveau est incroyable	325
Montagne ou vallée ?	326
Les illusions optiques	328
La perception du mouvement	329
Échantillonnage bayésien	330
Le scandale de l'induction	332
Apprendre à apprendre	333
La bénédiction de l'abstraction	334
Le bébé est un génie	336
Le langage	336
Apprendre à compter	338
La théorie de l'esprit	339
Innée ou acquis ?	340
20 Tout est fiction	343
La grotte de Platon	343
L'antiréalisme	344
La vie existe-t-elle ?	345
L'argent existe-t-il ?	346
La téléologie, impasse scientifique ?	349
Ce que la thèse de Church-Turing dit de la réalité	353
L'antiréalisme (instrumentaliste) est-il utile ?	354
Y a-t-il un monde extérieur au cerveau ?	356
Y a-t-il un chat dans un code binaire ?	357
L'antiréalisme du démon de Solomonoff	358
21 Aux origines des croyances	361
Le scandale des séries divergentes	361
Mais c'est faux, non ?	363
Élève officier	364
Mon périple en Asie	365
Tous des monstres en puissance ?	367
Les histoires ont plus d'effet que les chiffres	368
Les superstitions	370
L'évolution darwinienne des idéologies	371
Croire les superstitions est utile	373

La magie de YouTube	375
Le périple continue	376
22 Au-delà du bayésianisme	379
Le bayésien n'a pas de morale	379
La morale (sélectionnée par la sélection) naturelle	380
Inconscients de nos morales	382
Des bâtons et des carottes	385
La morale du plus grand nombre ?	386
La morale déontologique	388
Le savoir est-il une fin raisonnable ?	390
L'utilitarisme	392
La <i>conséquentialiste bayésienne</i>	394
Le mot de la fin	397

La théorie des probabilités n'est, au fond, que le bon sens réduit au calcul ; elle fait apprécier avec exactitude ce que les esprits justes sentent par une sorte d'instinct, sans qu'ils puissent souvent s'en rendre compte.

Pierre-Simon Laplace (1749-1827)

1

Mon voyage initiatique

Collé par un étudiant

À l'issue d'un cours de probabilités et statistiques que je donnais à l'École Polytechnique de Montréal, un étudiant *troll* vint me coller avec une énigme d'apparence triviale. Un homme a deux enfants. Au moins l'un d'eux est un garçon. Quelle est la probabilité que l'autre soit également un garçon ?

Après quelques secondes de réflexion, je réussis à trouver la bonne réponse à cette question — qui, comme on va le voir plus loin, n'est pas $1/2$. L'étudiant acquiesça, et enchaîna avec une seconde énigme. Supposez que vous appreniez maintenant que l'homme a au moins un garçon né un mardi. Que devient la probabilité que l'autre enfant soit également un garçon ?

Cette fois-ci, ma réponse fut une mauvaise réponse. L'étudiant m'avait collé.

Le réflexe usuel est certainement de penser que ces deux énigmes ne sont que des jeux mathématiques. Il y a en effet une bonne réponse, mais il ne s'agit d'une bonne réponse que dans un cadre mathématique rigide et restreint. On retrouve ce genre de problèmes en exercice ou à l'examen à l'école. Mais il ne s'agit *que* de mathématiques.

Pourtant, l'énigme de l'étudiant *troll* n'est qu'une version ultra-simplifiée de nombreuses réflexions qui encombrent nos quotidiens. Faut-il croire un diagnostic médical ? La présomption d'innocence est-elle justifiée ? Les juges font-ils de la discrimination raciale ? Faut-il craindre le terrorisme ? Peut-on

généraliser à partir d'un exemple ? À partir de mille exemples ? Un million ? L'argument d'autorité a-t-il une quelconque valeur ? Faut-il faire confiance aux marchés financiers ? Les OGM sont-ils nocifs ? En quoi la science aurait-elle plus « raison » que les « pseudo-sciences » ? Les robots vont-ils conquérir le monde ? Faut-il condamner le capitalisme ? Faut-il croire en l'existence de Dieu ? Qu'est-ce que le bien et le mal ?

Pour beaucoup, ces questions n'ont absolument rien à voir avec les mathématiques. Et en effet, les mathématiques seules sont impuissantes face à de telles questions. Vous ne résoudrez pas le problème de la faim dans le monde uniquement en prouvant des théorèmes. Mais il y a fort à parier que les mathématiques peuvent contribuer à mieux structurer notre réflexion, à mieux comprendre les tenants et les aboutissants et à fournir des solutions inattendues. Ce n'est sans doute pas un hasard si de nombreuses disciplines se voient de plus en plus mathématisées — y compris l'aide humanitaire¹.

Malgré le pulluler de nombreux modèles mathématiques, il semble que la plupart d'entre nous persiste à vouloir distinguer le « monde réel » des cours académiques que l'École nous force à suivre. En particulier, le monde réel, pense-t-on souvent, dépasse de loin le cadre des mathématiques, si bien que les théorèmes des mathématiques ne semblent jamais devoir, ni même pouvoir, s'appliquer au monde réel. À quel point faut-il être idiot pour penser que les mathématiques ont un quelconque mot à dire sur l'égalité devant la loi² ?

Cette défiance de l'utilité des mathématiques ne se réduit d'ailleurs pas à un réflexe de mauvais élève. Moi-même, pendant les années qui suivirent mon échec face à mon étudiant *troll*, je ne me suis pas rendu compte que cet échec mathématique révélait mon incapacité à raisonner correctement sur le monde réel. Moi-même, je ne compris pas qu'une meilleure compréhension de cette énigme m'aiderait à mieux écouter les conseils de mes amis baroudeurs pour mieux choisir ma prochaine destination de vacances — on y reviendra.

Sur les traces du bayésianisme

Certes, c'est le soir même que je résolus l'énigme de l'étudiant *troll* — au prix de mystérieux calculs obscurs. Mais ce n'est que trois ans plus tard, début 2016, au moment où je me suis intéressé de plus près au débat qui oppose les statisticiens fréquentistes aux statisticiens bayésiens³, que je commençai à vraiment méditer l'énigme de l'étudiant *troll* et, surtout, à la sortir d'un cadre purement mathématique.

¹  *A Set-Partitioning Formulation for Community Healthcare Network Design in Under-served Areas* | M. Cherkesly, M.E. Rancourt et K. Smilowitz (2017)

²  *Partager un gâteau, c'est pas du gâteau !* Démocratie 22 | Science4All | L.N. Hoang (2017)

³  *Les statistiques à l'heure du Big Data* | CESP Villejuif | L.N. Hoang (2016)

En particulier, pendant les deux ans qui suivirent, presque une fois par jour, je me mis à réfléchir à la formule magique qui résolvait cette énigme. Pour mon plus grand bonheur, petit à petit, cette formule mystérieuse commença à me révéler ses secrets. Lentement mais sûrement, cette formule lumineuse me séduisit, au point de changer la manière dont je concevais le monde, les sciences et la connaissance. Au fil des mois, je finis par être submergé par la sublime élégance de cette formule indomptable. Ce fut trop. Il me fallait dédier tout un livre à son sujet. C'est ainsi que fin 2016, je me lançai dans l'écriture du livre que vous venez d'entamer.

La formule indomptable dont je parle, c'est celle que j'aime appeler pompeusement la formule du savoir. Mais les mathématiciens, statisticiens et informatiens la connaissent davantage sous le nom de formule de Bayes.

La formule de Bayes est un théorème mathématique d'une simplicité remarquable. Il s'agit d'une équation compacte, que l'on enseigne dès le lycée. Sa preuve ne fait qu'une seule ligne, et ne repose que sur la connaissance de la multiplication, de la division et de la notion de probabilité. En particulier, elle paraît beaucoup plus facile à apprendre que de nombreux autres concepts de mathématiques que l'on demande aux lycéens et aux étudiants du supérieur de maîtriser.

Et pourtant. J'ose prétendre que même les meilleurs mathématiciens ne comprennent pas cette formule de Bayes — et il y a même des mathématiciens qui expliquent notre incapacité à saisir cette formule ! Sans aller jusque-là, il ne fait aucun doute à mes yeux que *je ne comprends toujours pas la formule de Bayes*. D'ailleurs, si j'avais compris la formule de Bayes au moment où j'enseignais le cours de probabilités et statistiques, j'aurais immédiatement vu le lien entre le fait qu'un garçon est né un mardi et le sexe de son frangin. J'aurais répondu du tac au tac à mon étudiant *troll*. Il ne m'aurait pas collé.

Depuis près de deux ans, je torture mon esprit pour ne plus jamais me faire coller ainsi. Je veux connaître, comprendre, sentir la formule de Bayes. J'ai déjà beaucoup appris, et je continue à apprendre. Je médite la formule de Bayes presque quotidiennement, comme s'il s'agissait d'une divinité à qui je me devais de consacrer une partie de mes journées. Et quel bonheur que de méditer cette formule ! Loin d'être un effort répétitif, ces méditations semblent avoir continuellement alimenté ma curiosité, me chuchotant, secrètement et au compte-gouttes, ses implications inattendues.

Au bout de longs mois de réflexions, j'ai fini par me laisser convaincre que peu d'idées étaient aussi profondes que la formule de Bayes. Je suis tombé amoureux de cette formule. Au point que je prétends volontiers aujourd'hui que la « rationalité » se réduit essentiellement à l'application de la formule de Bayes — auquel cas personne n'est rationnel ! Tel est, en tout cas, le fondement de ce que l'on pourrait appeler la philosophie de Bayes, ou *bayesianisme*.

Une philosophie unifiée du savoir

N’ayant pas encore eu le temps de vous présenter la formule de Bayes, je suis constraint de rester volontairement flou sur ce qu’est le bayésianisme pour l’instant. Mais s’il fallait le résumer en trois phrases maladroites, je donnerais la définition suivante. Le bayésianisme consiste à supposer que tout modèle, toute théorie ou toute conception de la « réalité » n’est que croyance, fiction ou poésie ; en particulier, « tous les modèles sont faux ». Les données empiriques doivent ensuite nous forcer à ajuster l’importance ou la *crédence* que l’on assigne aux différents modèles. De façon cruciale, la manière dont ces crédences sont ajustées doit obéir aussi rigoureusement que possible à la formule de Bayes.

J’ai longtemps rejeté la pertinence de cette philosophie du savoir. Elle semble discréditer toutes notions de réalité ou de vérité, pourtant si chères à tant de scientifiques. Cependant, elle semble aussi parfaitement coller avec ce que disait le physicien et prix Nobel Richard Feynman⁴ : « je peux vivre avec le doute et l’incertitude. Je peux vivre sans savoir. Je pense qu’il est beaucoup plus intéressant de vivre sans savoir que d’avoir des réponses qui pourraient être fausses. J’ai des réponses approximatives, j’ai des croyances plausibles avec différents degrés de certitude sur différents sujets. Mais je ne suis absolument sûr de rien. Et il y a beaucoup de choses dont je ne sais absolument rien. Mais je ne suis pas obligé d’avoir une réponse. Je ne me sens pas effrayé par le fait de ne pas savoir. »

Vous pourriez trouver ce point de vue exaltant. Ou vous avez peut-être envie de rejeter en bloc cette approche de la connaissance. Cependant, avant un éventuel rejet ou une potentielle adhésion au bayésianisme, je ne peux que vous encourager à d’abord prendre le temps de longuement méditer la formule de Bayes et ses conséquences.

Dans ce livre, malheureusement, le guide principal que je serai n’a qu’une compréhension très incomplète de cette équation. Pour nous aider dans nos réflexions, je vais invoquer un personnage fictif (féminin), la *pure bayésienne*, et on cherchera à imaginer comment cette *pure bayésienne* réagirait à divers contextes. Plutôt que moi-même, c’est cette *pure bayésienne* que je vous invite à mettre à l’épreuve. C’est d’ailleurs ce que l’on ne cessera de faire dans ce livre. On va multiplier les expériences de pensée, qui seront autant de défis que la *pure bayésienne* devra relever. Et on prendra le soin de guetter, juger et critiquer les diverses réactions de cette *pure bayésienne* — même si ces critiques se transformeront souvent bien vite en celles de notre intuition et de notre inlassable excès de confiance.

Si le premier bayésien de l’histoire digne de ce nom, le géant Pierre-Simon Laplace, n’avait qu’une description partielle de cette *pure bayésienne*, cela fait déjà un demi-siècle que tous les calculs, réflexions et prédictions de la *pure bayésienne* ont été rigoureusement décrits par le génialissime Ray Solomonoff.

⁴  The Feynman Series - Beauty | Reid Gower (2011)

Malheureusement, et comme on le verra en détails, la *pure bayésienne* que décrit Solomonoff semble nécessairement devoir enfreindre les lois de la physique — en particulier la thèse de Church-Turing dont on reparlera.

Voilà qui nous constraint à un bayésianisme inéluctablement approché, que je qualifierai de *pragmatique*. Ce *bayésianisme pragmatique*, qui se distingue fortement du bayésianisme pur par ses exigences de calculabilité (rapide), sera incarné par un autre personnage fictif (masculin), que j'appellerai *bayésien pragmatique*. Malheureusement (ou pas !), ma description du *bayésien pragmatique* sera très incomplète, puisque le bayésianisme pragmatique est encore un énorme champ de recherche très ouvert — et il n'est pas dit qu'il puisse un jour être entièrement résolu.

Comme vous commencez sans doute à le deviner, comprendre la *pure bayésienne* et le *bayésien pragmatique* ne sera pas chose aisée. Pour ce faire, il va nous falloir aborder de nombreux concepts fondamentaux des mathématiques, de la logique, des statistiques, de l'informatique, de l'intelligence artificielle, voire des notions de physique, de biologie, de neurosciences, de psychologie et d'économie. On devra parler de logarithme, de contraposition, de *p-value*, de complexité de Solomonoff et de réseaux de neurones, mais aussi d'entropie, de l'évolution darwinienne, des faux souvenirs, des biais cognitifs et des bulles financières. Et puis, on va aussi utiliser de nombreux exemples de l'histoire des sciences pour mettre nos deux héros fictifs au défi.

Oui, je sais, ça fait beaucoup à savoir pour comprendre la formule de Bayes...

Mais ça tombe bien, puisque j'aime expliquer les sciences modernes à mes heures perdues — à tel point que j'ai lancé ma chaîne YouTube appelée Science4All ! Du coup, plutôt que de lire ce livre comme un livre de philosophie, je vous invite à le lire (aussi) comme un livre qui promeut les sciences et les mathématiques. D'ailleurs, sur le chemin du bayésianisme, je n'hésiterai pas à faire quelques détours à travers les méandres des sciences, avec l'objectif secret de vous donner envie d'aller plus loin dans votre apprentissage des théories scientifiques !

Mais revenons-en à la philosophie pour l'instant. Comme vous l'aurez compris, j'ai fini par succomber aux sirènes bayésiennes. Après de longs mois de réflexions, sans l'avoir anticipé, le bayésianisme m'a séduit au point que je me sente obligé de vous en parler. J'ai sans cesse trouvé la *pure bayésienne* bougrement intelligente, si bien que j'aspire de plus en plus à lui ressembler... Et même bien après le début de l'écriture du livre, je n'ai cessé de découvrir, encore et encore, d'innombrables merveilles époustouflantes au sujet de ce qui est depuis devenu mon équation mathématique préférée.

Quand j'ai commencé l'écriture du livre, j'étais un bayésien enthousiaste. Déjà, je suis un bayésien convaincu. J'irai même jusqu'à me décrire bayésien extrémiste, notamment par opposition à d'autres qui se disent aussi bayésiens. Mais surtout, j'aimerais un jour devenir un bayésien compétent. Je rêve d'être capable d'appliquer la formule de Bayes, car je suis convaincu que ce n'est qu'ainsi que je pourrai enfin devenir un être rationnel !

Ce qui est amusant, et peut paraître embêtant, c'est que l'élan émotionnel que cette formule de Bayes a insufflé en moi a des airs de délire irrationnel. Je ne peux pas le nier. Je suis même sûr d'être victime d'un énorme biais cognitif qui a conduit à ma sacralisation de la formule de Bayes. Après tout, il m'est impossible d'être indifférent au fait d'avoir découvert par moi-même nombre de secrets de cette formule — même si bien d'autres les ont découverts un demi-siècle avant moi.

Ceci étant dit, conscient de ce biais, je vous promets avoir lutté — et je lutte encore — contre la *pure bayésienne*. J'ai sans cesse essayé de la mettre en défaut ; j'ai sans cesse essayé de gagner un débat contre elle. En vain.

Une alternative à la méthode scientifique

En mathématiques, dès qu'une conjecture semble tenir la route, on s'empresse de la prouver pour l'ériger au rang de théorème. Ce n'est pas loin d'être le cas du bayésianisme !

On verra ainsi que le théorème de Jaynes-Cox prouve que le bayésianisme est la seule généralisation de la logique aristotélicienne capable de traiter la notion de plausibilité de manière cohérente. Le théorème de complétude de Solomonoff, lui, prouve que la *pure bayésienne* finira par toujours repérer toutes les régularités dans un jeu de données, si ces régularités existent. Quant au théorème des gains espérés nécessaires des informations additionnelles, lui montre que la *pure bayésienne* ne perd jamais à acquérir plus de données. Enfin, la théorie statistique de la décision montre que les inférences bayésiennes sont essentiellement les seules méthodes d'apprentissage admissibles, dans le sens où une méthode n'est jamais dominée par une autre, si et seulement si, elle revient à appliquer la formule de Bayes⁵.

À ces théorèmes s'ajoutent de nombreux théorèmes dont on ne parlera malheureusement pas dans ce livre. Il y a par exemple le théorème de Teller⁶-Skyrms⁷, qui montre que seul un bayésien ne sera jamais déficitaire dans un problème de « pari hollandais⁸ ». Mieux encore, le théorème de Joyce⁹ montre que l'on gagne à rendre nos croyances conformes aux lois des probabilités, comme l'impose le bayésianisme. Ces différents résultats sont ainsi parfaitement illustrés par le paradoxe des deux enveloppes¹⁰.

⁵Ces théorèmes sont abordés respectivement dans les chapitres 3, 7, 9 et 12.

⁶  *Conditionalisation and observation* | Synthese | P. Teller (1973)

⁷  *Dynamic Coherence and Probability Kinematics* | Philosophy of Science | B. Skyrms (1987)

⁸  *Argent, risques et paradoxes* | Démocratie 12 | Science4All | L.N. Hoang (2017)

⁹  *A Nonpragmatic Vindication of Probabilism* | Philosophy of Science | J. Joyce (1998)

¹⁰  *Inégalité bayésienne* | Axiome 9 | T. Giraud et L.N. Hoang (2018)

J'ai été contraint d'énoncer ces théorèmes de manière grossière, car ceux-ci correspondent à des définitions et des hypothèses difficiles à expliquer brièvement. C'est bien là que le bât blesse. Tout puriste qui veut rejeter le bayésianisme saura pinailler et rejeter les hypothèses de ces théorèmes. Je ne prétends donc pas que ces théorèmes permettent de démontrer la nécessité du bayésianisme.

Plus généralement, il semble en effet impossible de se convaincre « de façon rationnelle » du fait que le *bayésianisme* est la bonne philosophie du savoir, la bonne théorie des théories ou la bonne définition de la rationalité. Après tout, pour se convaincre de la pertinence d'un concept, il faut avoir au préalable une philosophie du savoir qui mesure la pertinence des concepts. Pour théoriser les théories, il faut une théorie qui juge et discrimine les théories des théories. Pour parler de façon rationnelle de la rationalité, il faut avoir défini la rationalité de façon rationnelle... On a là un serpent qui se mord la queue.

Cette difficulté n'est absolument pas spécifique à la formule de Bayes. Toute philosophie du savoir semble vouée à souffrir de cette auto-référence. Les mathématiciens ont d'ailleurs lutté pendant des siècles pour développer des théories sans auto-référence. Sans véritable succès (merci Gödel !).

Ainsi, un adepte de la philosophie de Popper, que certains considèrent être une description de la *méthode scientifique*, voudra fonder son savoir sur la réfutabilité de ses connaissances. Or, ce principe même de réfutabilité ne semble pas réfutable. La philosophie de Popper semble donc carrément incohérente avec elle-même. Ou du moins, il ne semble pas possible d'accepter la philosophie de Popper selon les critères de Popper. Voilà ce qui a conduit beaucoup à tracer une nette délimitation entre sciences et philosophie, ou entre sciences et théologie. Toutefois, à bien y réfléchir, cette démarcation n'est qu'un pur artefact (indésirable ?) de la philosophie de Popper.

À ce niveau là, notre *pure bayésienne* se défend bien mieux. En effet, à défaut de pouvoir prouver la validité de sa pensée en dehors de son cadre de pensée, la *pure bayésienne* — pour qui, on le verra, tout est croyance — semble pouvoir parler du *bayésianisme* sans se contredire. Mieux encore, j'ai appliqué la formule de Bayes à mes crédences en le bayésianisme suite aux expériences de pensée avec la *pure bayésienne*. Mes calculs heuristiques n'ont fait que gonfler les crédences que j'avais en la philosophie bayésienne.

Mais il y a deux autres arguments plus convaincants qui m'ont conduit à préférer le bayésianisme à d'autres philosophies du savoir. Le premier est l'universalité du bayésianisme. Contrairement à la philosophie de Popper qui restreint le champ de la connaissance, en insistant par exemple sur la reproductibilité des expériences scientifiques¹¹ et sur la réfutabilité des théories, le bayésianisme n'a aucune restriction sur son champ d'applicabilité. Tout phénomène, qu'il soit sociologique, historique ou théologique, peut être analysé sous le prisme du bayésianisme. Le bayésianisme est une philosophie universelle du savoir.

¹¹On peut voir cette exigence de reproductibilité comme une nécessité imposée par le fréquentisme.

Le second argument est la rigueur, la concision et la clarté du bayésianisme. Celui-ci définit des règles d'inférence¹² si précises qu'à appliquer (même approximativement) ces règles semble suffire à « assez bien » apprendre de notre monde. Il s'agit là d'un idéal d'informaticien, qui n'aurait ensuite plus qu'à appuyer sur le bouton « démarrer » pour permettre à sa machine d'atteindre son objectif en ne faisant que suivre une liste d'instructions. Bien entendu, il s'agit là surtout d'une description d'une intelligence artificielle ! Et ce n'est sans doute pas un hasard si, depuis trois décennies, la formule de Bayes est au cœur de nombreuses recherches dans ce domaine.

Plus récemment, sous l'impulsion de chercheurs comme Josh Tenenbaum, Karl Friston et Stanislas Dehaene, le bayésianisme semble même émerger comme étant un cadre théorique incontournable pour comprendre le fonctionnement de notre propre intelligence. En particulier, en 2012, Dehaene introduit un cours de sciences cognitives au Collège de France intitulé *Le cerveau statisticien : la révolution Bayésienne en sciences cognitives*. « Beaucoup de biologistes sont sceptiques sur l'idée qu'il puisse exister en neurosciences [...] des théories générales, » rapporte Dehaene. « [Mais] il semble bien ici qu'on ait affaire à un cadre théorique qui puisse s'appliquer de façon extrêmement large », rajoute-t-il. « L'existence même de régularités très générales dans l'architecture du cortex pourrait remonter à cette hypothèse [selon laquelle] le cerveau est organisé pour effectuer des inférences statistiques bayésiennes. »

Il semble que le bayésianisme (pragmatique) soit la solution qu'a trouvé Dame Nature pour faire émerger une vie (à peu près) intelligente¹³...

Le mythe de l'objectivité

Pourtant, mystérieusement, le bayésianisme a longtemps été rejeté par plusieurs générations de scientifiques de premier rang. Pourquoi donc ? Ces grands scientifiques étaient-ils irrationnels ? Quelle est la motivation de leur rejet du bayésianisme ? Et si le rejet est injustifié, quel fut le sophisme fallacieux de tous ces grands scientifiques ?

Il se trouve que les deux siècles de guérillas épistémologiques auxquelles ce livre va chercher à mettre fin se résument tout bêtement à la notion d'objectivité. Mieux, on peut résumer l'opposition entre bayésiens subjectivistes et *fréquentistes* objectivistes à la question suivante : *qu'est-ce qu'une probabilité* ?

À titre personnel, cette question m'aura particulièrement marqué. Elle me fut posée lors de mon oral de TIPE au concours d'entrée de l'École Normale Supérieure (ENS). Cet oral est censé être la présentation d'un projet mené tout au long de l'année. J'étais particulièrement fier du mien. J'avais modélisé des

¹²On verra bientôt ce que ceci signifie.

¹³  *Les algorithmes du vivant* | TEDxSaclay | L.N. Hoang (2018)

matchs de football, estimé les niveaux des équipes et simulé diverses compétitions. En particulier, en s'appuyant sur 2 années de résultats sportifs, mes simulations avaient fait du Portugal, de la France et de l'Italie les trois grands favoris de la coupe du monde 2006, avec des probabilités de victoire finale de 20 %, 15 % et 10 %. Pas mal, quand on sait que ces trois équipes ont fini respectivement 4^e, 2^e et 1^{re} de la compétition !

Ce travail aura énormément plu aux examinateurs des concours de Centrale et des Mines. J'y ai récolté un excellent¹⁴ 19/20. Cependant, les simulations n'ont pas intéressé les examinateurs de l'ENS. Eux me coupèrent rapidement et voulurent uniquement savoir si je savais définir les probabilités.

Ma réponse fut fréquentiste. J'affirmai que la probabilité d'un événement était sa fréquence limite, quand l'expérience était répétée un nombre infini de fois. En particulier, toute fréquence empirique semble alors être une approximation d'une probabilité fondamentale et objective. Fréquentistes ou non, les puristes de l'ENS n'ont pas apprécié. Ils attendaient en fait de moi que je redécouvre une définition mathématique des probabilités, par exemple en disant d'elles qu'elles ne sont rien d'autre que des mesures sur une sigma-algèbre de mesure totale unitaire. J'ai eu 6/20 à cet oral.

Mais oublions mon sort. Attardons-nous davantage sur ce que la *pure bayésienne* appelleraient des erreurs de jeunesse.

J'ai grandi fréquentiste. J'ai baigné dans la quête de vérités, que celles-ci soient mathématiques ou scientifiques. J'ai accepté l'existence et la supériorité des résultats *objectifs*. Y compris en 2013, lorsque l'étudiant *troll* vint me coller, la majeure partie du cours que j'enseignais était essentiellement fréquentiste — et je pensais alors que c'étaient les *bonnes* statistiques à enseigner ! D'ailleurs, mon propre modèle des matchs de football était un archétype fréquentiste qui, à l'instar du paradoxe de Stein dont on reparlera, aurait pourtant gagné à subir une touche bayésienne.

Mais ce qu'il y a de plus étonnant, c'est que la nature même des probabilités que je calculais n'avait en fait rien de fréquentiste ! La fréquence à laquelle la France gagne la coupe du monde 2006 n'est pas du tout 15 %. Cette fréquence est égale à 0. En effet, il n'y a eu qu'une coupe du monde 2006. Et la France l'a perdue.

Mais si les 15 % prédits par le modèle ne sont clairement pas une fréquence, comment les interpréter ? Peut-on encore dire qu'il s'agit d'une probabilité ?

Oui, répond la *pure bayésienne*. Il s'agit de la probabilité que la France gagne la coupe du monde *selon le modèle mathématique*. Autrement dit, cette probabilité est *subjective* ; c'est l'opinion du modèle. Mais surtout, toute probabilité est de la sorte. Selon la *pure bayésienne*, aucune probabilité, ni aucune connaissance, n'est *objective* ; et quiconque qui prétendrait le contraire prendrait ses désirs subjectifs pour une réalité à imposer aux autres.

¹⁴Il s'agissait en fait de ma note au concours TIPE-ADS.

En effet, à bien y réfléchir, toute méthode de quête et d'organisation de la connaissance semble vouée à être biaisée par le simple choix de cette méthode plutôt qu'une autre — surtout dès qu'on invoque l'imprécis rasoir d'Ockham, l'état des connaissances scientifiques déjà « établies » ou les très problématiques *p-values*. Pire encore, la manière dont nous regardons, traitons et sélectionnons les données oriente inéluctablement les conclusions déduites de notre analyse de données. Comme on en parlera longuement, les faits sont souvent incroyablement trompeurs¹⁵.

Qui plus est, la mention explicite de la méthode suivie ne suffit pas. Comme les *data scientists* qui utilisent le *machine learning* pour inférer des informations utiles du *Big Data* l'apprennent vite, l'absence d'intervention humaine n'est absolument pas une garantie d'objectivité. Humains ou machines, il semble que nous soyons toujours contraints de raisonner à l'intérieur de modèles. Il semble que nos conclusions dépendront donc nécessairement de nos modèles. Voilà qui montre, prétend la *pure bayésienne*, que tout savoir est forcément *subjectif*.

Voilà qui devrait créer un inconfort en vous. Le bayésianisme semble tendre vers le relativisme. Si toutes les connaissances sont subjectives, est-ce à dire que tous les avis se valent ? La réponse est bien sûr non. On a beau voir chacun notre rouge, ceci ne veut pas dire que tous les avis sur la présence de rouge dans le drapeau français officiel se valent.

En particulier, ceux qui appliquent rigoureusement la formule de Bayes aux mêmes données finiront par mettre leurs crédences sur les mêmes modèles, surtout si ces données sont en grand nombre. Mais surtout, selon la *pure bayésienne*, même avec relativement peu de données, les modèles qui auront gagné les crédences des bayésiens seront beaucoup plus pertinents et *utiles* que les modèles favoris d'autres qui, exposés aux mêmes données, n'ont pas appliqué la formule de Bayes.

Notez qu'en particulier, le bayésianisme (notamment pragmatique) n'est pas une alternative à la modélisation ; cette philosophie cherche davantage à distinguer les modèles utiles. Le fondement du bayésianisme est en fait très bien résumé par cette citation sacro-sainte du bayésien George Box : « Tous les modèles sont faux, certains sont utiles ». Je la réutiliserai souvent ! Que cette citation soit « vraie » ou non, elle m'aura été incroyablement utile pour court-circuiter d'interminables débats qui étaient voués à demeurer dans l'impasse, et m'ennuyaient donc à mourir. Comme bien des bayésiens avant moi, j'ai fini par trouver beaucoup plus intéressant de juger l'utilité, notamment prédictive, des modèles. Pas leur vérité.

Or, selon la *pure bayésienne*, juger adéquatement l'utilité des modèles ne peut se faire qu'à l'aide de la formule de Bayes.

¹⁵  *Le paradoxe de Simpson* | Science Étonnante | D. Louapre (2015)

Les objectifs du livre

Même si je compte bien partager et justifier l'enthousiasme pour le bayésianisme dont j'ai été victime, et même si j'ai l'espérance invaincue que ceci amènera les mathématiciens, philosophes et scientifiques à questionner ce qu'ils croyaient savoir de leurs disciplines, le but de ce livre n'est pas de vous convertir. Ce que j'aimerais, c'est surtout partager avec vous quelques-uns des joyaux qui m'ont conduit à me tourner vers le bayésianisme. Et je suis prêt à parier — un réflexe typiquement bayésien — que vous allez être vous aussi surpris et, je l'espère, quelque peu séduits, par les conséquences stupéfiantes de la formule de Bayes, ainsi que par son omniprésence dans les mathématiques appliquées, dans nos propres façons de réfléchir et dans l'organisation même de nos sociétés.

Le bayésianisme explique pourquoi la communauté scientifique est bien plus fiable que chacun de ses membres et pourquoi nos crétins de cerveaux sont constamment victimes de l'effet d'ancrage. Il explique pourquoi il est souhaitable de combiner des modèles incompatibles et pourquoi le rasoir d'Ockham est un outil indispensable. Il pourrait même être la clé pour comprendre le fonctionnement de notre mémoire et l'utilité de nos rêves. À l'instar du biologiste Theodosius Dobzhansky selon qui « rien en biologie n'a de sens, excepté à la lumière de l'évolution », j'irai même jusqu'à prétendre que nombreux sont les mécanismes qui ne peuvent être compris qu'à travers les lentilles bayésiennes.

Ma découverte du bayésianisme a aussi été pour moi l'occasion de mesurer enfin toute l'étendue de mon ignorance. En particulier, le langage des probabilités permet de quantifier l'incertitude. Mais surtout, mon incapacité à appliquer la formule de Bayes, y compris dans des cas simplistes comme l'éénigme de l'étudiant *troll*, m'a forcé à reconnaître que je suis un piètre penseur. J'ai souvent eu une confiance irrationnelle et injustifiée en mon intuition, parfois accompagnée d'une étrange méfiance en la formule de Bayes. Mais à force de m'incliner devant la *pure bayésienne*, mon périple bayésien m'aura contraint à prendre conscience de mon inlassable excès de confiance — et à davantage placer mes crédences sur la formule de Bayes. Ce sera d'ailleurs là l'un des grands objectifs de ce livre. On cherchera à lutter contre nos excès de confiance et à prendre la mesure de l'étendue de notre ignorance.

Le reste du livre peut être grossièrement décomposé en quatre parties. Dans un premier temps, des chapitres 2 à 7, on va s'attaquer frontalement à la formule de Bayes et au *pur bayésianisme*. Ensuite, les chapitres 8 à 13 auront pour but de révéler la présence cachée de principes bayésiens dans des phénomènes dont on ne suspecterait pas l'aspect bayésien. Puis, les chapitres 14 à 19 étudieront le *bayésianisme pragmatique* et ses outils incontournables. Enfin, les trois derniers chapitres seront quelque peu à part. Le chapitre 20, intitulé « Tout est fiction », étudiera les conséquences philosophiques du bayésianisme, notamment en termes de réalisme. Le chapitre 21 retracera les origines de mes croyances et questionnera notre excès de confiance récurrent. Enfin, le chapitre 22 étudiera les conséquences du bayésianisme sur la philosophie morale.

Malheureusement, ce livre, comme tout livre fini, n'est absolument pas exhaustif. Je m'excuse d'avance pour ses indénombrables déficiences. En particulier, je ne prendrai pas le temps de comparer en détail le bayésianisme aux philosophies concurrentes. Mon objectif est plus modeste : j'aimerais vous aider à comprendre les aspects importants du bayésianisme. Ou du moins ce que j'en ai compris. En effet, comme le livre, mon cerveau est fini lui aussi. Je vous prie donc d'excuser toute l'étendue de mon ignorance. Je chercherai à relever et à croiser un maximum de réflexions qui me semblent pertinentes ; mais j'omettrai cependant nécessairement ce que j'ignore et ce dont j'ai mal mesuré l'importance.

Qui plus est, ce livre dépeint l'état de mes connaissances au moment de sa publication. Or, j'ose espérer continuer à progresser dans mon périple vers la maîtrise du bayésianisme. Pour m'accompagner dans ce périple, je vous invite à me suivre sur Twitter¹⁶ et à regarder ma chaîne YouTube Science4All, où, à partir de fin 2018, je commencerai une série de vidéos sur la formule de Bayes. Je vous invite aussi à suivre mes réflexions et celles du philosophe Thibaut Giraud, alias Monsieur Phi, sur le podcast Axiome que nous tenons ensemble. Si l'on y a pour but de discuter de tout ce qui a trait aux mathématiques, à la philosophie et aux différentes sciences, force est de constater que notre fascination commune pour les probabilités et la logique nous a maintes fois amenés à parler d'aspects du bayésianisme que je n'ai pas pu aborder dans ce livre. Comme l'interprétation bayésienne du paradoxe de Bertrand¹⁷.

À mes limitations cognitives s'ajoute le fait que ce livre se destine à un grand public et n'exige donc aucune connaissance préalable. Par conséquent, je ne ferai pas toujours preuve de la rigueur dont la *pure bayésienne* ne peut se passer — même si je ferai de mon mieux pour ne pas vous amener à des contre-sens. Ce livre reste un travail de vulgarisation.

Néanmoins, il y a de bonnes chances que vous ne compreniez pas tout. Parce que je ne voulais pas faire l'impasse sur les arguments les plus convaincants, je me suis permis de laisser traîner quelques sections de très haut niveau. Ces sections sont « étoilées ». Je préfère vous prévenir. Même les docteurs en mathématiques parmi vous risquent de sérieusement lutter pour saisir toutes les notions que je vous présenterai.

Ne vous précipitez pas dans la lecture. Prenez le temps de la réflexion. Mais n'abandonnez pas non plus. La difficulté de ce livre n'est *pas* croissante. Vous devriez pouvoir trouver du plaisir dans un chapitre sans avoir lu les précédents — même s'il est sans doute préférable de lire les chapitres dans l'ordre. Ceci n'est pas un livre de cours ; il n'y aura pas d'examen. Je ne vous demande pas de tout comprendre. Je vous conseille même vivement de sauter les passages trop compliqués et de poursuivre la lecture — quitte à revenir sur les difficultés sautées plus tard. Mon objectif n'est pas de faire de vous des experts du bayésianisme.

¹⁶Mon compte Twitter est @science_4_all.

¹⁷ Utilitarisme artificiel | Axiome 3 | T. Giraud et L.N. Hoang (2017)

Ce que j'aimerais, c'est avant tout que vous cherchiez la beauté et trouviez du plaisir dans le raisonnement bayésien, ainsi que dans les sciences utiles à la compréhension des fondements et des conséquences du bayésianisme. J'aimerais que vous vous mettiez dans la peau d'un explorateur qui part à la conquête de terres inconnues et y découvre diverses faunes, flores, cultures et paysages intrigants ; et qui ne prend pas nécessairement le temps d'apprendre toutes les subtilités de la langue des indigènes. J'aimerais que vous profitiez de *votre voyage*.

Si vous suivez mes pas, j'espère vous submerger d'enthousiasme, de fascination et de questionnements. Tel est l'objectif premier de ce livre.

Références en français

 *Le bayésianisme aujourd'hui : Fondements et pratiques* | Editions Matérielogiques | I. Drouet et collaborateurs (2016)

 *Le paradoxe de Simpson* | Science Étonnante | D. Louapre (2015)

 *La lune n'a PAS d'influence sur les naissances (Bayésianisme)* | Hygiène Mentale | C. Michel (2018)

 *Les statistiques à l'heure du Big Data* | CESP Villejuif | L.N. Hoang (2016)

 *Argent, risques et paradoxes* | Démocratie 12 | Science4All | L.N. Hoang (2017)

 *Partager un gâteau, c'est pas du gâteau !* | Démocratie 22 | Science4All | L.N. Hoang (2017)

 *La machine de Turing* | IA 4 | Science4All | L.N. Hoang (2017)

 *Les algorithmes du vivant* | TEDxSaclay | L.N. Hoang (2018)

 *Biodiversité algorithmique* | Axiome 1 | T. Giraud et L.N. Hoang (2017)

 *Optimisme probabiliste* | Axiome 2 | T. Giraud et L.N. Hoang (2017)

 *Utilitarisme artificiel* | Axiome 3 | T. Giraud et L.N. Hoang (2017)

 *Inégalité bayésienne* | Axiome 9 | T. Giraud et L.N. Hoang (2018)

Références en anglais

 *Probability Theory: The Logic of Science* | Washington University | E. Jaynes (1996)

 *Rationality: From AI to Zombies* | Machine Intelligence Research Institute | E. Yudkowsky (2015)

- ▣ *Conditionalisation and observation* | Synthese | P. Teller (1973)
 - ▣ *Dynamic Coherence and Probability Kinematics* | Philosophy of Science | B. Skyrms (1987)
 - ▣ *A Nonpragmatic Vindication of Probabilism* | Philosophy of Science | J. Joyce (1998)
 - ▣ *A Set-Partitioning Formulation for Community Healthcare Network Design in Underserved Areas* | M. Cherkesly, M.E. Rancourt and K. Smilowitz (2017)
-
- ▶ *The Feynman Series - Beauty* | Reid Gower (2011)
 - ▶ *The Universal Turing Machine* | ZettaBytes | R. Guerraoui (2016)
 - ▶ *Bayes: How one equation changed the way I think* | J. Galef (2013)
 - ▶ *Think Rationally via Bayes' Rule* | Big Think | J. Galef (2013)

Un des plus grands changements de paradigme dans ma façon de réfléchir, et chez beaucoup de gens que je connais, a été l'apprentissage de la règle de Bayes.

Julia Galef (1983-)

Nous sommes spontanément des crétins irrationnels incapables de réviser correctement nos croyances, et comprendre cette loi de Bayes peut vraiment contribuer à nous améliorer.

Thibaut Giraud (1986-)

Les statistiques bayésiennes sont difficiles dans le sens où penser est difficile.

Donald Berry (1940-)



Le théorème de Bayes

L'énigme des enfants

Revenons-en à l'énigme de l'étudiant *troll*. Un père a deux enfants. Au moins l'un d'eux est un garçon. Quelle est la probabilité que l'autre soit aussi un garçon ? Je vous invite à essayer de résoudre ce problème par vous-même. Même si vous échouez dans sa résolution, l'effort intellectuel que vous effectuez là vous sera sans doute utile pour la suite.

Je vais maintenant vous présenter une résolution de ce problème. L'approche la plus simple consiste à lister tous les cas possibles. Appelons les enfants Claude et Dominique. Il y a quatre possibilités :

- Claude et Dominique sont des garçons.
- Claude est un garçon. Dominique est une fille.
- Claude est une fille. Dominique est un garçon.
- Claude et Dominique sont des filles.

Ces quatre possibilités sont toutes équiprobables, c'est-à-dire qu'elles ont toutes la même probabilité. Enfin, pas tout fait. Les biologistes préciseront que 51 % des nouveaux-nés sont en fait des garçons — une découverte que fit Laplace en vertu de calculs bayésiens ! Mais faisons simple et supposons que la probabilité qu'un enfant soit un garçon est *a priori* de 50 %.

Sauf que l'on sait que Claude ou Dominique est un garçon. Les trois premières possibilités restent cohérentes avec cette nouvelle information, mais pas la quatrième. On peut donc barrer cette quatrième possibilité.

Mais maintenant, sachant que Claude ou Dominique est un garçon, dire que l'autre est un garçon aussi correspond ni plus ni moins à dire qu'à la fois Claude et Dominique sont des garçons. C'est le seul cas où l'un des enfants est un garçon, et l'autre aussi. Autrement dit, ce que l'on cherche à calculer, c'est la même chose que la probabilité que les deux enfants soient tous les deux des garçons, sachant que l'un des deux est un garçon.

Ceci correspond à l'une des trois possibilités restantes. Du coup, la probabilité recherchée est égale à $1/3$. Et non $1/2$! Surpris ?

La première fois que j'ai entendu cette démonstration — c'était bien avant que l'étudiant *troll* ne me la pose — je me souviens ne pas avoir été convaincu. Il n'est pas si clair que le raisonnement que l'on a donné est valide. A-t-on vraiment le droit de barrer la dernière possibilité, et de considérer que les trois premières restent équiprobables ?

Je pourrais vous sortir la tête de l'eau et vous donner la bonne façon de penser le problème dès à présent — qui consiste bien sûr en l'application de la formule de Bayes ! — mais je pense qu'il est bon de garder la tête encore un peu plus sous l'eau pour l'instant.

Le problème de Monty Hall

Passons donc au problème de Monty Hall. Ce grand classique de la théorie des probabilités est inspiré d'un jeu télévisé américain appelé *Let's Make a Deal* et présenté par un dénommé Monty Hall dans les années 1960. À la fin du jeu, un candidat doit choisir un rideau parmi trois. Derrière l'un des rideaux se trouve une voiture. Derrière les deux autres se trouvent des chèvres. Une fois le choix du rideau effectué par le candidat, Monty Hall fait monter le suspense. Parmi les rideaux que le candidat n'a pas choisis, il y a forcément au moins un rideau derrière lequel se trouve une chèvre. Monty Hall révèle un tel rideau.

Il reste alors deux rideaux. Derrière l'un d'eux se trouve la voiture. Derrière l'autre se trouve une chèvre. Monty Hall propose alors au candidat un choix. Le candidat peut conserver son rideau. Ou il peut changer de rideau. Que doit faire le candidat ? Doit-il suivre son instinct initial ? Ou doit-il ne pas pécher par obstination ?

À l'instar de l'énigme de l'étudiant *troll*, il semblerait qu'on a là un cas similaire où une possibilité a été supprimée. Il est donc tentant de penser que la position de la voiture demeure équiprobable. Il est tentant de penser que changer d'avis ou non n'a aucune importance.

Si c'est ce que vous pensez, sachez que vous faites là une erreur que de nombreux mathématiciens de premier rang ont faite avant vous. Le problème de Monty Hall a dérouté beaucoup de gens très intelligents. En 1990, quand Marilyn vos Savant proposa une solution correcte à ce problème dans la revue *Parade*, 10 000 lecteurs, dont 1 000 détenteurs d'une thèse, écrivirent à la revue en affirmant que vos Savant s'était trompée.

Même le très grand mathématicien Paul Erdős, l'homme qui publia le plus dans l'histoire des mathématiques, refusa de croire la preuve rigoureuse de vos Savant. Ce n'est qu'au moment d'observer les résultats de simulations qu'Erdős, consterné, s'inclina. Le grand Erdős ne comprenait pas la formule de Bayes. Il n'est pas le seul.

J'avais 13 ans quand j'ai découvert le problème de Monty Hall. Je ne connaissais pas la formule de Bayes. Mais il y a un raisonnement assez convaincant qui me fut accessible. En effet, si vous choisissez un rideau et si vous savez que vous ne changerez pas d'avis, tout se déroule comme si Monty Hall n'avait pas fait monter le suspense en révélant l'un des rideaux avec une chèvre. Votre probabilité de trouver la voiture est alors la même que la probabilité que le rideau initialement choisi cache la voiture. Elle est égale à $1/3$. Vous avez donc une chance sur trois de gagner si vous ne changez pas d'avis. Étrangement, je fus assez convaincu de ce résultat, mais je restais incapable de déterminer la probabilité de gagner en changeant de rideau.

Pourtant, si vous perdez en conservant votre rideau, c'est bien que c'est l'autre rideau qui cache la voiture, ce rideau que Monty Hall vous avait proposé au moment où il vous a demandé si vous vouliez changer d'avis. En fait, ce qu'il se passe, c'est que, 2 fois sur 3, une chèvre se cache derrière le rideau que vous aviez initialement choisi. Lorsque c'est le cas, au moment où il ne reste plus que deux rideaux, la voiture se retrouve forcément derrière l'autre rideau. Vous gagnez alors en changeant de rideau. Deux fois sur trois.

Les mathématiques sont indiscutables. Vous doublez vos chances de gagner la voiture en changeant de rideau ! Notre *pure bayésienne* gagnerait là deux fois plus souvent que celui qui, n'ayant pas pris la peine et le soin de bien réfléchir à ce problème, conserverait le rideau initialement choisi.

Si vous n'êtes toujours pas convaincus par ce raisonnement, je vous invite à effectuer l'expérience vous-même, à l'instar d'Erdős. Dans un excellent documentaire sur la BBC, le mathématicien Marcus du Sautoy posa le problème de Monty Hall au comédien Alan Davies. Incrédule, Alan Davies crut avoir sa chance dans un jeu de Monty Hall répété en ne changeant jamais de rideau, contrairement à Marcus du Sautoy qui lui changeait toujours de rideau. Sur 20 tentatives, Alan Davies ne gagna que 2 fois. Marcus du Sautoy gagna 16 fois. Certes, les chiffres ne correspondent là pas vraiment aux $1/3$ et $2/3$ que prédit la théorie bayésienne — la faute à la loi des petits nombres ! — mais ceci eut le mérite de convaincre Alan Davies de son erreur. Ou du moins, du fait qu'il n'avait pas vraiment compris les explications de Marcus du Sautoy.

Sur ce coup, Alan Davies n'y a perdu qu'un peu de son honneur. Il en aurait perdu un peu plus s'il savait que les pigeons étaient, eux, bien plus capables de comprendre le problème de Monty Hall¹. Mais parfois, la « mécompréhension » de la formule de Bayes a des conséquences autrement plus dramatiques. La vie de Sally Clark en a fait les frais de la manière la plus tragique qui soit.

Le procès de Sally Clark

En 1996, le nouveau-né de Sally Clark décéda au bout de deux semaines. *Bis repetita* un an plus tard, avec la mort de son second nouveau né. Sally Clark fut poursuivie en justice pour double homicide. Le pédiatre Roy Meadow fut auditionné. Il affirma que la probabilité d'une double mort infantile par cause naturelle était d'un sur 73 millions. Ce témoignage condamna Sally Clark.

Cependant, trois ans plus tard, on découvrit que le Dr Alan Williams, en charge de l'autopsie, avait manqué de rapporter les conclusions de son analyse : le second enfant était bel et bien mort de cause naturelle. Sally Clark fut enfin libérée. Mais non sans séquelles. Elle subit de graves troubles psychiatriques et mourut quatre ans plus tard par coma éthylique.

Outre le manquement du Dr Williams, la cause des maux de Sally Clark peut se réduire à une mauvaise application de la formule de Bayes, connue sous le nom de sophisme du procureur. Le juge — et peut-être vous aussi... — a confondu la probabilité d'une double mort infantile par cause naturelle et la probabilité de l'innocence de Sally Clark. Pourtant, la rareté d'un élément incriminant n'est pas nécessairement incriminante.

On aura l'occasion de revenir en longueur sur ce sophisme du procureur, puisqu'il s'agit d'un sophisme présent dans la plupart des interprétations classiques de la méthode scientifique. Mais il est bon d'insister là-dessus dès maintenant. La rareté d'un élément incriminant peut n'être qu'une conséquence de la rareté du cas du suspect. Le cas de Sally Clark est déjà exceptionnellement rare. La probabilité d'une double mort infantile est déjà extrêmement faible. Par conséquent, la probabilité d'une double mort infantile par cause naturelle ne peut être qu'extrêmement faible aussi.

En fait, les calculs approchés du professeur de mathématiques Ray Hill de Salford University montrèrent que la probabilité d'une double mort infantile par causes naturelles, bien que faible, demeurait 5 à 10 fois plus grande que la probabilité d'un double homicide. Autrement dit, la formule de Bayes force à préférer de loin l'hypothèse de morts naturelles au double homicide — au passage, Hill pointa aussi une erreur grossière dans le calcul du pédiatre Meadow qui n'avait pas pris en compte la corrélation entre les morts des deux nouveau-nés, ce qui rendait la conclusion bayésienne encore plus favorable à l'hypothèse des morts naturelles.

¹  *Les pigeons, rois des cons ?* Science de Comptoir | V. Delattre et M. Guillet (2016)

Le bayésianisme jugé illégal

Le procès de Sally Clarke est un échec cuisant de la justice britannique. Mais au lieu d'une prise de conscience de l'importance de la formule de Bayes, cet épisode tragique conduisit au contraire à une méfiance accrue des statistiques. À tel point qu'en 2010, un juge britannique évinça le théorème de Bayes des cours de justice. Être bayésien devant le juge est devenu illégal ! La formule du savoir est interdite des tribunaux !

Mais il est difficile d'en vouloir à ce juge. Si même Erdős a du mal à appliquer la formule de Bayes, est-il vraiment raisonnable de demander à des juges et des jurés de fonder leur raisonnement sur cette équation ?

Si la *pure bayésienne*, elle, est capable de naviguer à travers les méandres d'un complexe système judiciaire, il faut certainement faire attention à ne pas envahir les tribunaux de flopées de statistiques que personne n'est vraiment capable d'interpréter, ou dont le taux de contre-sens est trop élevé. Dans un article publié dans *The Annual Review of Statistics and Its Applications*, Fenton, Neil et Berger font d'ailleurs la remarque suivante : « il y a une autre raison rarement articulée mais désormais prédominante à l'utilisation limitée [de la formule de Bayes] : c'est parce que la plupart des exemples de l'approche bayésienne simplifient à outrance les arguments légaux qu'elle modélise, dans le but de permettre d'effectuer les calculs à la main ». Conscient de cette difficulté, les auteurs proposent une théorie bayésienne plus sophistiquée, celle des réseaux bayésiens, dont on reparlera plus tard. Pour l'heure, force est de constater l'irrationalité de notre système judiciaire et l'extrême difficulté de la corriger.

Le calcul bayésien simplifié à outrance est très limité. Quant au calcul bayésien correct, il est bien trop complexe pour être effectué, même par les meilleurs d'entre nous. L'objectif de ce livre n'est donc absolument pas de faire de vous des machines à appliquer la formule de Bayes. Ce serait peine perdue.

Cependant, je pense que les exemples simples sont à la portée de tous — à condition de vraiment prendre le temps de la réflexion — et qu'ils peuvent servir de référence ou d'entraînement pour des raisonnements approximativement bayésiens dans les cas plus pragmatiques. Je ne ferai pas de vous des *purs bayésiens*. Mais j'ose espérer faire de vous de meilleurs penseurs. J'ose espérer vous aider à (être capable de) bayesianiser vos intuitions.

J'espère aussi vous montrer notre incapacité à intuiter les calculs probabilistes. Or, selon la *pure bayésienne*, *être rationnel, c'est obéir aux lois des probabilités*. Quand on contemple notre confusion devant la formule de Bayes, tout en se convaincant qu'elle est la solution à tous nos maux épistémologiques, on est certainement plus à même de douter de toutes nos convictions. C'est, je pense, la bonne réaction à avoir quand on découvre à quel point on est de mauvais penseurs. Il faut accepter le fait que nous sommes constamment en excès de confiance. Il nous faut absolument diminuer nos crédences en nos intuitions et en nos raisonnements non-bayésiens.

Le théorème de Bayes

Assez divagué. Il est enfin temps de vous révéler mon équation préférée des mathématiques. Présentons la formule de Bayes. Pour cela, on va prendre un quatrième exemple qui nous vient du monde médical.

Imaginez qu'un diagnostic affirme que vous êtes atteint d'Ebola. Or vous savez que vous rentrez de vacances au Nigeria. Vous demandez naturellement quel est le degré de fiabilité du diagnostic. On vous répond que lorsqu'une personne est saine, le taux de diagnostics corrects est de 90%. Devez-vous commencer à écrire votre testament ?

La réponse de la *pure bayésienne* est univoque : non. Même en Afrique noire où cette terrible maladie a fait le plus de dommages, pas plus d'une personne sur 10 000 a été victime d'Ebola. Du coup, *a priori*, vous qui n'avez que brièvement séjourné au Nigeria n'avez pas plus d'une chance sur 10 000 d'avoir contracté la maladie. On peut succinctement noter $\mathbb{P}[\oplus]$ cette probabilité. On l'appelle aussi *l'a priori*.

Supposons maintenant que vous avez appris que le diagnostic médical était défavorable. Ce qui compte désormais est la probabilité d'avoir Ebola sachant que le diagnostic est défavorable, que l'on va noter $\mathbb{P}[\oplus|\textcolor{red}{\checkmark}]$, où $\textcolor{red}{\checkmark}$ est un symbole qui indique que le diagnostic est défavorable. Par opposition, on utilisera le symbole $\textcolor{green}{\checkmark}$ pour parler du cas où le test médical indique que vous n'avez pas Ebola.

Que signifie vraiment la probabilité dite *conditionnelle* $\mathbb{P}[\oplus|\textcolor{red}{\checkmark}]$? Le postulat fondamental de la théorie des probabilités suppose que cette probabilité conditionnelle est reliée aux probabilités des événements \oplus et $\textcolor{red}{\checkmark}$ comme suit :

$$\mathbb{P}[\oplus|\textcolor{red}{\checkmark}] = \frac{\mathbb{P}[\oplus \text{ et } \textcolor{red}{\checkmark}]}{\mathbb{P}[\textcolor{red}{\checkmark}]}.$$

Autrement dit, la probabilité d'être malade sachant que l'on a eu un diagnostic défavorable n'est autre que la proportion des cas où l'on est malade et l'on a un diagnostic défavorable, comparée à l'ensemble de tous les cas où le diagnostic est défavorable.

Il est bon de signaler que, de nos jours, même le plus anti-bayésien des statisticiens accepte ce postulat. En fait, on peut y voir là la définition de la probabilité conditionnelle. Et comme toute définition, elle ne peut pas être fausse. Cependant, on peut se demander si c'est une définition pertinente (et utile) des probabilités conditionnelles. En particulier, on peut se demander si elle est conforme à notre langage naturel, et à la manière dont il faudrait réfléchir. Et bien justement, l'acte de foi de la *pure bayésienne* est que, non seulement cette définition est proche de notre langage naturel, mais surtout, il s'agit là de la bonne manière de penser. *Être bayésien, c'est faire du langage des probabilités conditionnelles le fondement de tout savoir.*

En suivant les pas de notre *pure bayésienne*, acceptons donc que la probabilité conditionnelle $\mathbb{P}[\text{⊕}| \text{⊖}]$ décrit bel et bien la probabilité d'être malade sachant que le diagnostic est défavorable. Cependant, ce n'est pas la quantité qui vous a été communiquée. Le chiffre de 90 % qui vous a été communiqué, c'est la probabilité d'un diagnostic correct lorsque vous n'avez pas Ebola. Autrement dit, 90 % est la probabilité $\mathbb{P}[\text{⊖}| \text{⊕}]$ d'avoir un diagnostic favorable sachant que vous n'avez pas contracté Ebola (le symbole ⊕ indique que vous êtes sain). Les 10 % restants correspondent alors à la probabilité $\mathbb{P}[\text{⊖}| \text{⊖}]$ d'un diagnostic défavorable sachant que vous n'avez pas contracté Ebola.

Pour déterminer votre probabilité d'être malade sachant le diagnostic défavorable, il nous faut prouver et utiliser la théorème de Bayes. Pour y arriver, écrivons la définition de la probabilité inverse $\mathbb{P}[\text{⊖}| \text{⊕}] = \mathbb{P}[\text{⊖ et } \text{⊕}] / \mathbb{P}[\text{⊕}]$. Le remarquez-vous ? Le numérateur est le même que dans la définition de la probabilité inverse $\mathbb{P}[\text{⊕}| \text{⊖}]$! En particulier, la probabilité des deux événements simultanés s'écrit $\mathbb{P}[\text{⊖ et } \text{⊕}] = \mathbb{P}[\text{⊕}] \mathbb{P}[\text{⊖}| \text{⊕}]$. Ceci revient à dire que la probabilité d'être malade et d'avoir un diagnostic défavorable est égale à la probabilité de d'abord tomber malade, puis d'avoir un diagnostic défavorable sachant que l'on est déjà malade.

On a quasiment fini notre preuve du théorème de Bayes. Il nous suffit d'utiliser l'équation ci-dessus dans la définition de la probabilité conditionnelle $\mathbb{P}[\text{⊕}| \text{⊖}]$. On obtient alors la formule plus importante de la philosophie du savoir que présente ce livre, la formule de Bayes. Prenez le temps d'en savourer l'élégance calligraphique et le *pattern* suivi par les symboles.

$$\mathbb{P}[\text{⊕} | \text{⊖}] = \frac{\mathbb{P}[\text{⊖} | \text{⊕}] \mathbb{P}[\text{⊕}]}{\mathbb{P}[\text{⊖}]}$$

Autrement dit, pour déterminer la probabilité d'avoir Ebola sachant que le test est défavorable, il nous faut multiplier la probabilité que le test soit défavorable sachant que l'on a Ebola (cela demande un peu d'imagination !) par la probabilité d'avoir Ebola *a priori*, puis diviser tout ça par la probabilité d'avoir un diagnostic défavorable.

Comme annoncé dans le premier chapitre introductif, tout ce qu'il faut savoir faire, ce sont des multiplications et des divisions ! Quoi de plus simple ?

Ce qui rend cette formule si difficile à comprendre, ce ne sont bien sûr pas les calculs qu'elle requiert, mais l'interprétation de chacun de ses termes — en tout cas dans les exemples simplistes de ce chapitre. Il est très facile de faire un contre-sens au moment de penser à ces termes. Je ne peux que vous inviter à longuement les méditer.

Les composants de la formule de Bayes

Dans l'expression de droite, la probabilité $\mathbb{P}[\Theta]$ est l'*a priori*. C'est ce que l'on aurait pu (ou dû) penser avant d'avoir reçu le résultat du diagnostic. Dans notre cas, on l'a estimée en comparant le nombre de cas recensés d'Ebola à la population des pays d'Afrique noire. Mais il ne s'agit là que d'une grossière estimation. D'ailleurs, on n'a même pas pris en compte la durée du séjour au Nigeria, qui, à n'en pas douter, est un facteur déterminant sur la croyance *a priori*. Tout aussi importante est la fréquence d'interaction avec des individus locaux, et l'exposition à des individus malades. Toutes ces contributions sont incroyablement difficiles à quantifier. On se contentera de notre grossière estimation ici.

L'autre quantité au numérateur de l'expression de droite est la probabilité $\mathbb{P}[\text{R}|\Theta]$ de recevoir un diagnostic médical défavorable sachant que l'on a contracté Ebola. Ce terme requiert de l'imagination. Il demande de s'extirper du monde réel et d'imaginer une sorte de monde alternatif où l'on sait que l'on a contracté Ebola. Dans ce monde alternatif, aura-t-on un diagnostic médical défavorable ? Telle est la question dont $\mathbb{P}[\text{R}|\Theta]$ est la réponse.

Contrairement à nous, la *pure bayésienne* est non seulement capable de se mettre dans la peau des autres et d'imaginer ce qu'ils imaginent ; mais c'est en fait ce qu'elle fait à longueur de journée ! C'est tout l'art des fameuses *expériences de pensée*. Ces expériences sont en fait indispensables à la philosophie bayésienne. Sans ces expériences, on serait incapables d'estimer des termes comme $\mathbb{P}[\text{R}|\Theta]$. On serait ensuite incapable d'appliquer la formule de Bayes. Selon la *pure bayésienne*, on serait donc irrationnel.

Malheureusement, il arrive bien trop souvent que certains refusent de pleinement accepter temporairement les prémisses contre-intuitives d'une théorie pour en explorer les conséquences. Trop souvent, les débats opposent des partis qui ne voient le monde que sous le prisme de leur épistémologie, de leurs modèles de la réalité, de leur théologie ou de leurs principes moraux. Sans fondements communs, de tels débats sont alors voués à dégénérer. Trop souvent, les débats sautent l'étape du calcul de termes comme $\mathbb{P}[\text{R}|\Theta]$.

De telles quantités sont communément appelées *vraisemblances* par les statisticiens. Cependant, cette terminologie me semble dangereuse car elle risque fortement de conduire à des contre-sens. En effet, il est bon de garder en tête que la vraisemblance est la vraisemblance des données observées étant donné certaines hypothèses sur le monde, laquelle est très distincte de la vraisemblance des hypothèses sachant les données. Pour éviter d'éventuelles confusions, il me semble préférable de parler de *termes d'expériences de pensée* — même s'il m'arrivera de céder à la terminologie communément admise.

Enfin, il reste le dénominateur de l'expression de droite, la probabilité $\mathbb{P}[\text{R}]$ d'un diagnostic défavorable. Ce terme est un monstre. C'est la plus grosse difficulté de la formule de Bayes. C'est ce terme qui fait passer des nuits blanches à de

nombreux chercheurs en probabilité (ou à leurs intelligences artificielles). Ils l'appellent la marginale ou la fonction de partition. C'est ce terme qu'il m'est le plus difficile à visualiser et à comprendre — même si dans certains cas simplistes de ce chapitre, cette quantité est remarquablement simple à déterminer.

Pour calculer la probabilité $\mathbb{P}[\text{D}]$ d'un diagnostic défavorable, il nous faut distinguer deux cas : le cas où le diagnostic est défavorable parce que l'on a Ebola, et le cas où le diagnostic est défavorable à cause de l'imperfection du diagnostic. Et pour chacun des deux cas, il nous faut multiplier *l'a priori* du cas par la probabilité que le cas imaginé conduise à un diagnostic défavorable. Autrement dit, on va utiliser ce que l'on appelle parfois la *loi des probabilités totales*² :

$$\mathbb{P}[\text{D}] = \mathbb{P}[\text{D}|\text{E}] \mathbb{P}[\text{E}] + \mathbb{P}[\text{D}|\text{G}] \mathbb{P}[\text{G}].$$

En particulier, la fonction de partition $\mathbb{P}[\text{D}]$ requiert le calcul de deux expériences de pensée qui correspondent à deux cas différents. C'est ce qui rend son calcul difficile. Le bayésien doit développer cette subtile gymnastique cérébrale, qui consiste à raisonner dans différentes versions mutuellement incompatibles de la réalité. C'est sans doute cela qui rend la formule de Bayes si difficile à appliquer et à comprendre.

Bayes au secours du diagnostic

On peut enfin combiner la loi des probabilités totales à la formule de Bayes, ce qui nous donne

$$\mathbb{P}[\text{E}|\text{D}] = \frac{\mathbb{P}[\text{D}|\text{E}] \mathbb{P}[\text{E}]}{\mathbb{P}[\text{D}|\text{E}] \mathbb{P}[\text{E}] + \mathbb{P}[\text{D}|\text{G}] \mathbb{P}[\text{G}]}.$$

On connaît maintenant presque toutes les quantités de la formule de droite. On a vu que $\mathbb{P}[\text{E}]$ pouvait être estimé à 1 sur³ 10 000, d'où l'on déduit que $\mathbb{P}[\text{G}]$ est au moins 9 999 sur 10 000. Et puis, on remarque que $\mathbb{P}[\text{D}|\text{G}]$ correspond au taux d'erreurs du test pour des individus sains, dont on a vu qu'il était de 10 %. Enfin, il reste le terme $\mathbb{P}[\text{D}|\text{E}]$ qui est la fiabilité du test pour des individus atteints d'Ebola. Constatons simplement que ce terme, qui est une probabilité, ne peut pas être supérieur à 1. En combinant tout ceci, on peut enfin conclure avec le calcul suivant :

$$\mathbb{P}[\text{E}|\text{D}] \approx \frac{1 \cdot 0,0001}{1 \cdot 0,0001 + 0,1 \cdot 0,9999} \approx 0,001.$$

²Comme la formule de Bayes, la loi des probabilités totales se déduit de la définition des probabilités conditionnelles, ainsi que du fait que la probabilité de deux événements incompatibles est la somme des probabilités des événements.

³En fait, il s'agit d'une borne supérieure.

Autrement dit, maintenant que vous connaissez le résultat défavorable du diagnostic, vous nous donnez une probabilité de moins de 1 sur 1 000 d'être vraiment atteint d'Ebola. Cette probabilité reste incroyablement négligeable. Ce n'est donc pas la peine de vous mettre à écrire votre testament.

Que s'est-il passé ? Pourquoi le résultat final est-il si faible ? Comment guider notre intuition pour qu'elle puisse sentir la petitesse du résultat sans avoir à faire confiance aveuglément aux calculs ? Je vous invite vraiment à y réfléchir de vous-même.

Il est utile de divaguer quelque peu sur la fonction de partition $\mathbb{P}[\textcolor{red}{\checkmark}]$ au dénominateur. On a vu que cette quantité se décomposait en deux cas : le cas où on a été victime d'Ebola, et le cas où on a été victime d'un mauvais diagnostic. Ces deux cas n'ont absolument pas la même probabilité. En fait, le cas où on a été victime d'Ebola est 1 000 fois moins probable que le cas où on a été victime d'un mauvais diagnostic. La différence entre ces deux cas est si grande, qu'il est raisonnable de négliger le cas où on a été victime d'Ebola dans le calcul de la fonction de partition.

À ce moment-là, la formule de Bayes n'est plus qu'un ratio entre un numérateur inchangé qui calcule le cas où on a été victime d'Ebola, et un dénominateur qui calcule le cas où on a été victime d'un mauvais diagnostic. La formule de Bayes compare donc les deux alternatives qui expliquent le résultat défavorable du diagnostic. La petitesse de notre résultat final traduit alors le fait que le cas où on a été victime d'Ebola est beaucoup plus improbable que le cas où on a été victime d'un mauvais diagnostic.

Une autre façon d'interpréter la formule de Bayes est comme un transfert de crédence. Ainsi, les théories dont le terme d'expérience de pensée est faible perdent leurs crédences au profit de celles dont le terme d'expérience de pensée est important. Dans notre cas, $\mathbb{P}[\textcolor{red}{\checkmark}|\textcolor{green}{\checkmark}] = 10\%$ est 10 fois plus petit que $\mathbb{P}[\textcolor{red}{\checkmark}|\textcolor{black}{\checkmark}] \approx 1$. La crédence de $\textcolor{black}{\checkmark}$ augmentera alors au profit de celle de $\textcolor{green}{\checkmark}$ d'un facteur 10. Cependant, *a priori*, $\textcolor{black}{\checkmark}$ était environ 10 000 fois moins probable. *A posteriori*, $\textcolor{black}{\checkmark}$ devient alors seulement 1 000 fois moins probable⁴.

En pratique, de nos jours, le personnel médical va chercher à vous protéger d'une peur bleue inutile, en combinant divers tests médicaux aussi indépendants que possibles, et ne vous annoncera un diagnostic défavorable que si les résultats de ces nombreux tests sont effectivement défavorables. Autrement dit, le personnel médical va essayer de diminuer autant que possible la probabilité $\mathbb{P}[\textcolor{red}{\checkmark}|\textcolor{green}{\checkmark}]$ d'un diagnostic défavorable pour des individus pourtant sains.

⁴Formellement, ceci correspond au calcul suivant :

$$\frac{\mathbb{P}[\textcolor{black}{\checkmark}|\textcolor{red}{\checkmark}]}{\mathbb{P}[\textcolor{green}{\checkmark}|\textcolor{red}{\checkmark}]} = \frac{\mathbb{P}[\textcolor{red}{\checkmark}|\textcolor{black}{\checkmark}] \mathbb{P}[\textcolor{black}{\checkmark}]}{\mathbb{P}[\textcolor{red}{\checkmark}|\textcolor{green}{\checkmark}] \mathbb{P}[\textcolor{green}{\checkmark}]} = 10 \cdot \frac{\mathbb{P}[\textcolor{black}{\checkmark}]}{\mathbb{P}[\textcolor{green}{\checkmark}]}.$$

Bayes au secours de Sally Clark

Pour mieux comprendre la formule de Bayes, appliquons-la maintenant au cas de Sally Clark. Ce que l'on aimerait savoir, c'est la probabilité de son innocence, sachant que ses deux nouveau-nés sont morts. Écrivons la formule de Bayes.

$$\mathbb{P}[\text{活着} | \text{双死}] = \frac{\mathbb{P}[\text{双死} | \text{活着}] \mathbb{P}[\text{活着}]}{\mathbb{P}[\text{双死} | \text{活着}] \mathbb{P}[\text{活着}] + \mathbb{P}[\text{双死} | \text{凶手}] \mathbb{P}[\text{凶手}]}$$

Comme tout à l'heure, il y a trois quantités auxquelles il nous faut réfléchir : l'*a priori* $\mathbb{P}[\text{活着}]$ sur l'innocence de Sally Clark, et les termes d'expériences de pensée $\mathbb{P}[\text{双死} | \text{活着}]$ et $\mathbb{P}[\text{双死} | \text{凶手}]$ qui conduisent aux morts des deux nouveau-nés en fonction de l'innocence et de la culpabilité de Sally Clark.

Commençons par l'*a priori* sur l'innocence de Sally Clark. Il se doit d'être très important. Cet *a priori*, c'est la probabilité qu'une personne tirée au hasard n'ait pas tué ses deux nouveau-nés. Or, la quasi-totalité des êtres humains n'ont pas tué deux de leurs nouveau-nés ! En fait, le docteur Hill estimait la probabilité *a priori* de la culpabilité de Sally Clark à environ 1 sur 500 millions !

Voilà qui justifie la *présomption d'innocence*. En l'absence d'éléments à conviction et pour des crimes graves, toute personne est beaucoup plus probablement innocente que coupable. La présomption d'innocence est donc le bon *a priori* à avoir ! Cependant, il ne faut pas pousser la présomption d'innocence au-delà de son champ d'applicabilité. Elle correspond uniquement à l'*a priori* en l'absence d'éléments à conviction. Si beaucoup de tels éléments semblent incriminer un suspect, il est possible que la probabilité de son innocence chute en deçà de sa probabilité de culpabilité. Il est aussi possible que ces éléments ne suffisent pas.

Notre *pure bayésienne* nous forcerait à appliquer la formule de Bayes pour mieux comprendre le niveau de suspicion adéquat au vu des éléments à charge. Et surtout, à moins de devoir expliquer les calculs qui ont mené à ce niveau de suspicion, elle n'invoquerait pas seulement la présomption d'innocence pour conclure. Car cette notion suppose l'absence d'éléments à charge.

Passons aux *termes d'expériences de pensée*. La probabilité $\mathbb{P}[\text{双死} | \text{活着}]$ de la double mort des nouveau-nés sachant que Sally Clark est innocente correspond au cas de morts par causes naturelles. C'est le chiffre (sous-estimé) d'une chance sur 70 millions dont on a parlé plus haut. Enfin, la probabilité $\mathbb{P}[\text{双死} | \text{凶手}]$ de la double mort des nouveau-nés sachant que Sally Clark est coupable est égale à 1.

On peut maintenant combiner tout ceci, ce qui nous fait faire le calcul suivant :

$$\mathbb{P}[\text{活着} | \text{双死}] \approx \frac{1/70 \text{ millions} \cdot 1}{1/70 \text{ millions} \cdot 1 + 1 \cdot 1/500 \text{ millions}} \approx 0,88.$$

Autrement dit, même avec une sous-estimation de la double mort par causes naturelles, la probabilité d'innocence de Sally Clark reste très élevée. Il reste

beaucoup plus probable qu'elle soit innocente que coupable. Il ne semble alors pas raisonnable de condamner Sally Clark — même si, comme on le verra dans le dernier chapitre, cette décision morale sort du cadre du bayésianisme... .

Jusque-là, on a appliqué la formule de Bayes à deux cas concrets. Comme souvent en mathématiques appliquées, ces cas concrets sont en fait déjà trop compliqués pour être bien compris. Après tout, l'*a priori* de l'Ebola $\mathbb{P}[\Theta]$ et le calcul de la probabilité de causes naturelles $\mathbb{P}[\Theta | \text{EBOLA}]$ sont en fait très difficiles à estimer correctement. Il ne faut pas perdre de vue que nos résultats finaux sont donc forcément approximatifs. « Tous les modèles sont faux. »

La *pure bayésienne* n'a ainsi aucune certitude sur les résultats numériques que l'on a obtenus. Elle a même l'intelligence de calculer ses crédences en les différents résultats qu'elle a obtenus pour différentes valeurs de $\mathbb{P}[\Theta]$ et $\mathbb{P}[\Theta | \text{EBOLA}]$. Mais si un juge lui demande un résultat unique, elle calculera la moyenne des résultats qu'elle a obtenus, pondérés par leur crédence. Elle constatera aussi que la probabilité d'avoir Ebola au vu d'un diagnostic défavorable ne peut qu'être très faible, alors que le résultat du cas de Sally Clark est moins robuste, et donc moins conclusif.

L'énigme des enfants enfin résolue !

Je vous propose enfin d'appliquer la formule de Bayes au problème de l'étudiant *troll*. Souvenez-vous. Parmi deux enfants, Claude et Dominique, il y a au moins un garçon. Quelle est la probabilité que l'autre soit un garçon aussi ? On a déjà vu que ceci revenait à la probabilité que Claude et Dominique soient des garçons, sachant que Claude ou Dominique est un garçon.

Pour simplifier les expressions, notons $C\sigma$ le fait que Claude est un garçon et $D\sigma$ le fait que Dominique est un garçon. On utilisera aussi le symbole \varnothing pour le cas où Claude ou Dominique est une fille. La formule de Bayes s'écrit alors

$$\mathbb{P}[C\sigma \text{ et } D\sigma | C\sigma \text{ ou } D\sigma] = \frac{\mathbb{P}[C\sigma \text{ ou } D\sigma | C\sigma \text{ et } D\sigma] \mathbb{P}[C\sigma \text{ et } D\sigma]}{\mathbb{P}[C\sigma \text{ ou } D\sigma]}$$

L'*a priori* $\mathbb{P}[C\sigma \text{ et } D\sigma]$ est la probabilité *a priori* que Claude et Dominique soient tout deux des garçons. Cette probabilité est égale à $1/4$.

Le terme d'*expérience de pensée* $\mathbb{P}[C\sigma \text{ ou } D\sigma | C\sigma \text{ et } D\sigma]$ est la probabilité que Claude ou Dominique soit un garçon sachant que tous deux sont des garçons. Or le fait que Claude ou Dominique est un garçon est une conséquence du fait que tous deux sont des garçons. Cette probabilité est donc égale à 1.

Enfin, il reste la *fonction de partition* $\mathbb{P}[C\sigma \text{ ou } D\sigma]$. Elle correspond à 3 cas sur 4 des combinaisons de sexes possibles de Claude et Dominique.

On peut d'ailleurs justifier la validité de notre raisonnement du début du chapitre. Si les trois hypothèses garçon-garçon, garçon-fille et fille-garçon demeurent équiprobables après élimination de l'hypothèse fille-fille, c'est parce que les termes d'expérience de pensée de ces trois hypothèses sont toutes égales à 1. Il n'y a donc pas eu de transfert de crédence entre ces trois hypothèses⁵.

On a ainsi toutes les données nécessaires. Il ne nous reste plus qu'à effectuer le calcul. On obtient alors $1 \cdot 1/4 / 3/4$, ce qui est égal à $1/3$. C'est bien la réponse que l'on avait calculée « à la main » et peu rigoureusement au début de ce chapitre.

Quelques mots d'encouragement

Les calculs bayésiens de ce chapitre ne sont pas évidents. À chaque fois, il nous a fallu toute une page pour les expliciter. Ceci a de quoi effrayer même les plus matheux d'entre vous. Oui, la formule de Bayes est difficile à appliquer, même dans les cas les plus simples. Il est encore plus difficile de la comprendre. À l'instar des mathématiques en général, le niveau d'abstraction et de sophistication de cette formule a de quoi rebouter les moins courageux d'entre nous.

Je ne peux que vous encourager à ne pas baisser les bras. La formule de Bayes est dure à comprendre pour tout le monde. Même de grands mathématiciens ont bien du mal à l'appliquer à un cas aussi simpliste que le problème de Monty Hall. Même avec la meilleure volonté qui soit, vous ne comprendrez pas entièrement la formule de Bayes. Mais vous pouvez grandement progresser dans sa compréhension. Et pour cela, il faut lutter. Il faut se battre. Il ne faut pas abandonner. Le prix à payer est un énorme effort intellectuel ; mais la récompense est absolument formidable. Selon la *pure bayésienne*, la faculté à enfin bien raisonner (assez) juste est au bout du périple.

Toutefois, la lecture de ce livre ne suffira pas. Les mathématiques s'apprennent en pratiquant, en laissant l'esprit manipuler des objets abstraits et en cherchant constamment à clarifier par soi-même les concepts mathématiques. Il faut penser mathématique pour devenir bon en mathématiques. Je vous invite donc à faire et refaire les raisonnements bayésiens de ce chapitre à vos heures perdues, sous la douche, dans un transport ou en randonnée. Et quand vous vous sentirez prêts, je vous invite aussi à vous attaquer au problème de Monty Hall, puis à la deuxième partie de l'éénigme de l'étudiant *troll* — la bonne réponse est 13/27.

Luttez donc. Mais le conseil le plus important que je me permettrais de vous donner est de constamment chercher à y trouver du plaisir. C'est sur ce plaisir que j'insisterai le plus dans les chapitres à venir. En particulier, il y a quelque

⁵Une variante du problème consiste à tirer l'un des enfants au hasard, et à constater qu'il s'agit d'un garçon. À ce moment-là, les termes d'expériences de pensée des cas garçon-fille et fille-garçon ne sont plus égaux à 1, ce qui crée un transfert de crédence et conduit à une conclusion différente. La vidéo suivante parle de cette variante :

▶ *Le problème des deux enfants* | Math un peu ça (2017)

chose d'absolument fascinant dans le fait que la formule de Bayes soit à la fois si compacte, si piégieuse et si utile pour comprendre le monde. Par exemple, on a déjà vu qu'elle expliquait l'inconclusivité d'excellents tests médicaux et la pertinence de la présomption d'innocence en droit ! Et ce n'est qu'un début !

L'élégance des équations bayésiennes et de leurs conséquences m'a conduit à d'innombrables réflexions jouissives. La philosophie du savoir qui en ressort n'a cessé de me remplir de joie et de bonheur. C'est pour ça que j'ai fini par considérer que la formule de Bayes est la plus belle équation des mathématiques.

Références en français

- ▶ *La loi de Bayes (1/2) - Argument frappant* | Monsieur Phi | T. Giraud (2016)
- ▶ *La loi de Bayes (2/2) - Argument frappant* | Monsieur Phi | T. Giraud (2016)
- ▶ *La lune n'a PAS d'influence sur les naissances (Bayésianisme)* | Hygiène Mentale | C. Michel (2018)
- ▶ *Le paradoxe des trois portes* | Math&Magique (2016)
- ▶ *Quart d'Heure Insolite : le paradoxe de Monty Hall* | R. Taillet (2015)
- ▶ *Le problème des deux enfants* | Math un peu ça (2017)
- ▶ *Les pigeons, rois des cons ?* Science de Comptoir | M. Guillet, I. Hamchiche et V. Delattre (2016)

Références en anglais

- 🚩 *Are Birds Smarter Than Mathematicians? Pigeons (*Columba livia*) Perform Optimally on a Version of the Monty Hall Dilemma* | Journal of Comparative Psychology | W. Herbranson and J. Schroeder (2010)
- 🚩 *Bayes and the law* | Annual Review of Statistics and Its Application | N. Fenton, M. Neil and D. Berger (2016)
- 🌐 *Conditional Probabilities: Know what you Learn* | Science4All | L.N. Hoang (2013)
- 🌐 *A Formula for justice* | The Guardian (online) | A. Saini (2011)
- ▶ *A visual guide to Bayesian thinking* | J. Galef (2015)
- ▶ *Your brain is not a Bayes net (and why that matters)* | J. Galef (2016)
- ▶ *Fundamentals: Bayes' Theorem* | Critical Thinking | Wireless Philosophy | I. Olasov (2016)
- ▶ *Alan and Marcus go forth and multiply* | BBC (2009)
- ▶ *Monty Hall Problem* | Numberphile | L. Goldberg (2014)
- ▶ *The Monty Hall Problem* | Singingbanana | J. Grime (2009)
- ▶ *The Monty Hall Problem - Explained* | AsapSCIENCE (2012)
- ▶ *Are you REALLY sick? (false positives)* | Numberphile | L. Goldberg (2016)
- ▶ *The Bayesian Trap* | Veritasium | D. Muller (2017)

La logique nous amène plus près des cieux que n'importe quelle autre étude.

Bertrand Russell (1872-1970)

Toute extension autorisée de la logique aristotélique à une théorie de la plausibilité est isomorphe à la théorie des probabilités bayésiennes.

Petri Myllymäki

3

Logiquement...

Deux modes de raisonnement

Imaginez qu'on vous affirme : « si une carte a une dame d'un côté, alors l'autre côté est bleu ». Autrement dit, on considère l'hypothèse « $\text{dame} \rightarrow \text{bleu}$ ». Vous avez quatre cartes devant vous. La première est une dame, la seconde est un dix, la troisième une carte bleue et la dernière est rouge. Quelle(s) carte(s) faut-il retourner pour tester l'hypothèse ?



Figure 3.1. De gauche à droite, les cartes sont dame, dix, bleue et rouge.

Cette question fut posée à de nombreux sujets. Seulement 4 % des milliers de personnes interrogées ont su donner la bonne réponse. Je vous invite à y réfléchir. Ne tombez pas dans le(s) piège(s). Puis, une fois votre réponse choisie, je vous invite fortement à aller voir l'excellente vidéo d'Hygiène Mentale¹.

¹ *Raisonnez de façon correcte (Testez votre logique)* | Hygiène Mentale | C. Michel (2016)

La philosophie des sciences et du savoir se divise traditionnellement en deux types de raisonnements très distincts : la déduction et l'induction. Il est souvent enseigné que le chercheur scientifique doit combiner ces deux types de raisonnement dans une approche alors appelée « hypothético-déductive ». C'est l'approche que vous devriez avoir en partie adoptée pour l'éénigme ci-dessus. En partant d'une hypothèse, vous devriez avoir déduit ses conséquences, et vous devriez les avoir testées.

En bon mathématicien, je suis vite tombé amoureux de la partie déductive de ce raisonnement. Mais, en bon mathématicien, j'ai toujours trouvé la partie inductive de la *méthode scientifique* insatisfaisante. Je l'ai souvent trouvée *ad hoc*, bancale et très distante du quotidien des chercheurs. Pire, j'ai souvent eu l'impression que de nombreux tenants de la *méthode scientifique* avaient un objectif politique dans le choix de leurs descriptions de la *méthode scientifique*, à savoir distinguer les *sciences* des *pseudo-sciences*. J'ai souvent senti que, derrière leur « définition » de la science, se cachait une espèce d'hooliganisme scientifique, c'est-à-dire une rationalisation fallacieuse motivée par un désir de défense de la communauté scientifique. En particulier, cet hooliganisme a tendance à balayer sous le tapis la difficulté du problème de l'induction, qui me semble être un problème pourtant majeur.

Ne vous y méprenez pas. La distinction entre la fiabilité des sciences et celle des pseudo-sciences est importante à faire. J'insisterai longuement dessus dans les chapitres suivants. Mais je préfère le dire clairement dès maintenant. Mon inconfort avec la « *méthode scientifique* » n'entraîne pas mon rejet des conclusions scientifiques — et encore moins mon acceptation des alternatives pseudo-scientifiques. En particulier, nous verrons plus tard qu'un principe bayésien nous permet d'accorder une très grande crédence aux consensus scientifiques.

Mais avant d'en arriver là, revenons-en à la distinction entre déduction et induction. Cette distinction semblera évidente à tout scientifique bien formé aujourd'hui. Pourtant, curieusement, la *pure bayésienne* ne fait pas cette distinction. Pour elle, il n'y a qu'un seul type de raisonnement. Il n'y a que la formule de Bayes. En particulier, tout notre système de déduction n'est qu'un cas particulier de la formule de Bayes, tandis que l'induction, telle qu'elle est souvent présentée, est une approximation fallacieuse de la formule de Bayes.

Le jour où je me suis rendu compte de ceci, j'ai été stupéfait. C'est cette découverte, plus que toute autre, qui m'a convaincu de me lancer dans l'écriture de ce livre !

Dans ce chapitre, on va s'intéresser uniquement à la déduction. On abordera l'induction dans le chapitre suivant. Ici, on verra que cette déduction, aussi appelée logique, est en fait beaucoup plus sophistiquée, contre-intuitive et obscure que ce qu'on pourrait croire naïvement. En fait, contrairement à ce que même des scientifiques très éduqués pourraient penser, il existe *plusieurs* logiques déductives. Et nous verrons que la logique bayésienne n'a rien à envier à celles que l'on enseigne aujourd'hui.

Les règles de la logique

L'exemple classique de déduction logique est le syllogisme d'Aristote. Ce syllogisme considère les deux prémisses suivantes :

- Tous les hommes sont mortels.
- Socrate est un homme.

Aristote affirme que de ces deux prémisses découle la conclusion suivante² :

- Par conséquent, Socrate est mortel.

La logique d'Aristote paraît implacable. Elle paraît même naturelle, indiscutablement vraie. Il y a plusieurs années, quand un ami me mit au défi de douter du syllogisme d'Aristote, je m'avouai vaincu.

Cependant, ce syllogisme d'Aristote a inspiré de nombreux philosophes, logiciens et mathématiciens, qui l'ont ensuite décortiqué pour déterminer des règles logiques. Ces règles logiques sont la règle de substitution et le *modus ponens*. À l'instar des mathématiques modernes, le syllogisme d'Aristote repose en fait sur ces deux règles logiques. Pour les comprendre, il est bon de commencer par un cas plus simple que le syllogisme d'Aristote. Prenons l'exemple de deux événements :

: Il pleut.

: J'ai pris mon parapluie.

Chaque événement peut être vrai ou faux. À partir de ces deux événements, que l'on appelle aussi *variables booléennes*, on peut fabriquer de nouveaux événements, que l'on appelle *formules logiques*. Par exemple, on peut construire les formules « non », « ou » ou encore « et », et même des formules plus compliquées comme « (non) ou ». Pour comprendre les formules, il est utile de construire des tables de vérité qui explicitent les valeurs logiques des formules en fonction de celles des variables booléennes.

Tableau 3.1. Table de vérité de « il pleut ou j'ai pris mon parapluie »

	= ✓	= ✗
= ✓	✓	✓
= ✗	✓	✗

Par exemple, dans le tableau ci-dessus, la ligne du milieu, qui correspond à , est le cas où il pleut, tandis que la colonne du milieu est le cas où j'ai mon

²En fait, il semble qu'Aristote n'ait pas envisagé de tels syllogismes, parce que sa théorie ne voulait pas traiter du particulier (et donc se refusait à raisonner avec la seconde prémissse). Ceci s'oppose à l'approche des stoïciens qui avaient développé la logique des propositions que l'on verra plus tard. Pour plus de précisions, je vous renvoie vers l'excellent livre *Les Métamorphoses du calcul* de Gilles Dowek (Le Pommier, 2007).

parapluie. La case du milieu correspond donc à la valeur logique de « $\text{pluie ou parapluie}$ » lorsque pluie et parapluie sont tous les deux vrais. Cette case affirme que si pluie et parapluie sont vrais, alors « $\text{pluie ou parapluie}$ » est vrai. Je vous invite à prendre le temps d'analyser cette table de vérité par vous-même, et à déterminer les tables de vérité d'autres formules logiques.

Jusque-là, on n'a étudié que des formules à deux variables booléennes. Mais on peut aller plus loin et considérer des formules à trois, huit, ou un très grand nombre de variables booléennes. Les tables de vérité se doivent alors d'être beaucoup plus grandes pour lister toutes les combinaisons possibles des valeurs des variables booléennes. Je vous invite à calculer par vous-même le nombre d'entrées de ces tables de vérité géantes, ainsi que le nombre de tables de vérité différentes. Mais je vous déconseille fortement de lister toutes les différentes tables de vérité pour trois variables booléennes ou plus. Il y a 256 tables de vérité pour 3 variables booléennes... Et il y a à peu près autant de tables de vérité pour 8 variables booléennes que de particules dans l'univers !

Il y a une formule logique particulièrement importante en pratique. Il s'agit de la formule « $(\text{non pluie}) \text{ ou parapluie}$ », que l'on écrit aussi plus communément « $\text{pluie} \rightarrow \text{parapluie}$ ». Cette formule se lit intuitivement « $\text{pluie} \rightarrow \text{parapluie}$ » ou « Pour toute pluie, il y a parapluie », ou encore « si pluie alors parapluie ». Je vous invite à longuement méditer la table de vérité de cette formule, qui a de quoi surprendre celui qui n'y a pas assez réfléchi.

Tableau 3.2. Table de vérité de « s'il pleut alors je prends mon parapluie »

	$\text{parapluie} = \checkmark$	$\text{parapluie} = \times$
$\text{pluie} = \checkmark$	\checkmark	\times
$\text{pluie} = \times$	\checkmark	\checkmark

Si la formule $\text{pluie} \rightarrow \text{parapluie}$ est particulièrement importante, c'est parce qu'elle est au cœur de la déduction logique. La déduction logique consiste justement à partir de prémisses pluie pour en déduire des conclusions parapluie . Si l'implication est vraie et si les prémisses sont vraies, alors les conclusions le seront elles aussi. Plus formellement, on a la formule logique $((\text{pluie} \rightarrow \text{parapluie}) \text{ et pluie}) \rightarrow \text{parapluie}$. Cette formule est ce que l'on appelle le *modus ponens*³.

Je vous invite à remplacer la flèche d'implication par sa définition en termes de « ou », « et » et « non » que l'on a définis plus haut. En jouant avec la formule logique ainsi obtenue, ou en dressant une table de vérité, vous devriez conclure que, peu importent les valeurs de pluie et parapluie , le *modus ponens* est toujours vrai. On dit que le *modus ponens* est une tautologie, car il est vrai pour toutes valeurs logiques des variables booléennes.

³En fait, c'est plus compliqué que cela. Pour être rigoureux, il faudrait distinguer le *alors* et le *et* du méta-langage des symboles \rightarrow et et du langage dans lequel on exprime la logique. Voir :

▶ *Le paradoxe de Lewis Carroll* | Grain de philo | Monsieur Phi | T. Giraud (2017)

Contrairement à son sens commun, en logique, une tautologie n'a potentiellement rien « d'évident ». Elle n'est pas forcément une lapalissade ou une trivialité. À l'instar du *modus ponens*, certaines tautologies échappent à la plupart d'entre nous, et requièrent le temps de la réflexion.

Les dames sont-elles toutes bleues ?

Revenons-en aux cartes d'Hygiène Mentale. Souvenez-vous, l'hypothèse à tester est « $\text{dame} \rightarrow \text{bleu}$ ». Quatre cartes sont devant vous. La première est une dame, la deuxième est un dix, la troisième une carte bleue et la dernière est rouge. Quelle(s) carte(s) faut-il retourner pour tester l'hypothèse ?

La première carte est une dame. Le *modus ponens* permet d'effectuer une pré-diction. En effet, on a la tautologie « $((\text{dame} \rightarrow \text{bleu}) \text{ et } \text{dame}) \rightarrow \text{bleu}$ ». Si l'hypothèse à tester est vraie, et puisque la première carte est une dame, alors son derrière est bleu. L'hypothèse prédit donc que le derrière de la première carte est bleu. Si le derrière de la carte n'est pas bleu, on aura alors réfuté l'hypothèse.

Pour la deuxième carte, en revanche, force est de constater que l'on ne peut rien déduire sur la valeur de bleu à partir de la prémissse « $((\text{dame} \rightarrow \text{bleu}) \text{ et non dame})$ ». La carte peut alors très bien être « ni dame ni bleu » ou « bleu mais pas dame ». De même, dans le cas de la troisième carte, rien sur bleu ne peut être déduit de la prémissse « $((\text{dame} \rightarrow \text{bleu}) \text{ et bleu})$ ». Les cas où la carte est une dame et où la carte n'est pas une dame sont tout deux compatibles avec l'hypothèse. Prenez le temps de vous en convaincre. Une façon d'y arriver est de dresser les tables de vérités de ces formules logiques.

En revanche, l'hypothèse à tester émet bel et bien une pré-diction univoque pour la dernière carte. En effet, si l'autre côté d'une carte qui n'est pas bleu est une dame, alors on aura une carte qui est une dame mais dont l'autre côté n'est pas bleu. Ceci contredirait l'hypothèse. Il faut donc retourner cette dernière carte pour tester l'hypothèse.

À l'instar du cas de la première carte, le cas de la dernière carte correspond lui aussi à une tautologie. Il s'agit de la tautologie « $((\text{corbeau} \rightarrow \text{noir}) \text{ et non noir}) \rightarrow \text{non corbeau}$ ». C'est le *modus tollens*. Cette tautologie peut aussi se réécrire sous la forme « $(\text{corbeau} \rightarrow \text{noir}) \rightarrow (\text{non noir} \rightarrow \text{non corbeau})$ ». Autrement dit, l'implication « $\text{corbeau} \rightarrow \text{noir}$ » implique l'implication « $\text{non noir} \rightarrow \text{non corbeau}$ », que l'on appelle la *contraposée*. En fait, ces deux implications sont même équivalentes.

L'équivalence entre une implication et sa contraposée est l'une des nombreuses tautologies contre-intuitives de la logique. Considérons ainsi l'hypothèse qui dit que « tous les corbeaux sont noirs ». La contraposée de cette hypothèse affirme que « tout ce qui n'est pas noir n'est pas un corbeau ». Or, puisqu'une hypothèse est équivalente à sa contraposée, confirmer la contraposée revient à confirmer l'hypothèse. En particulier, chaque pomme rouge confirme l'hypothèse selon

laquelle les corbeaux sont noirs ! Cette conclusion extrêmement contre-intuitive mais logiquement implacable est ce que l'on appelle le *paradoxe du corbeau noir* ou *paradoxe de Hempel*. D'expérience, même les détenteurs de thèses en mathématiques ont du mal à l'anticiper et à l'accepter !

Un autre exemple de tautologie logique est celui de la disjonction de cas, qui correspond à la tautologie « $((\text{bleu} \rightarrow \top) \text{ et } (\text{non bleu} \rightarrow \top)) \rightarrow \top$ ». On peut aussi citer le raisonnement par l'absurde, ou *reductio ad absurdum*, dont la tautologie associée est « $(\text{bleu} \rightarrow (\top \text{ et non } \top)) \rightarrow \text{non bleu}$ ».

Quantificateurs et prédictats

La logique des propositions vue jusque-là est très riche et déjà très contre-intuitive. Cependant, son langage est trop restreint. En effet, chaque formule logique de la logique des propositions ne peut faire intervenir qu'un nombre fini de variables booléennes. Or, en mathématiques comme en sciences, il est fréquent de vouloir étudier l'ensemble des possibles, lequel peut être infini.

Par exemple, comme il y a une infinité de nombres, on peut former une infinité de variables booléennes avec ces nombres. Typiquement, il y a une infinité de propositions de la forme « n est un nombre pair ». Une pour chaque nombre entier n . Plutôt que de donner un nom différent pour chacune de ces propositions, on peut alors considérer une proposition $\text{Pair}(n)$ qui dépend du nombre n . On dit alors de Pair qu'il s'agit d'un prédictat. On peut former des prédictats plus compliqués encore, par exemple $\text{Addition}(m, n, p)$ qui signifie « $m + n = p$ ».

Un prédictat ne peut pas être *vrai*. Ce que l'on peut dire d'un prédictat, c'est s'il est toujours vrai, toujours faux, au moins une fois vrai ou au moins une fois faux. Ces quatre affirmations correspondent aux différents quantificateurs. Ainsi, on dit que $P(n)$ est toujours vrai si, pour tout n , $P(n)$ est vrai. Cette phrase s'écrit brutalement en logique des prédictats sous la forme compacte $\forall n P(n)$. Le symbole « \forall » peut alors se lire « pour tout ». C'est le *quantificateur universel*. De façon similaire, si $P(n)$ est toujours faux, on écrit $\forall n (\text{non } P(n))$. Enfin, le fait que $P(n)$ soit vrai (respectivement, faux) pour au moins une valeur de n s'écrit $\exists n P(n)$ (respectivement, $\exists n (\text{non } P(n))$). Le symbole « \exists » est le *quantificateur existentiel*. Il se lit « il existe ».

De façon cruciale, une formule dont toutes les variables des prédictats ont été quantifiées universellement ou existentiellement devient une proposition, qui peut être vraie ou fausse. Ainsi, la phrase « n est pair » n'est ni vraie ni fausse. En revanche, les phrases quantifiées « $\forall n (n \text{ pair})$ » et « $\exists n (n \text{ pair})$ » sont vraies ou fausses. Je vous laisse deviner laquelle est vraie et laquelle est fausse.

On peut alors s'amuser à combiner les symboles de la logique pour former des propositions plus intéressantes, comme « $\forall n ((n \text{ pair}) \rightarrow (n + 1 \text{ impair}))$ », « $\forall n \exists m (m > n)$ » ou « $\forall n \forall p \exists q \exists r (n = pq + r) \text{ et } 0 \leq r < p$ ». Dans sa forme la

plus pure, le boulot du mathématicien est alors de déterminer des formules de la logique des prédictats qui sont des tautologies. Quand une tautologie non-triviale est découverte, on l'appelle théorème.

Le syllogisme d'Aristote réinterprété

On peut enfin adresser le syllogisme d'Aristote. Les objets d'étude de ce syllogisme ne sont pas les nombres, mais les êtres. La première prémissse, qui affirmait « tous les hommes sont mortels », décrit une relation entre deux prédictats sur les êtres. Réécrite en termes logiques, cette prémissse s'écrit alors $\forall x(\text{Homme}(x) \rightarrow \text{Mortel}(x))$, où $\text{Homme}(x)$ veut dire que x est un homme, et $\text{Mortel}(x)$ que x est un mortel. La seconde prémissse affirmait « Socrate est un homme ». Il s'agit donc d'une prémissse qui porte sur une valeur particulière de x , égale à Socrate. On peut le réécrire $\text{Homme}(\text{Socrate})$, qui est donc une variable booléenne qu'on a supposé vraie.

Pour conclure, on veut invoquer l'implication « $\text{Homme}(x) \rightarrow \text{Mortel}(x)$ » qui correspond au cas où x est Socrate. Pour ce faire, les logiciens ont inventé une règle logique appelée la règle de substitution. Appliquée au cas de Socrate, cette règle nous dit que, si Socrate est un objet de notre théorie et si on a la prémissse « $\forall x(\text{Homme}(x) \rightarrow \text{Mortel}(x))$ », alors la formule logique « $\text{Homme}(\text{Socrate}) \rightarrow \text{Mortel}(\text{Socrate})$ » est vraie, parce qu'elle est ce que l'on obtient lorsque l'on a substitué la variable indéterminée x par l'objet de la théorie Socrate.

On peut maintenant conclure à l'aide du *modus ponens*. En effet, puisque l'implication « $\text{Homme}(\text{Socrate}) \rightarrow \text{Mortel}(\text{Socrate})$ » est vraie et puisque la prémissse $\text{Homme}(\text{Socrate})$ de l'implication est vraie, on en déduit que la conclusion de l'implication est vraie elle aussi. On obtient alors la conclusion $\text{Mortel}(\text{Socrate})$, qui est bien la conclusion d'Aristote. On vient de démontrer la validité du syllogisme d'Aristote à l'aide de fondements logiques.

Vous vous demandez sans doute si on ne s'est pas franchement creusé la tête pour pas grand-chose. Après tout, ne savait-on pas déjà que le syllogisme d'Aristote était juste ? Oui, certes. Mais il faut bien voir que sa validité repose sur l'acceptation de règles logiques. Il peut vous sembler irréaliste de rejeter ces règles logiques. Pourtant, la rigueur du logicien le force à questionner ces règles. Et aussi étonnant que cela puisse paraître, certains logiciens, dits *intuitionnistes* ou *constructivistes*, rejettent désormais certaines règles logiques, dont la table de vérité est pourtant remplie de vrai. Par opposition, les logiciens classiques sont parfois dits *platoniciens*.

La divergence entre platoniciens et intuitionnistes est particulièrement explicite dans leurs interprétations respectives du théorème d'incomplétude de Gödel. Mais pour comprendre cela, il nous faut faire un détour par l'axiomatisation.

L'axiomatisation

Pour déterminer la valeur d'une formule logique finie, il nous fallait préciser les valeurs des variables booléennes. Cependant, dans le cas de la logique des prédictats, on ne peut pas se permettre de prendre un temps infini à lister toutes les valeurs de chaque prédictat. L'approche utilisée est alors forcément une approche axiomatique. Autrement dit, à l'instar du syllogisme d'Aristote, on va partir de prémisses appelées axiomes, desquelles devront être déduits les conséquences logiques. Formellement, les mathématiques se résument alors à déterminer des tautologies de la forme Axiomes → Théorème.

Prenons l'exemple des axiomes de Peano, qui fondent une théorie des nombres entiers naturels, c'est-à-dire 0, 1, 2, 3... Le premier axiome postule l'existence d'un objet dans cette théorie, que l'on appelle communément 0. Puis, le second axiome de cette théorie affirme, en gros, que tout nombre possède un successeur. Il y a plusieurs autres axiomes avancés par Peano que je ne détaillerai pas ici⁴. Ce qui est assez magique, c'est qu'à partir des quelques axiomes de Peano, on peut déduire une liste monstrueuse de théorèmes des mathématiques.

Cependant, les axiomes de Peano sont limités à la théorie des nombres entiers. Or, de nombreux objets mathématiques intéressants ne sont pas des nombres entiers. Il y a aussi les nombres réels, les courbes géométriques ou encore les probabilités. De nos jours, la plupart des mathématiciens préfèrent donc les axiomes de Zermelo-Fraenkel (ZF), et y ajoutent (ou non) l'axiome du choix (C). La quasi-totalité des théorèmes prouvés en mathématiques s'écrivent donc ZFC → Théorème.

Le théorème de Gödel s'applique à tous les ensembles d'axiomes qui généralisent les axiomes de Peano. Mieux encore, le théorème de Gödel s'applique à toutes les théories qui se fondent sur la logique des prédictats, qui reposent sur un ensemble fini (ou calculable) d'axiomes et qui sont capables de décrire l'addition et la multiplication des nombres entiers naturels⁵. Le théorème de Gödel affirme alors que, dans toutes ces théories, il existe des formules dont les axiomes ne peuvent déterminer ni la vérité ni la fausseté⁶. Et ceci inclut ZF et ZFC.

Platoniciens versus intuitionnistes

Un mathématicien platonicien interprète alors ce théorème comme une définition des axiomes. Pour le platonicien, les entiers naturels, ou les ensembles, ont une réalité dans un monde idéal, et toute proposition sur ce monde est forcément vraie ou fausse. Malheureusement, la finitude des mots et des symboles nous contraint à n'avoir qu'une description partielle de ce monde idéal. Il y

⁴  1+1=2 (en arithmétique de Peano) | Infini 13 | Science4All | L.N. Hoang (2016)

⁵ Avec quelques autres détails techniques mineurs que l'on va passer sous silence ici.

⁶  Les théorèmes d'incomplétude de Gödel | Infini 18 | Science4All | L.N. Hoang (2016)

a donc des théorèmes vrais concernant ce monde idéal, mais la finitude de nos axiomes nous empêche de les prouver. Pour le platonicien, le théorème de Gödel montre qu'il existe des théorèmes vrais sans démonstration.

Le mathématicien intuitionniste a une interprétation différente de ce théorème. Pour l'intuitionniste, les mathématiques sont un jeu de construction. Ainsi, le premier axiome de Peano est avant tout un outil qui nous permet de construire le nombre 0. Quant au second axiome de Peano, c'est une sorte de machine à qui on donne un nombre entier naturel, et qui utilise ce nombre pour fabriquer un nouveau nombre entier naturel.

Mieux encore, notamment dans le cadre de la théorie des types, une alternative moderne à la logique des prédictats, l'intuitionniste considère que les « preuves mathématiques » sont des objets de sa théorie, et qu'elles doivent donc être construites elles aussi. En particulier, la question que se pose l'intuitionniste est celle de la constructibilité de ses objets. Pas celle de la vérité de ses théorèmes. Pour l'intuitionniste, le théorème de Gödel affirme que dans toute théorie, il existe des théorèmes pour lesquels aucune démonstration, pour ou contre le théorème, ne peut être construite. Et ceci ne lui pose aucun problème métaphysique, puisque la question de la vérité du théorème n'a qu'une importance secondaire.

Le cœur du débat entre platoniciens et intuitionnistes peut se résumer à la loi du tiers exclu. Cette loi affirme que la proposition « P ou non P » est une tautologie. Il semble suffire de dresser une table de vérité pour s'en rendre compte. Si P est vraie, alors « P ou non P » est vraie aussi. Et si P est fausse, alors non P est vraie, et donc « P ou non P » est vraie aussi.

Cependant, pour l'intuitionniste, il demeure une troisième possibilité : le cas où P n'est ni démontrable ni réfutable. On dit alors de P qu'il est indécidable. Et bien, si P est indécidable, on voit alors que ni P ni non P n'est vraie : la proposition « P ou non P » est alors indécidable elle aussi. En effet, en l'absence de preuves de P et de non P , il est impossible de construire une preuve de « P ou non P ». Cette proposition n'est donc pas une tautologie pour l'intuitionniste.

L'opposition entre platoniciens et intuitionnistes ne se limite donc pas à une interprétation du théorème de Gödel. L'intuitionniste rejette toutes les preuves non-constructives que le platonicien a prouvées. Parmi les plus connus de ces théorèmes, on peut citer le paradoxe de Banach-Tarski⁷, le théorème de la base incomplète ou l'unicité des clôtures algébriques.

⁷  Deux (deux ?) minutes pour... le théorème de Banach-Tarski | El jj | J. Cottanceau (2016)

La logique bayésienne*

Qu'en est-il de notre *pure bayésienne*? À quelle logique croit-elle? L'une des découvertes les plus excitantes de mes méditations bayésiennes est que la *pure bayésienne* a en fait sa propre logique déductive, qui n'est ni classique ni intuitionniste. On peut donc parler de logique bayésienne, qui est un cas particulier de la formule de Bayes. Dans cette logique, la vérité d'un événement comme \heartsuit correspond au cas extrême où cet événement intervient avec probabilité 1, c'est-à-dire quand $\mathbb{P}[\heartsuit] = 1$. Par ailleurs, le fait qu'un événement \heartsuit implique un événement \clubsuit s'écrit $\mathbb{P}[\clubsuit|\heartsuit] = 1$. Autrement dit, en termes bayésiens, \heartsuit implique \clubsuit , si et seulement si, la probabilité de \clubsuit sachant \heartsuit est égale à 1.

Le *modus ponens* et le *modus tollens*, à l'instar de d'autres règles logiques, sont eux aussi des cas particuliers de la formule des probabilités totales et de la formule de Bayes. Souvenez-vous que le *modus ponens* est la tautologie « $((\heartsuit \rightarrow \clubsuit) \text{ et } \heartsuit) \rightarrow \clubsuit$ ». La version bayésienne correspond à supposer que $\mathbb{P}[\clubsuit|\heartsuit] = 1$ et $\mathbb{P}[\heartsuit] = 1$. De ces deux égalités, on en déduit, via des calculs que je vous invite à effectuer de vous-même, que $\mathbb{P}[\clubsuit] = 1$. De la même manière, en vous appuyant notamment sur la formule de Bayes, je vous invite à prouver le *modus tollens*, la contraposition et le tiers exclu.

La logique bayésienne semble donc ne rien avoir à envier aux logiques communément admises. Cependant, elle n'est pas équivalente à ces logiques classiques. En particulier, il y a une petite différence entre l'implication de la logique classique « $\heartsuit \rightarrow \clubsuit$ » et l'identité bayésienne $\mathbb{P}[\clubsuit|\heartsuit] = 1$, à savoir le cas où \heartsuit est faux. En effet, lorsque \heartsuit est faux, la formule logique $\heartsuit \rightarrow \clubsuit$ est vraie, même si \clubsuit est faux. Si vous pensez que c'est étrange, vous n'êtes pas le seul⁸! Cependant, si $\mathbb{P}[\heartsuit] = 0$, alors l'expression bayésienne $\mathbb{P}[\clubsuit|\heartsuit]$ est mal définie.

Ce qui est amusant, c'est que l'interprétation bayésienne de l'implication est beaucoup plus naturelle que son interprétation en logique classique. En effet, la phrase « si la France avait gagné la coupe du monde 2006, alors les poules auraient des dents » est vraie en logique classique. Pourtant, elle a de quoi heurter le sens commun. On a plutôt envie de dire que cette phrase n'a pas de valeur logique ; qu'elle n'a aucun sens. C'est ce que conclut la logique bayésienne en affirmant que $\mathbb{P}[\clubsuit|\heartsuit]$ n'est pas définie lorsque $\mathbb{P}[\heartsuit] = 0$.

La logique bayésienne s'étend aussi très naturellement à la logique des prédictats, mais, encore une fois, elle est légèrement distincte de la logique classique. Pour comprendre cette distinction, il faut au préalable considérer que les objets de la théorie sont tirés aléatoirement. Considérons donc une loi de probabilité sur l'ensemble des objets de la théorie. La quantification universelle « $\forall x A(x)$ » se traduit alors par l'identité⁹ $\mathbb{P}[A(x)] = 1$, lorsque x est tiré selon la loi de

⁸ *La logique, c'est pas logique* | Image des Maths CNRS | P. Colmez (2010)

⁹ En théorie de la mesure, l'équivalence entre la logique classique et l'équivalent bayésien que je présente n'est pas une équivalence, parce que certains objets (voire tous les objets) peuvent avoir une probabilité nulle d'être tirée (et puis, il faut définir des sigma-algèbres et

probabilité. En logique bayésienne, cette identité peut tout simplement s'écrire $\mathbb{P}[A] = 1$. À l'inverse, la quantification existentielle « $\exists x A(x)$ » se traduit en $\mathbb{P}[A] > 0$.

La loi de substitution pour le quantificateur universel a alors un équivalent bayésien que l'on peut déduire de la formule de Bayes. Ainsi, si $\mathbb{P}[A] = 1$ et si y est un objet de la théorie¹⁰, alors $\mathbb{P}[A(y)] = \mathbb{P}[A|y] = 1$. En revanche, la quantification existentielle a un équivalent bayésien distinct de son sens en logique classique. En logique bayésienne, si $\mathbb{P}[A] > 0$, alors, tout ce que l'on peut dire, c'est qu'il existe un objet y de la théorie pour lequel l'événement $A(y)$ a une probabilité strictement positive de survenir, c'est-à-dire $\mathbb{P}[A|y] > 0$.

Au-delà du vrai ou faux

La magie de la logique bayésienne, cependant, c'est qu'elle nous permet d'aller plus loin que la logique classique en nous autorisant à manipuler divers niveaux de certitudes et de les combiner. On peut même démontrer qu'il s'agit de la seule logique qui y parvient¹¹. C'est ce qu'affirme le théorème de Jaynes-Cox et ses variantes¹², qui déduisent la logique bayésienne de certains prérequis naturels à la logique des plausibilités.

J'irai même jusqu'à prétendre que la logique bayésienne permet de comprendre pourquoi tant de règles logiques nous semblent contre-intuitives. Pour cela, il est utile d'imiter certains algorithmes de *machine learning* qui, comme les machines de Boltzmann dont on parlera plus tard, s'interdisent de considérer des probabilités égales à 0 ou à 1. Après tout, en pratique, quand on parle de notre monde réel, il est raisonnable de ne jamais rien exclure.

Dès lors, si la phrase « si la France avait gagné la coupe du monde 2006, alors les poules auraient des dents » nous semble fausse, c'est peut-être parce que l'on n'attribue pas tout à fait une probabilité nulle à ce que la France ait gagné la coupe du monde 2006. Peut-être s'est-on endormi la veille de la finale et a-t-on rêvé avoir vécu des années après la défaite de la France. Ou peut-être notre mémoire confond-elle la finale de la coupe du monde 2006 avec celle de l'EURO 2000 — on reviendra sur la fragilité de notre mémoire ! Et qui sait si la victoire de l'Italie ne sera pas un jour annulée pour des raisons de dopage ou autre ?

Pour la *pure bayésienne*, la défaite de la France à la coupe du monde 2006 est très probablement vraie, mais on ne peut pas complètement exclure le fait

tout ça...). Pour faire simple, à travers ce livre, vous pouvez considérer que les probabilités bayésiennes correspondent à des ensembles dénombrables, avec des lois à support total, de sorte que tout objet x a une probabilité strictement positive d'être tiré.

¹⁰De probabilité non nulle.

¹¹La logique floue est parfois avancée pour être un candidat à cela, mais le flou qui y est décrit ne correspond pas à des probabilités (et donc pas à un incertain épistémique).

¹²From Propositional Logic to Plausible Reasoning: A Uniqueness Theorem | International Journal of Approximate Reasoning | K. Van Horn (2017)

qu'elle puisse en fait être fausse. Dès lors, la phrase « si la France avait gagné la coupe du monde 2006, alors les poules auraient des dents » serait incorrecte, dans la mesure où la probabilité que les poules auraient des dents sachant que la France avait gagné la coupe du monde 2006 n'est pas égale à 1. Autrement dit, cet apparent paradoxe disparaît dès que l'on refuse de tomber dans une logique binaire où tout est vrai ou faux, et que l'on s'autorise enfin à juger à l'aide de niveaux de crédences ! *En logique comme en politique, la bipolarisation pousse aux sophismes.*

Justement. L'une des nombreuses idées bayésiennes qui aura grandement excité mes neurones fut l'explication de notre malaise vis-à-vis de la contraposée. Certes, une hypothèse est vraie, si et seulement si, sa contraposée est vraie. Cependant, une hypothèse peut être très probablement vraie sans que sa contraposée le soit.

C'est ce que l'on a vu dans le cas Sally Clark ! Quand une mère est innocente, alors ses enfants vont très probablement ne pas mourir dès leur naissance. Autrement dit, $\mathbb{P}[\text{non } \blacksquare | \heartsuit]$ est très proche de 1. Cependant, si les enfants meurent dès leur naissance, alors la mère demeure néanmoins très probablement innocente. Autrement dit, la probabilité contraposée $\mathbb{P}[\blacksquare | \blacksquare]$ est proche de 0. On a là une explication du fait que la contraposée n'est pas intuitive : en dehors du monde platonicien auquel la logique des propositions s'applique, il est plus raisonnable de considérer des niveaux de crédence plutôt que de trancher la vérité des propositions. Dès lors, la contraposée n'est plus valide.

Se détacher de la bipolarisation vrai ou faux permet également de comprendre d'autres intuitions que la logique classique considérerait être des sophismes. Supposons « $\heartsuit \rightarrow \clubsuit$ ». Autrement dit, on suppose qu'à chaque fois qu'il pleut, je prends mon parapluie. On a alors envie de dire que s'il ne pleut pas, alors j'ai moins de chances de prendre mon parapluie. Cette remarque intuitive n'a pas de pendant en logique classique. Mais elle est un théorème en logique bayésienne. Ce théorème dit que si $\mathbb{P}[\clubsuit | \heartsuit] = 1$, alors $\mathbb{P}[\clubsuit | \text{non } \heartsuit] \leq \mathbb{P}[\clubsuit]$. Plus généralement, certes, l'absence de preuves n'est pas une preuve de l'absence. Mais cette absence de preuve ne peut qu'augmenter la suspicion de l'absence.

Embrasser la logique bayésienne et ses incertitudes permet aussi d'élucider le mystère du corbeau noir. En effet, une analyse bayésienne montre que chaque pomme rouge confirme bel et bien que les corbeaux sont noirs, mais ne le confirme que très, très, très faiblement. Beaucoup, beaucoup, beaucoup plus faiblement que le fait d'observer un corbeau noir, notamment car le nombre d'objets qui ne sont pas des corbeaux est beaucoup plus grand que le nombre de corbeaux. Confirmer la contraposée confirme l'implication originale, mais cet effet peut être si minime qu'il devient largement négligeable. C'est le cas du corbeau noir !

Cette conclusion a un corollaire immédiat concernant le problème des cartes et de la dame nécessairement bleue : les deux cartes que l'on a jugées inutiles à retourner pour réfuter l'hypothèse $\clubsuit \rightarrow \text{B}$ peuvent en fait être utiles pour cor-

roborer ou questionner la contraposée de l'hypothèse, et donc l'hypothèse. Bien entendu, une telle corroboration sera minime, au point d'être essentiellement négligeable.

De façon plus générale, quand il s'agit de confirmer ou réfuter une théorie scientifique, le langage binaire de la logique classique est inapproprié. Il omet l'étendue de la confirmation ou l'ampleur du rejet. D'ailleurs, comme l'affirme Eliezer Yudkowsky, le calcul bayésien des crédences des théories n'est pas une marche. *L'apprentissage est une danse*. À l'instar d'un cours de bourse ou de la température moyenne sur Terre, les crédences de la *pure bayésienne* ne cessent de fluctuer au gré de ses observations. Au cours de cet apprentissage, même les crédences des meilleures théories n'augmenteront pas de manière monotone. Même ces meilleures théories subiront très probablement de nombreuses (légères) pertes de crédences, notamment parce que, par chance, certaines observations colleront parfaitement avec des théories concurrentes. Cependant, sur le long terme, si une théorie est vraiment plus pertinente que ses concurrentes, la tendance de sa crédence sera à la hausse.

Malheureusement, la rigidité de la *méthode scientifique* est incapable de décrire cette danse des crédences. Le langage approprié semble être davantage celui des probabilités de notre *pure bayésienne*.

Vers une cohabitation de théories incompatibles

Un autre point fort du langage probabiliste de la *pure bayésienne* est de permettre de penser plusieurs théories à la fois et, surtout, d'en combiner les prédictions. Cette technique a été reprise avec grand succès en *machine learning* sous le nom d'*ensembling* ou de *bagging*. En pratique, son efficacité est stupéfiante. Combiner différentes théories incompatibles semble bien souvent fournir de meilleures prédictions que celle de la meilleure théorie ! *Une forêt de modèles incompatibles est plus sage que chacun de ses arbres*¹³.

On peut décrire cette approche comme suit. La *pure bayésienne* pense selon une théorie « *Theorie* » en conditionnant les probabilités par *Theorie*. Ainsi, si dans le cadre de *Theorie*, on sait que ☁ implique ↑ et que ☁ arrive avec probabilité $1/2$, on en déduit $\mathbb{P}[\text{☁ et } \uparrow | \text{Theorie}] = \mathbb{P}[\uparrow | \text{☁ et Theorie}] \cdot \mathbb{P}[\text{☁} | \text{Theorie}] = 1 \cdot 1/2 = 1/2$.

La *pure bayésienne* est d'ailleurs capable de calculer certaines probabilités dans de nombreuses théories différentes. Reprenons l'exemple de Sally Clark. La *pure bayésienne* attribue différentes valeurs aux probabilités $\mathbb{P}[\text{☠} | \text{Theorie}]$ et $\mathbb{P}[\text{red face} | \text{Theorie}]$ de doubles morts par causes naturelles et $\mathbb{P}[\text{red face} | \text{Theorie}]$ de la culpabilité *a priori* de Sally Clark, dans différentes théories *Theorie* qu'elle imagine. Ceci la conduit à différentes conclusions de la probabilité $\mathbb{P}[\text{red face} | \text{Theorie}]$ de l'innocence de Sally Clark, sachant la double mort de ses nouveau-nés.

¹³  *La sagesse des forêts* | IA 17 | Science4All | L.N. Hoang (2017)

Si la *pure bayésienne* est pressée par un juge qui lui demande un résultat unique, elle calculera le résultat moyen qu'elle aura obtenu, avec des pondérations proportionnelles aux crédences qu'elle attribue aux différentes théories *Theorie*. De façon (presque) formelle, cette moyennisation n'est autre que la loi des probabilités totales. Elle correspond à l'égalité

$$\mathbb{P}[\text{Je suis un juge} | \text{Theorie}] = \sum_{\text{Theorie}} \mathbb{P}[\text{Theorie}] \mathbb{P}[\text{Je suis un juge} | \text{Theorie et Theorie}],$$

où le symbole \sum signifie que le terme de droite est une somme de (beaucoup de) termes de la forme $\mathbb{P}[\text{Theorie}] \mathbb{P}[\text{Je suis un juge} | \text{Theorie et Theorie}]$, pour différentes théories *Theorie*.

Se pose alors la question du calcul des probabilités $\mathbb{P}[\text{Theorie}]$. Ces probabilités sont les crédences de la *pure bayésienne* en ses différentes théories. Pour qu'il s'agisse bien d'une moyennisation, il faut bien sûr que la somme de ces probabilités soient égales à 1.

Mais surtout, ces probabilités ne sont pas arbitraires. En fait, leur calcul est l'objet central d'étude de ce livre. Comme souvent, ce calcul repose sur la formule de Bayes. Et c'est justement le sujet du prochain chapitre.

Références en français

 *Les Métamorphoses du calcul : Une étonnante histoire des mathématiques* | Le Pommier | G. Dowek (2007)

 *Logicomix* | Bloomsbury Publishing and Bloomsbury | A. Doxiadis, C. Pa-padimitriou, A. Papadatos et A. Di Donna (2010)

 *La logique, c'est pas logique* | Image des Maths CNRS | P. Colmez (2010)

 *Logique & Raisonnement* | e-penser | B. Benamran (2016)

 *Raisonnez de façon correcte (Testez votre logique)* | Hygiène Mentale | C. Michel (2016)

 *La contraposée (sans maths)* | Wandida | E.M. El Mhamdi (2013)

 *La négociation logique (sans maths)* | Wandida | E.M. El Mhamdi (2013)

 *L'axiomatisation* | Passe-Science | T. Cabaret (2016)

 *Deux (deux ?) minutes pour l'hôtel de Hilbert* | El jj | J. Cottanceau (2016)

 *Deux (deux ?) minutes pour... le théorème de Banach-Tarski* | El jj | J. Cottanceau (2016)

 *Les théorèmes d'incomplétude de Gödel* | Science Étonnante | D. Louapre (2016)

- ▶ *Comment démontrer n'importe quoi* | Grain de philo | Monsieur Phi | T. Giraud (2017)
- ▶ *Le scepticisme - Le trilemme d'Agrippa* | Grain de philo | Monsieur Phi | T. Giraud (2017)
- ▶ *Le fondationnalisme - Quelle base pour l'édifice des connaissances ?* | Grain de philo | Monsieur Phi | T. Giraud (2017)
- ▶ *L'axiomatique - Les Éléments d'Euclide* | Grain de philo | Monsieur Phi | T. Giraud (2017)
- ▶ *Le paradoxe de Lewis Carroll* | Grain de philo | Monsieur Phi | T. Giraud (2017)
- ▶ *La règle des règles* | Grain de philo | Monsieur Phi | T. Giraud (2017)

- ▶ *L'infini et les fondations mathématiques* (Playlist) | Science4All | L.N. Hoang (2016)
- ▶ *$1+1=2$ (en arithmétique de Peano)* | Infini 13 | Science4All | L.N. Hoang (2016)
- ▶ *Les théorèmes d'incomplétude de Gödel* | Infini 18 | Science4All | L.N. Hoang (2016)
- ▶ *Les maths : invention ou découverte ?* | Infini 22 | Science4All | L.N. Hoang (2017)
- ▶ *La théorie des types* | Infini 24 | Science4All | L.N. Hoang (2017)
- ▶ *La sagesse des forêts* | IA 17 | Science4All | L.N. Hoang (2017)

Références en anglais

- ➲ *Probability Theory: The Logic of Science* | Washington University | E. Jaynes (1996)
- ➲ *Homotopy Type Theory: Univalent Foundations of Mathematics* | Institute for Advanced Studies | The Univalent Foundations Program (2013)

- ➲ *Reasoning about a Rule* | The Quarterly Journal of Experimental Psychology | P. Wason (1968)
- ➲ *From Propositional Logic to Plausible Reasoning: A Uniqueness Theorem* | International Journal of Approximate Reasoning | K. Van Horn (2017)

- ▶ *5 Stages of Accepting Constructive Mathematics* | Institute for Advanced Studies | A. Bauer (2014)
- ▶ *The Banach-Tarsky Paradox* | VSauce | M. Stevens (2015)
- ▶ *Computer Science ∩ Mathematics (Type Theory)* | Computerphile | T. Altenkirch (2017)
- ▶ *The Netflix Prize* | ZettaBytes | A.M. Kermarrec (2017)

- 🌐 *Type Theory: A Modern Computable Paradigm for Math* | Science4All | L.N. Hoang (2014)
- 🌐 *Homotopy Type Theory and Inductive Types* | Science4All | L.N. Hoang (2014)
- 🌐 *Univalent Foundations of Mathematics* | Science4All | L.N. Hoang (2014)

Toute connaissance dégénère en probabilité; et cette probabilité est plus ou moins grande, en fonction de notre expérience de la vérité ou de la fausseté de notre compréhension, et en fonction de la simplicité ou de la complexité de la question.

David Hume (1711-1776)

Notre cerveau a une fâcheuse tendance à penser que [...] si dans cette hypothèse les résultats sont peu probables, alors l'hypothèse a peu de chance d'être vraie. Ce qui est faux.

Christophe Michel (1974-)

4

Il faut (bien) généraliser !

Le mouton noir d'Écosse

Un biologiste, un physicien et un mathématicien partent en vacances en Écosse pour la première fois. Alors qu'ils sont encore dans le train vers Édimbourg, ils voient un mouton noir. Le biologiste s'exclame : « Incroyable ! Les moutons sont noirs en Écosse ! » Le physicien, agacé, corrige le biologiste : « Tout ce que l'on peut dire, c'est qu'il y a au moins un mouton noir en Écosse. » Le mathématicien, placide, rajoute : « En fait, tout ce que l'on peut dire, c'est qu'au moins la moitié d'un mouton est noir en Écosse. »

Cette plaisanterie a de quoi faire sourire. Si le biologiste s'est peut-être laissé trop emporté, on a envie de dire que le physicien est un peu trop dans la retenu, tandis que le mathématicien est d'une rigueur ridiculement exagérée. Après tout, si la moitié du mouton que l'on voit est bien noir, il semble déraisonnable de ne pas généraliser la noirceur du mouton à son autre moitié.

On pourrait d'ailleurs prolonger l'histoire avec un philosophe qui rajoutera : « Mais qui nous dit que l'on est vraiment en Écosse ? Vous êtes peut-être dans vos lits en train de rêver. Pire encore, un démon a peut-être implanté tous vos souvenirs. Alors que vous croyez vivre une vie sur Terre, vous êtes en fait à la merci de ce démon qui ne fait que vous jouer des tours. Ou si ça se trouve, nous sommes dans une simulation et rien de ce qui nous entoure n'est réel... »

Le problème que soulève la plaisanterie est en fait l'un des plus redoutables problèmes de l'épistémologie en particulier, et de la philosophie en général.

Une brève histoire de l'épistémologie

Le penseur le plus influent sur cette question est sans doute David Hume, un philosophe Écossais du XVIII^e siècle. D'autres philosophes de son temps avaient cru prouver des vérités nécessaires, à l'instar d'un Descartes qui prétendait avoir prouvé l'existence de Dieu via un raisonnement sur la notion de perfection. Il est plus parfait d'exister que de ne pas exister, disait Descartes. Or Dieu est l'être absolument parfait. Donc Dieu existe. Un tel raisonnement est malheureusement un florilège d'erreurs logiques, d'axiomatisation déficiente et de raisonnement motivé.

A contrario, dans ses œuvres magistrales *A Treatise on Human Nature*, puis *An Enquiry Concerning Human Understanding*, Hume affirma qu'il était impossible de déduire quoi que ce soit d'absolu et de général sur notre monde sur la simple base de nos observations. L'empirisme ne peut pas conduire à des vérités *nécessaires*. Peu importe le nombre d'observations que l'on en fait, on ne pourra jamais conclure que le soleil se lèvera *tous* les jours. Les observations passées, même univoques, ne permettent aucune prédiction sans réserve sur le futur.

Néanmoins, Hume argua aussi que de telles généralisations sont souvent assez justes, ou du moins, utiles à faire. Et pour Hume, la raison pour laquelle de telles généralisations sont assez justes, c'est en vertu d'un principe d'uniformité de la Nature. Les lois de la Nature ne semblent pas évoluer. Et si elles évoluent, à l'instar de l'activité nucléaire de notre Soleil, elles semblent évoluer suffisamment lentement pour nous autoriser certaines généralisations pour un futur pas trop éloigné.

En particulier, en vertu de ce principe d'uniformité (que l'on justifera partiellement à l'aide de la thèse de Church-Turing), il est alors possible d'effectuer des prédictions sur ce qui est probable. D'un éclair de génie, Hume pressent le rôle central de la théorie des probabilités dans la résolution du problème de l'induction. Si Laplace est le père du bayésianisme, Hume en est sans doute le grand-père — et Solomonoff en est l'enfant prodige !

Mais les germes bayésiens de Hume n'ont pas fleuri. Pendant deux siècles, rares sont ceux qui ont eu l'idée de formaliser et mathématiser ses idées. Pire encore, en 1934, Karl Popper prit le contrepied de Hume en publiant *The Logic of Scientific Discovery*. Popper y dépeint ce qui était, selon lui, la philosophie des sciences. Pour Popper, toute théorie en science se doit avant tout d'être réfutable par l'expérience. C'est ce que l'on appelle le principe de *réfutabilité* : une théorie est scientifique, si et seulement si, elle impose des contraintes sur des observations expérimentales concevables qui la rendent ainsi sujette à une éventuelle réfutation. Il s'agit ensuite de répéter ces expériences pour réfuter définitivement ladite théorie scientifique, ou, le cas échéant, pour la corroborer. Mais selon Popper, une corroboration n'est pas une validation de la théorie !

Une brève histoire de la planétologie

Cependant, les philosophes des sciences d'aujourd'hui rétorquent souvent que les beaux principes de Popper ne correspondent pas tout à fait à la réalité de la recherche scientifique. Illustrons cela avec l'exemple de la planétologie.

En 1821, l'astronome Alexis Bouvard remarqua des anomalies dans la trajectoire d'Uranus. Cette septième planète du système solaire ne semblait pas se mouvoir selon les lois de la gravité de Newton. Les lois de Newton semblaient violées par Uranus. Mais Bouvard ne rejeta pas les lois de Newton.

Bouvard, suivi de John Couch Adams puis Urbain Le Verrier, préférèrent postuler l'existence d'une huitième planète du système solaire. Bouvard, Adams et Le Verrier préféraient croire en l'existence d'entités non observées que d'écouter la philosophie de Popper (même si Popper n'était pas né à ce moment-là). La *pure bayésienne* dirait que leur crédence en la théorie newtonienne était tout simplement plus importante que leur crédence en l'absence d'une huitième planète.

Étrangement, nos trois théoriciens avaient raison ! À la suite de calculs savants, Adams et Le Verrier parvinrent même à déterminer la position précise de cette huitième planète. Adams demanda aux astronomes de l'observatoire de Cambridge d'effectuer la détection de cette huitième planète. Mais Sir Airy se contenta d'émettre des doutes sur les calculs d'Adams. De son côté, Le Verrier, consterné devant le peu d'enthousiasme en France, s'adressa plutôt à l'observatoire de Berlin. Le soir même, Johann Gottfried Galle confirma la prédiction stupéfiante de le Verrier et découvrit Neptune !

On pourrait croire que la morale est qu'il ne faut jamais remettre Newton en question. Mais l'histoire des sciences semble s'être amusée à nous donner le tour-nis. Le même Le Verrier, fort de son succès avec Neptune, étudia des anomalies dans la trajectoire de Mercure. Ces anomalies le conduisirent à prédire l'existence d'une zéro-ième planète du système solaire. Il l'appela Vulcain.

Sauf que personne ne détecta Vulcain. Peut-être cette planète était-elle trop près du Soleil pour être détectée, la luminosité du Soleil prenant le pas sur toute autre luminosité ? Ou peut-être fallait-il, cette fois, remettre Newton à sa place ? C'est cette réflexion audacieuse qu'eut un certain Albert Einstein.

En 1915, après huit longues années de sophismes douteux, de calculs hasardeux et d'éclairs de génie, Einstein publia une nouvelle théorie révolutionnaire de l'espace, du temps et de la gravité. Cette théorie, c'est la *relativité générale*. Son point de départ était une pensée à la fois obscure et lumineuse. Dès 1907, alors employé dans un bureau de brevets, Einstein osa une réflexion qui a suscité la moquerie chez nombre de mes auditeurs sur YouTube : et si la gravité n'était pas une force ? Et si la gravité n'était qu'une illusion ? Et si la gravité n'était qu'un artefact, dû à l'accélération vers le haut du sol et à notre égocentrisme qui nous constraint à nous placer dans le référentiel non-inertiel du sol¹ ?

¹  *Le sol accélère-t-il vraiment vers le haut ?* My4Cents (Chenonceau) | Science4All |

Cette pensée, appelée le *principe d'équivalence*, est ce qu'Einstein nommera la pensée la plus heureuse de sa vie². Mais ce n'est peut-être pas celle qui lui aura donné les plus dangereuses palpitations cardiaques. Huit ans plus tard donc, en 1915, Einstein établit de nouvelles équations de la gravité qui reposaient sur une toute nouvelle conception (non-euclidienne !) de l'espace et du temps. Mais surtout en novembre 1915, Einstein prouva par le calcul que ses mystérieuses — mais ô combien élégantes ! — équations de ladite courbure de l'espace-temps expliquaient parfaitement les anomalies de Mercure ! Il en était désormais convaincu : sa théorie est *juste* ! Dès l'hiver 1916, quelques mois plus tard, sa théorie fut enseignée par le grand mathématicien David Hilbert à l'université de Göttingen³.

On a là encore un petit paradoxe apparent pour le poppérien. Comment Einstein et Hilbert, peut-être les deux plus grands génies du moment, ont-ils pu croire en une théorie qu'aucune observation ne confirmait ? Ce n'est d'ailleurs que quatre ans plus tard que l'observation expérimentale des aberrations optiques d'une éclipse solaire vint conforter la théorie d'Einstein. Un étudiant demanda alors au génie allemand ce qu'il aurait fait s'il en était autrement. Einstein répondit : « J'aurais été désolé pour le bon seigneur. La théorie est correcte. »

Albert Einstein, la superstar des sciences, lui qui avait un intérêt profond pour les travaux de philosophes comme Kant et Mach, n'aurait pas accepté la réfutation par l'expérience. Einstein ne semblait pas obéir à la philosophie de Popper.

Les sciences contre Popper ?

Les exemples de l'inadéquation de la philosophie de Popper que j'ai donnés jusque-là sont loin d'être exhaustifs. Il suffit en fait de prêter attention aux détails de l'Histoire des sciences pour se rendre compte que la méthodologie de Popper n'est pas la norme. De la loi de la biogenèse de Pasteur à la théorie de l'évolution de Darwin, du *Principia Mathematica* d'Isaac Newton à la classification périodique de Mendeleïev, de la mécanique quantique à la théorie des cordes, les inventeurs de ces théories semblent avoir été convaincus par leurs théories bien avant d'avoir pris le temps de les tester comme Popper l'aurait exigé.

L'un des derniers exemples en date est le cas des neutrinos plus rapides que la vitesse de la lumière. En 2011, l'expérience OPERA avait annoncé avoir détecté de tels neutrinos. Or, de tels neutrinos contredisent les fondements de la théorie de la relativité restreinte. Pour la quasi-totalité des physiciens dont la crédence en ces fondements est énorme, il est bien plus facile d'imaginer une erreur de mesure expérimentale.

L.N. Hoang (2016)

²  *L'apesanteur et la pensée la plus heureuse d'Einstein* | Relativité 17 | Science4All | L.N. Hoang (2016)

³  *Et Einstein découvrit la gravité...* Relativité 20 | Science4All | L.N. Hoang (2016)

Dans son livre *The Big Picture*, le physicien Sean Carroll s'amuse à remarquer qu'il s'agit là de l'un des nombreux cas où c'est l'expérience qui fut réfutée par la théorie. Ceci peut paraître paradoxal, ou antinomique à la méthode scientifique. Ça l'est. Mais comme on le verra dans le prochain chapitre, pour notre *pure bayésienne*, le raisonnement des physiciens est en fait parfaitement raisonnable.

Popper lui-même s'était rendu compte que sa philosophie ne devait pas être appliquée *stricto sensu*. Après tout, tout résultat d'expériences est sujet à des erreurs de mesures et des aléas, qui rendent le résultat de l'expérience au moins un peu aléatoire. Il fallait donc adapter la philosophie de Popper pour bien prendre en compte les erreurs statistiques inhérentes aux expériences scientifiques.

Le fréquentisme*

Les héros de la théorie des erreurs statistiques s'appellent Karl Pearson, Egon Pearson (son fils), Jerzy Neyman et surtout Ronald Fisher. Autour des années 1920, ces statisticiens de grand génie développèrent un cadre de pensée, le *fréquentisme*, qui a depuis envahi toutes les sciences. Le *fréquentisme* suppose que les probabilités sont des mesures de fréquences. Pour un *fréquentiste*, comprendre les probabilités, c'est avant tout comprendre comment les erreurs s'annulent quand la taille des échantillons devient suffisamment grande.

L'argument séduisant des *fréquentistes*, à l'instar de la philosophie de Popper, est avant tout l'objectivité de leurs méthodes. En particulier, Pearson, Neyman et Fisher se vantaient d'offrir des méthodes rigoureuses, prédéterminées et applicables à tout problème. Contrairement aux méthodes bayésiennes que les *fréquentistes* dénigraient, la méthode fréquentiste n'offrait pas la possibilité aux tenants des théories de biaiser les conclusions de l'expérience à l'aide d'*a priori* mal définis.

L'un des objets centraux de la philosophie *fréquentiste* est le concept de test statistique par *p-value*. Un test statistique consiste à tester la plausibilité d'une théorie T . Pour les *fréquentistes*, il y a une sorte de présomption de plausibilité⁴. Conformément à la philosophie de Popper, le test statistique va ensuite chercher à rejeter la plausibilité de la théorie T à l'aide d'une expérience. Appelons d les données collectées par l'expérience. Si ces données d sont hautement improbables selon la théorie T , alors les *fréquentistes* proposent de rejeter T .

En fait, quand on essaie de traduire ce raisonnement en termes plus mathématiques, on se rend compte que cette démarche a un défaut de taille : si l'on considère des données très précises, alors toute donnée est hautement improbable. En effet, si j'obtiens $d = 0,1583197412 \pm 10^{-10}$, et si ma théorie disait qu'il fallait obtenir un nombre entre 0 et 1, alors la probabilité d'obtenir exactement d jusqu'au 10^e chiffre après la virgule est d'une chance sur deux milliards. Il faudrait alors rejeter la théorie.

⁴Un bayésien *troll* appellerait ça un *a priori* subjectif !

Pour être plus raisonnables, les *fréquentistes* proposèrent d'associer à toute donnée d l'ensemble D des données qui sont « plus improbables encore » que d , vis-à-vis de la théorie T . Par exemple, si la théorie T dit qu'il faudrait obtenir une donnée $d \approx 0$, mais si on obtient la donnée $d > 0$, alors « D pire que d » va typiquement être l'ensemble des données D qui prennent des valeurs supérieures à d , voire l'ensemble des données D qui s'éloignent de 0 de plus de d unités.

La fameuse *p-value* p associée à la théorie T , à la donnée d et au test statistique que l'on a considéré est alors définie comme étant la probabilité d'obtenir une donnée D comme d ou pire. Autrement dit, on a l'équation suivante :

$$p = \mathbb{P}[D \text{ pire que } d | T].$$

Si on la compare à la formule de Bayes, on voit que cette *p-value* ressemble au terme *d'expérience de pensée* dont on a parlé au chapitre 2. Ce terme mesure à quel point la théorie T est capable de « bien » expliquer les données observées.

Intuitivement, plus p est petit, plus la donnée d semble incohérente avec la théorie T , et plus il est tentant de rejeter la théorie T . Fisher proposa de rejeter les théories pour lesquelles p prenait des valeurs inférieures à 5 %. De nos jours, lorsque les nouvelles technologies permettent la collecte de milliards, voire de millions de milliards de données comme c'est le cas dans certaines expériences de physique, le seuil est habituellement ramené à 0,00003 %.

Quels que soient les détails, il est indéniable que le principe de Fisher a été incroyablement fécond dans les sciences de la deuxième moitié du XX^e siècle. Par exemple, en 2012, le *Large Hadron Collider* du CERN a annoncé avoir détecté le boson de Higgs. En fait, ce qu'il aurait fallu dire pour être pointilleux, c'est que le CERN a montré qu'il était hautement improbable qu'il observe ce qu'il a observé, si on supposait que le boson de Higgs n'existe pas dans le modèle standard de la physique des particules. Autrement dit, la probabilité des données du CERN (ou pire), sachant que le boson de Higgs n'existe pas, est inférieure à 0,00003 %. Voilà qui a conduit les chercheurs du CERN à rejeter l'inexistence du boson de Higgs. Ou comme le diraient les médias, ils ont ainsi accepté l'existence du boson de Higgs⁵.

La méthode des *fréquentistes* a régné sans partage dans le monde des sciences du XX^e siècle. Fisher, en particulier, y est pour beaucoup. Ses vives critiques de la formule de Bayes, et son acharnement malsain à taire tout avis contraire à son génie, rendaient toute alternative à sa philosophie *fréquentiste* taboue. Ainsi affirmait-il : « la théorie de la probabilité inverse [à savoir le théorème de Bayes] est fondée sur une erreur, et doit être entièrement rejetée. » S'il fallait romancer l'Histoire des statistiques de façon manichéenne en faisant de notre *pure bayésienne* l'héroïne, Fisher, plus que tout autre *fréquentiste*, serait le grand méchant.

⁵  *Pas de maths, pas de chocolats !* Scilabus | V. Lalande (2015)

 *[Preuves scientifiques] P-valeur ou je fais un malheur !* La statistique expliquée à mon chat | L. Maugeri, G. Grisi et N. Uyttendaele (2018)

Ceci étant dit, au-delà de son caractère arrogant, nerveux et méprisant, et au-delà de ses convictions eugénistes et racistes, Fisher n'en reste pas moins un brillant mathématicien et l'un des penseurs les plus influents du XX^e siècle. Grâce à sa rigueur et son génie, les sciences du XX^e siècle, notamment celles dites molles, ont énormément progressé et ont grandement gagné en crédibilité. Les statistiques de Fisher ont fait énormément de bien.

Néanmoins, la *pure bayésienne* a de nombreuses objections. En fait, l'approche *fréquentiste* n'a ni queue ni tête pour elle. Pourquoi faudrait-il accepter cette présomption de plausibilité ? Pourquoi supposer *a priori* que toutes les théories se valent ? N'y a-t-il quand même pas certaines théories plus simples ou plus structurées, et donc plus prometteuses que d'autres ? Ne faudrait-il pas prendre en compte les succès passés de la théorie ? Pourquoi faudrait-il considérer les données pires que d ? Y a-t-il toujours une façon *naturelle* de déterminer l'ensemble des données pires que d ? Pourquoi un rejet serait-il définitif ? Que faire si on a rejeté toutes les théories ? Ne devrait-on pas plutôt comparer les théories entre elles ? Pourquoi avoir choisi le seuil de 5 % ? Pourquoi celui de 0,00003 % ? N'est-ce pas complètement arbitraire ? Et que peut-on dire si très peu de données sont disponibles ? Comment parler de la vie dont on ne connaît que l'exemplaire terrestre, ou de l'univers dont on ne dispose que d'une version ? Comment traiter le cas du mouton noir d'Écosse ?

Les statisticiens contre la *p-value*

La *pure bayésienne* n'est pas la seule à critiquer Popper et les *fréquentistes*. Récemment, la *p-value* a même reçu une bien mauvaise presse de la part des statisticiens. L'une des causes principales de cette mauvaise réputation est le biais de sélection causé par le fait que seuls les travaux conclusifs sont publiés. Pire, on assiste de plus en plus à des stratégies de *p-hacking*, sur lesquelles on reviendra un peu plus tard. Peu importent les causes, on assiste à une surabondance de résultats publiés faux, estimée à au moins 25 % par Valen Johnson.

En fait, la présence d'erreurs dans les résultats scientifiques est très certainement largement supérieure, surtout lorsqu'on considère aussi les erreurs non statistiques. L'informaticien Leslie Lamport va même jusqu'à suggérer qu'une publication mathématique sur trois, pourtant acceptée par un comité de relecture, contient au moins un théorème faux⁶ !

Plus étonnant encore, si l'on prenait la *p-value* vraiment au sérieux, alors on devrait finir par rejeter toutes les théories scientifiques. Y compris celles qui sont vraies. En effet, à en croire la plupart des descriptions de la méthode scientifique, toute théorie doit être testée, encore et encore. Or, avec un seuil

⁶  Comment écrire une démonstration au 21ème siècle | Math Park | Institut Henri Poincaré | L. Lamport (2016)

de 0,00003 %, chaque expérience a une probabilité 0,00003 % de rejeter une théorie vraie qu'elle teste. Par conséquent, même une théorie vraie sera rejetée au bout de quelques millions d'expériences. En fait, clairement, *si nos théories scientifiques tiennent encore, c'est uniquement parce qu'elles n'ont pas encore été testées suffisamment*. Mais si ces théories ne cessent d'être testées, le jour de leur rejet viendra. Inéluctablement. Étrange, non ? La méthode scientifique est vouée à rejeter toutes les théories vraies⁷ !

Les statisticiens se sont montrés de plus en plus véhéments au fil des années. En 2010, Tom Siegfried expliqua son scepticisme vis-à-vis des publications scientifiques, affirmant : « C'est le secret inavoué le plus sombre des sciences : la "méthode scientifique" de tester des hypothèses par analyse statistique repose sur une piètre fondation. » En 2014, Regina Nuzzo ajoute : « [Le problème] vient de la nature étonnamment piégieuse de la *p-value*, qui n'est ni aussi fiable ni aussi objective que ce que la plupart des scientifiques suppose. »

Les critiques vives et statistiquement fondées de nombreux statisticiens ont culminé en 2016 dans un communiqué⁸ de l'Association américaine de statistiques (ASA) : « La communauté statistique a été très concernée par les problèmes de reproductibilité et de réplicabilité des conclusions scientifiques. Sans entrer dans les définitions et distinctions entre ces termes, on observe l'émergence de beaucoup de confusion et même de doutes sur la validité des sciences. Un tel doute peut conduire à des décisions radicales, comme celle des éditeurs de *Basic and Applied Social Psychology*, qui ont choisi de bannir les *p-values*. [...] La mécompréhension et la mauvaise utilisation de l'inférence statistique n'est que l'une des causes de la "crise de reproductibilité", mais pour notre communauté, il s'agit d'une cause importante. »

Le *p-hacking*

Il y a bien sûr ceux qui ne comprennent pas suffisamment bien la *p-value*. Mais il y a aussi ceux qui la comprennent trop bien et y voient l'opportunité de booster leur carrière, à une époque où le dicton « *publish or perish* » règne dans le monde académique. Or, pour publier, trop souvent, il faut, voire il suffit, d'obtenir des *p-value* en dessous du seuil de 5 %. Obtenir de telles valeurs de *p* pour des théories valides est assez improbable. Mais ce n'est pas si improbable. Lorsqu'on cherche à rejeter des théories pourtant valides, la probabilité d'obtenir de telles valeurs de *p* est en fait, par construction, de 5 % justement. Et ça, ça veut dire que, en moyenne, une expérience sur 20 va obtenir une *p-value* publiable ! Autrement dit, pour obtenir des résultats scientifiquement publiables, il suffit de multiplier les expériences. C'est ça, le *p-hacking*.

⁷Pour un seuil de 5 %, il suffit d'environ 20 expériences (seulement !) pour rejeter une théorie vraie. Et environ 100 expériences suffisent pour un seuil de 1 %.

⁸ *The ASA's Statement on p-values: Context, Process, and Purpose* | The American Statistician | R. Wasserstein et N. Lazar (2016)

Ce danger du *p-hacking* est particulièrement bien illustré par l' excellente bande dessinée xkcd de Randall Munroe intitulée *Significant*. Munroe y imagine une suspicion qu'une dragée cause l'acné. Un scientifique effectue des expériences et conclut que la théorie T selon laquelle une dragée ne cause pas l'acné est associée à une *p-value* supérieure à 5%. Autrement dit, l'expérience ne permet pas de rejeter T . Jusque-là, tout va bien.

Mais une seconde rumeur dit en fait qu'il n'y a que les dragées d'une certaine couleur qui causent l'acné. Or il y a 20 couleurs de dragées, donc 20 expériences indépendantes à réaliser. Sans surprise, l'une des 20 expériences, celle pour les dragées vertes par exemple, produit une *p-value* inférieure à 5 %. Voilà qui permet de rejeter l'hypothèse selon laquelle les dragées vertes ne causent pas l'acné. Le lendemain, les journaux annoncent alors en gros titre que la science a démontré que les dragées vertes causent l'acné !

À l'échelle des sciences sur toute la Terre, le nombre d'expériences indépendantes est très largement supérieur à 20. Les journaux sensationnels trouveront donc tous les jours des milliers de publications scientifiques surprenantes à reprendre. Ce n'est ainsi pas surprenant qu'une grosse proportion de ces publications finit par être contredite par d'autres travaux, voire rétractée par leurs auteurs.

Outre la multiplicité des expériences, il y a une méthode tout aussi efficace pour obtenir des résultats publiables : il suffit de cumuler les données expérimentales jusqu'à ce que ces données soient conclusives. De façon étrange, il a été démontré que toute théorie finit par être rejetée par la *p-value* si l'on cumule les données expérimentales⁹ jusqu'à ce que l'on puisse conclure ! Autrement dit, si vos données ne permettent pas de rejeter la théorie, vous pourrez toujours finir par la rejeter, simplement en allant collecter toujours plus de données.

Vous voyez le problème ? En choisissant le moment d'arrêter l'expérience, on introduit un énorme biais de sélection. Sauf qu'*a posteriori*, si votre publication ne précise pas la manière dont la quantité de données collectées a été déterminée, il n'y aura rien à y redire. Votre publication sera alors aux normes de la « méthode scientifique » de la *p-value*. Or, cumuler les données jusqu'à obtenir une statistique conclusive, c'est malheureusement une approche très répandue¹⁰...

Certains proposent dès lors de réduire les seuils de la *p-value*. Mais même en physique et même avec le seuil extrême de 0,00003 %, le problème reste entier. Des artefacts de l'analyse statistique et la multiplicité des expériences scientifiques ont ainsi conduit à la découverte du pentaquark en 2003 — qui fut même confirmée par d'autres expériences indépendantes qui torturaient leurs données — avant d'être finalement rejetée par la communauté scientifique qui ne réussissait pas à répliquer l'expérience originale¹¹.

⁹Le nombre de données nécessaires pour être sûr de rejeter une hypothèse est toutefois en général exponentiel. Cependant, Johari, Pelekis et Walsh ont montré que l'augmentation de la probabilité de rejet est déjà énorme même avec des quantités de données raisonnables.

¹⁰Johari et ses collaborateurs proposent une variante de la *p-value* pour corriger le tir.

¹¹Le pentaquark semble finalement avoir été réellement découvert en 2015.

Ces aléas de la « méthode scientifique » par *p-value* remettent sérieusement en cause la fiabilité des sciences. Je ne peux que vous conseiller deux excellents résumés de toutes ces controverses pour en savoir plus : le billet de blog de Science Étonnante¹² et la vidéo de Veritasium¹³.

Ce qu'en dit un cours de statistique

Il est intéressant de voir ce qu'en dit un livre de cours de statistiques modernes très reconnu. Dans le sien, le statisticien Larry Wasserman écrit : « Les résultats d'études observationnelles commencent à devenir crédibles quand :

- (i) les résultats sont répliqués dans plusieurs études,
- (ii) chacune des études a contrôlé les facteurs de confusion plausibles,
- (iii) il y a une explication scientifique plausible de l'existence d'un lien causal. »

On reviendra ultérieurement sur les facteurs de confusion. Pour l'heure, sachez qu'il s'agit d'une difficulté redoutable qui s'ajoute aux déficiences de la *p-value*.

Ce sur quoi j'aimerais insister, c'est le flou artistique que dépeint Wasserman. La citation que je vous ai donnée est minée de termes volontairement imprécis et sujets à interprétations, comme « commencent à devenir », « plusieurs études », « plausibles », « explication scientifique », « lien causal ». Le mot « crédible » ressemble même à une invitation au bayésianisme !

Notez que je ne dis pas cela pour critiquer Larry Wasserman ou son livre. Son cours est excellent. En fait, il me semble que tout bon cours de statistiques *fréquentistes* se doit de souligner ce flou artistique et d'insister sur la prudence dont il faut faire preuve au moment d'interpréter des données statistiques.

Ce flou artistique donne toutefois des impressions d'inconclusivité des sciences. À être trop prudent, on risque de suggérer que rien ne peut être au-delà de tout doute raisonnable, ce qui risque d'amener certains à douter des bienfaits de la vaccination, de l'existence du réchauffement climatique et de l'effet nocif du tabac sur la santé. Même moi, j'ai un sentiment d'inconfort quand un scientifique affirme avoir *prouvé* l'existence du boson de Higgs. Ou qu'il ne fait *aucun doute* que notre univers est quantique. Notre *pure bayésienne* aussi.

Il nous faut un langage mieux approprié pour décrire différents niveaux de certitude ou de crédence en diverses affirmations que l'on pourrait vouloir émettre. La version heuristique ou simplifiée de ce langage est celui des « plausibles », « crédibles », « hautement probables » et « au-delà de tout doute raisonnable ». La traduction plus rigoureuse de ces degrés de crédence est inéluctablement un langage similaire (ou *isomorphe*) aux probabilités bayésiennes.

¹²  Comment être sûr qu'un résultat scientifique est vrai ? Science Étonnante | D. Louapre (2013)

¹³  Is Most Published Research Wrong? Veritasium | D. Müller (2016)

La formule du savoir

Pour la *pure bayésienne*, toute la philosophie du savoir doit se réduire au calcul des crédences bayésiennes. Savoir, c'est assigner des crédences adéquates aux différentes théories. Et pour ce faire, il existe une formule magique. Oui, je parle de la formule de Bayes. En particulier, la version la plus fondamentale de la formule de Bayes, celle que l'on pourrait vraiment appeler la formule du savoir, est, je pense, la merveilleuse formule suivante :

$$\mathbb{P}[\text{Theorie}|\text{Data}] = \frac{\mathbb{P}[\text{Data}|\text{Theorie}] \mathbb{P}[\text{Theorie}]}{\mathbb{P}[\text{Data}|\text{Theorie}] \mathbb{P}[\text{Theorie}] + \sum_{\text{Alter}} \mathbb{P}[\text{Data}|\text{Alter}] \mathbb{P}[\text{Alter}]},$$

où *Alter* désigne les théories alternatives à *Theorie*. Je vous renvoie vers le chapitre 2 pour mieux vous familiariser avec cette formule.

Je ne peux que vous inviter à longuement méditer, encore et encore, sur cette formule absolument remarquable. Je vous invite notamment à imaginer comment la *pure bayésienne* utilise cette formule pour adresser les objections qu'elle a soulevées vis-à-vis de la philosophie de Popper.

Pour commencer, l'équivalent de la *p-value* est le *terme d'expérience de pensée* $\mathbb{P}[\text{Data}|\text{Theorie}]$. Ce terme est crucial. Il mesure la capacité d'une théorie à prédire les données observées. Cependant, pour la *pure bayésienne*, ce n'est qu'une partie de l'équation.

L'autre terme fondamental est l'*a priori* $\mathbb{P}[\text{Theorie}]$. Ce terme est absolument incontournable. Comme on le verra dans de futurs chapitres, c'est cet *a priori* qui nous permet d'apprendre sur notre quotidien malgré la petitesse des échantillons auxquels on a accès. Mieux encore, on verra que l'*a priori* bayésien, combiné à l'informatique théorique, implique le rasoir d'Ockham, ce principe philosophique qui affirme que les théories plus simples sont plus crédibles !

Mais surtout, il y a le dénominateur, aussi appelé *fonction de partition*. Ce dénominateur est égal à $\mathbb{P}[\text{Data}]$, que l'on a décomposé à l'aide de la loi des probabilités totales. L'une des composantes de ce dénominateur est le numérateur lui-même. Mais la fonction de partition comporte aussi des termes similaires associés à des théories alternatives. Autrement dit, cette fonction de partition met en compétition les différentes théories — et garantit par là que la somme des crédences sera toujours égale à 1.

La *pure bayésienne* assigne une grande crédence à une théorie, si et seulement si, cette théorie est beaucoup plus crédible que ses concurrentes. Ceci veut dire que la *pure bayésienne* n'attachera pas une grande crédence à une théorie qui explique des observations simples à expliquer, si d'autres théories plus simples les expliquent tout aussi bien ou mieux. À l'inverse, dans le cas des phénomènes difficiles à expliquer, la théorie qui gagnera les crédences de la *pure bayésienne*

n'aura peut-être pas à parfaitement expliquer ces phénomènes, si aucune autre théorie n'y arrive.

Une autre chose à souligner, c'est ce que j'entends là par **Data**. Cette variable ne désigne pas l'ensemble des résultats d'une expérience scientifique. Cette variable désigne l'ensemble de toutes les données observationnelles auxquelles la *pure bayésienne* a eu accès dans toute sa vie. En particulier, ceci veut dire qu'aucune expérience ne doit être pensée de façon isolée.

D'ailleurs, malgré les protocoles fréquentistes qu'ils utilisent, les scientifiques se fondent beaucoup plus sur la philosophie de cumul des données que ce que la « méthode scientifique » laisse parfois entendre. C'est pour cela que les articles scientifiques commencent toujours avec une longue revue de littérature qui montre que le travail présenté n'est qu'une contribution à un plus large champ de recherche. Comme on le verra notamment dans le prochain chapitre, les scientifiques semblent davantage raisonner selon les principes du bayesianisme que selon ceux de Popper.

L'apprentissage cumulatif

En pratique, on ne détermine pas les crédences en une théorie en se remémorant toutes les données collectées dans notre vie — notamment parce que, comme on le verra, notre mémoire limitée est aussi très déficiente. Notre apprentissage est davantage cumulatif. Et bien, la formule de Bayes permet justement d'intégrer de nouvelles données collectées pour affiner nos crédences. Ce procédé correspond à ce que l'on appelle l'*inférence bayésienne*. Cette inférence, ou mise à jour des crédences, repose sur la décomposition partielle de la formule de Bayes qui suit :

$$\mathbb{P}[T|\text{NewData et } D] = \frac{\mathbb{P}[\text{NewData}|T \text{ et } D]\mathbb{P}[T|D]}{\mathbb{P}[\text{NewData}|T \text{ et } D]\mathbb{P}[T|D] + \sum_A \mathbb{P}[\text{NewData}|A \text{ et } D]\mathbb{P}[A|D]},$$

où, pour plus de concision, on a noté T pour **Theorie**, A pour **Alter** et D pour **Data**. Le cas particulièrement intéressant est alors celui où la nouvelle donnée **NewData** a été obtenue indépendamment des données D , auquel cas¹⁴ on obtient la formule d'inférence bayésienne qui suit :

$$\mathbb{P}[T|\text{NewData et } D] = \frac{\mathbb{P}[\text{NewData}|T]\mathbb{P}[T|D]}{\mathbb{P}[\text{NewData}|T]\mathbb{P}[T|D] + \sum_A \mathbb{P}[\text{NewData}|A]\mathbb{P}[A|D]}.$$

¹⁴Techniquement, il faut même que, dans toute théorie T ou A , **NewData** soit indépendant de D .

Cette formule est la même que la formule de Bayes que l'on a vue précédemment, sauf que l'*a priori* est là un *a priori* calculé à partir des données D collectées avant la nouvelle donnée **NewData**. Autrement dit, en pratique, au moment d'acquérir une nouvelle donnée, le *bayésien pragmatique* va remplacer l'*a priori* fondamental $\mathbb{P}[T]$ par sa crédence actuelle $\mathbb{P}[T|D]$ en la théorie T , et les *a priori* fondamentaux $\mathbb{P}[A]$ sur les alternatives par leurs crédences actuelles $\mathbb{P}[A|D]$. C'est avec ces crédences actuelles que le *bayésien pragmatique* va appliquer la formule de Bayes. Dans de futurs chapitres, on verra que ce principe se retrouve à la fois au cœur de l'évolution darwinienne, de la fiabilité du consensus scientifique et d'algorithmes de *machine learning* en temps réel.

Ces calculs montrent aussi que, selon notre *pure bayésienne*, il est tout à fait possible d'étudier l'Histoire de façon rationnelle. Pour certains, l'Histoire, mais aussi la phylogénie et la cosmologie, sont des disciplines non-scientifiques car elles ne peuvent pas se prêter à la reproductibilité des expériences. Il est intéressant de voir que cette réflexion est un pur artefact de la philosophie de Popper et des statistiques fréquentistes.

Pour la *pure bayésienne*, il n'y a aucune ligne de démarcation fondamentale à tracer entre ces disciplines qui essaient de retracer le passé de notre univers et d'autres disciplines étudiant des lois dont on suspecte l'invariance au cours du temps. Dans les deux cas, il s'agit de collecter des données pertinentes et d'effectuer des inférences bayésiennes pour déterminer les crédences adéquates en diverses théories.

En particulier, pour la *pure bayésienne*, ce qui distingue les « sciences » des « pseudo-sciences » n'est pas tant la réfutabilité des hypothèses de ces disciplines¹⁵. C'est surtout la justesse de l'application de la formule de Bayes par les tenants. Les scientifiques l'appliquent beaucoup, beaucoup, beaucoup mieux — et, comme on le verra, la communauté scientifique l'applique encore mieux que ses membres !

Revenons-en à Einstein

Pour mieux comprendre la philosophie de la *pure bayésienne*, revenons-en au cas d'Einstein. La première chose à remarquer est que, bien que trop abstraite pour la quasi-totalité des physiciens de son temps, la relativité générale d'Einstein est remarquablement simple. En effet, elle se résume à dire que l'espace-temps de dimension 4 a une courbure déterminée par l'équation $G_{\mu\nu} \propto T_{\mu\nu}$. Les détails importent peu. Ce qu'il faut noter, c'est que cette équation comporte un unique paramètre, qui est le coefficient de proportionnalité de l'équation. Ce coefficient décrit l'intensité avec laquelle la matière $T_{\mu\nu}$ affecte la courbure $G_{\mu\nu}$ de l'espace-temps.

¹⁵Quoique, comme on le verra au chapitre 7, toute bonne théorie doit être prédictive, quitte à avouer l'étendue de son ignorance lors d'une prédiction en prédisant 50-50 quand il n'y a que deux alternatives.

De son côté, la loi de la gravité de Newton $M_1\vec{a} \propto m_1m_2\vec{r}/r^3$ comporte aussi un coefficient qui décrit l'intensité avec laquelle les masses des corps affectent la force gravitationnelle entre ces corps. Cependant, la loi de la gravité de Newton doit, de plus, supposer que la masse inertielle M_1 qui lutte contre l'accélération des objets est égale (ou proportionnelle) à la masse gravitationnelle m_1 qui cause la force de gravité. Ainsi, même si la loi de Newton semble plus facile à concevoir et même si elle se prête mieux aux calculs, la loi de Newton est en fait plus arbitraire que la théorie d'Einstein.

Aujourd'hui, on sait même à quel point la théorie d'Einstein est tout sauf arbitraire. On a ainsi prouvé que les combinaisons linéaires du tenseur métrique et du tenseur d'Einstein étaient les seuls tenseurs de valence 2 comportant uniquement des dérivées au plus secondes du tenseur métrique. Malheureusement, expliquer ce théorème mathématique prendrait trop de temps¹⁶. Ce qu'il suffit de retenir, c'est que l'équation d'Einstein n'a absolument rien d'arbitraire. Cette équation est une conséquence inéluctable du simple postulat selon lequel la courbure de l'espace-temps est la cause de ce que nous appelons erronément la « gravité ». Voilà qui explique d'ailleurs qu'Hilbert a découvert la même équation que celle d'Einstein de façon indépendante. Ce caractère inéluctable fait de la théorie d'Einstein une théorie formidablement plus crédible *a priori* que la loi de Newton. C'est pour ça que, même avec quasiment aucune donnée observationnelle, il n'est pas déraisonnable de préférer la théorie d'Einstein à celle de Newton.

Au moment où Einstein avait déterminé son équation, il ne savait pas encore qu'elle était la conséquence inéluctable de son postulat initial. Mais l'élégance de son équation, combinée à des années de frustration à travailler avec de mauvaises équations, lui a sans doute mis la puce à l'oreille. Ce n'est donc pas étonnant que sa crédence en sa nouvelle théorie à ce moment-là était d'ores et déjà supérieure à celle en la loi de Newton, ou du moins suffisamment importante pour se donner l'ambition de renverser Newton. Mais il y a bien sûr un élément observationnel en particulier qui allait complètement faire pencher la balance.

La relativité générale d'Einstein se comporte comme la loi de Newton lorsque la gravité est « faible ». En fait, dans tout notre système solaire, pour lequel les données de l'époque étaient limitées, il n'y a que près du soleil que la gravité est suffisamment forte pour que les prédictions de l'équation d'Einstein diffèrent de celles de Newton. Or, la théorie de Newton ne semble pas expliquer la trajectoire de Mercure, qui se trouve être la planète évoluant dans le champ de gravité le plus intense (ou, comme le dirait Einstein, dans la région de l'espace-temps la plus courbée).

Certes, il pourrait y avoir une planète comme Vulcain pour expliquer la trajectoire de Mercure. Mais c'est en vain que de nombreux observateurs avaient cherché Vulcain depuis des décennies. L'absence de preuve, si elle n'est pas la preuve de l'absence, ne peut affecter la crédence en Vulcain que négativement.

¹⁶  *Les tenseurs de la relativité générale* | Hardcore | Science4All | L.N. Hoang (2016)

Cependant, la théorie d'Einstein arrive à expliquer formidablement bien la trajectoire de Mercure, dans le cas hautement plus probable où aucune planète non observée n'affecte sa trajectoire. Même si la crédence *a priori* en la théorie d'Einstein était la même que celle en la mécanique newtonienne, le fait que la théorie d'Einstein explique parfaitement la trajectoire de Mercure sans hypothèse hautement improbable ne laisse plus aucun doute à notre *pure bayésienne* : la formule de Bayes implique une crédence en la théorie d'Einstein formidablement plus grande qu'en celle de Newton.

Nul doute qu'Einstein et Hilbert avaient intuité ce raisonnement bayésien. Dès novembre 1915, contrairement à l'entièreté de la communauté scientifique, tout deux étaient d'ores et déjà persuadés que la relativité générale était désormais, et de loin, la théorie de la gravité la plus crédible. La *pure bayésienne* approuve.

Références en français

⌚ *Comment être sûr qu'un résultat scientifique est vrai ?* Science Étonnante | D. Louapre (2013)

▶ *Comment écrire une démonstration au 21ème siècle* | Math Park | Institut Henri Poincaré | L. Lampert (2016)

▶ *Pas de maths, pas de chocolat !* Scilabus | V. Lalande (2015)

▶ *Pourquoi vous perdez au casino : rencontre avec la loi des grands nombres* | La statistique expliquée à mon chat | L. Maugeri, G. Grisi et N. Uyttendaele (2016)

▶ *[Preuves scientifiques] P-valeur ou je fais un malheur !* La statistique expliquée à mon chat | L. Maugeri, G. Grisi et N. Uyttendaele (2018)

▶ *La lune n'a PAS d'influence sur les naissances (Bayésianisme)* | Hygiène Mentale | C. Michel (2018)

▶ *Les statistiques à l'heure du Big Data* | CESP Villejuif | L.N. Hoang (2016)

▶ *L'Histoire de la planétologie* | Relativité 2 | Science4All | L.N. Hoang (2016)

▶ *L'apesanteur et la pensée la plus heureuse d'Einstein* | Relativité 17 | Science4All | L.N. Hoang (2016)

▶ *La relativité générale* | Relativité 18 | Science4All | L.N. Hoang (2016)

▶ *Et Einstein découvrit la gravité...* Relativité 20 | Science4All | L.N. Hoang (2016)

▶ *Le sol accélère-t-il vraiment vers le haut ?* My4Cents (Chenonceau) | Science4All | L.N. Hoang (2016)

▶ *Albert Einstein, la superstar des sciences* | Science4All | L.N. Hoang (2016)

▶ *Les tenseurs de la relativité générale* | Hardcore | Science4All | L.N. Hoang (2016)

Références en anglais

- ➲ *A treatise of human nature* | Courier Corporation | D. Hume (1738)
- ➲ *An Enquiry Concerning Human Understanding* | London: A. Millar | D. Hume (1738)
- ➲ *The logic of scientific discovery* | Routledge | K. Popper (2005)
- ➲ *All of Statistics: A Concise Course in Statistical Inference* | Springer Science & Business Media | L. Wasserman (2013)
- ➲ *The Big Picture: On the Origin of Life, Meaning and the Universe Itself* | Dutton | S. Carroll (2016)

- ➲ *Statistical Methods for Research Workers* | Genesis Publishing Pvt Ltd | R. Fisher (1925)
- ➲ *On the Problem of the Most Efficient Tests of Statistical Hypotheses* | Breakthroughs in Statistics | J. Neyman and E. Pearson (1933)
- ➲ *Why Most Published Research Findings are False* | PLoS Med | J. Ioannidis (2005)
- ➲ *Revised Standards for Statistical Evidence* | Proceedings of the National Academy of Sciences | V. Johnson (2013)
- ➲ *Statistical Errors* | Nature | R. Nuzzo (2014)
- ➲ *Editorial* | Basic Applied Social Psychology | D. Trafinow and M. Marks (2015)
- ➲ *The Reproducibility Crisis in Science: A Statistical Counterattack* | Significance | R. Peng (2015)
- ➲ *The ASA's Statement on p-values: Context, Process, and Purpose* | The American Statistician | R. Wasserstein et N. Lazar (2016)
- ➲ *Can I take a Peek? Continuous Monitoring of Online a/b Tests* | Proceedings of the 24th International Conference on World Wide Web | ACM | R. Johari (2015)
- ➲ *Always Valid Inference: Bringing Sequential Analysis to a/b Testing* | R. Johari, L. Pekelis and D. Walsh (2015)

- ➲ *Significant* | xkcd | R. Munroe
- ➲ *Hypothesis Test with Statistics: Get it Right!* | Science4All | L.N. Hoang (2013)

- *Is Most Published Research Wrong?* | Veritasium | D. Muller (2016)
- *Scientific Studies* | Last Week Tonight | J. Oliver (2017)

Tout savoir humain commence par l'intuition, se poursuit ensuite pour devenir concepts, et finit en tant qu'idées.

Emmanuel Kant (1724-1804)

Selon le théorème de Bayes, aucune théorie n'est parfaite. Au lieu de cela, il s'agit d'un travail en cours, toujours sujet à davantage de raffinements et de tests.

Nate Silver (1978-)

5

Gloire aux préjugés

Le problème de Linda

Linda a 31 ans. Elle est célibataire, franche et très intelligente. Elle a fait des études de philosophie. Quand elle était étudiante, elle était très concernée par les problèmes de discrimination et de justice sociale, et a aussi participé à des manifestations anti-nucléaires. Qu'est-ce qui est le plus probable ?

1. Linda est caissière de banque.
2. Linda est caissière de banque et active dans le mouvement féministe.

Je vous invite à prendre le temps de la réflexion et à donner votre réponse de vive voix avant de poursuivre la lecture.

Ce problème, fameusement intitulé le « problème de Linda », a été proposé par Amos Tversky et Daniel Kahneman, deux chercheurs en psychologie qui cherchaient à mieux comprendre la manière dont les gens réfléchissent. Pour ce travail et bien d'autres, Kahneman gagnera d'ailleurs le prix Nobel d'économie en 2002. Dans son excellent livre, il rajoute que si Tversky n'était pas décédé en 1996, les deux chercheurs auraient alors sans doute partagé ce prix Nobel.

Si le problème de Linda est devenu si connu, c'est parce que le taux d'échec à ce problème est énorme. Entre 85 % et 91 % des personnes interrogées par Tversky et Kahneman donnèrent la mauvaise réponse dans diverses réplications de l'expérience. Nos brillants cerveaux humains font là très nettement moins bien qu'un chimpanzé qui répondrait au hasard !

Certaines critiques ont évoqué l'ambiguïté de la réponse 1 qui pourrait laisser entendre que Linda n'est pas active dans le mouvement féministe. Cependant, même lorsque la réponse 1 est clarifiée et remplacée par « Linda est caissière de banque et elle est ou non active dans le mouvement féministe », le taux d'échec restait de 57 % — encore moins bien que le chimpanzé.

Si vous n'avez pas résolu le problème, vous êtes peut-être surpris qu'il existe une *bonne* et une *mauvaise* réponse. C'est pourtant le cas. En effet, la réponse 2 est un cas particulier de la réponse 1. Autrement dit, si 2 est vrai, alors 1 est vrai aussi. Ou en termes de diagrammes de Venn, l'ensemble des cas où la réponse 2 est vraie est un sous-ensemble de l'ensemble des cas où la réponse 1 est vraie. En termes probabilistes, tout ce que j'ai dit là se résume à l'inégalité $\mathbb{P}[\text{Banque et Feministe}] \leq \mathbb{P}[\text{Banque}]$. La probabilité que deux événements se produisent tous les deux est toujours inférieure à la probabilité que l'un des événements se produise. Les mathématiques sont là implacables. La réponse 1 est la *bonne* réponse.

Si vous n'avez pas donné la bonne réponse, vous pouvez donc trouver refuge dans le fait que vous faites partie de la majorité. Peu parviennent à effectuer le raisonnement mathématique ci-dessus. Tversky et Kahneman pensent qu'au lieu de cela, les gens effectuent un raisonnement par association. La question qu'ils se posent n'est pas tant celle de la probabilité des réponses 1 et 2, mais celle de la représentativité des réponses 1 et 2 vis-à-vis de la description de Linda en préambule. La réponse 2 semble ainsi plus « représentative » des femmes de 31 ans, célibataires, ayant fait des études et participé à des mouvements contre la discrimination.

Les préjugés au secours de Linda*

Il y a peut-être une autre façon plus éclairante de comprendre l'incroyable taux d'échec au problème de Linda, qui nous ramène tout droit au débat qui oppose les *fréquentistes* aux *bayésiens*. Pour le *pur fréquentiste* féru de *p-value*, le test d'hypothèse consiste à étudier la probabilité des données sachant l'hypothèse. Ainsi, le *pur fréquentiste* va davantage s'intéresser aux probabilités $\mathbb{P}[\text{Preamble}|1]$ et $\mathbb{P}[\text{Preamble}|2]$, où 1 et 2 désignent les réponses 1 et 2, respectivement.

Il est alors raisonnable de penser que le préambule est plus probable si l'on suppose 2 que si l'on suppose 1. De façon formelle, on a l'inégalité $\mathbb{P}[\text{Preamble}|1] \leq \mathbb{P}[\text{Preamble}|2]$. En termes statistiques, le préambule est plus *vraisemblable* pour la réponse 2, ou encore, la réponse 2 est le *maximum de vraisemblance*.

Malheureusement, la « méthode scientifique » dont on a parlé dans le chapitre précédent s'intéresse bien souvent uniquement à ces *vraisemblances*. La terminologie n'aide pas. Elle nous pousse au contre-sens. Elle nous invite à confondre

la *vraisemblance* des données sachant les hypothèses avec la probabilité des hypothèses. Pour la *pure bayésienne*, c'est là que réside le sophisme des *purs fréquentistes*.

Cependant, ces vraisemblances, que j'ai préféré appeler *termes d'expérience de pensée* au chapitre 3, ne sont que l'une des composantes de la formule de Bayes. Pour la *pure bayésienne*, la quantité importante est la probabilité inverse. Dans le cas du problème de Linda, cette probabilité inverse est celle des réponses sachant le préambule. Comme vous commencez à le comprendre, cette probabilité inverse se déduit de la formule de Bayes. Pour la réponse 1, cette formule du savoir s'écrit comme suit :

$$\mathbb{P}[1|\text{Preamble}] = \frac{\mathbb{P}[\text{Preamble}|1]\mathbb{P}[1]}{\mathbb{P}[\text{Preamble}]} = \frac{\mathbb{P}[\text{Preamble}|\text{Banque}] \cdot \mathbb{P}[\text{Banque}]}{\mathbb{P}[\text{Preamble}]}.$$

De façon similaire, pour la réponse 2, la probabilité de la réponse sachant le préambule s'écrit

$$\mathbb{P}[2|\text{Preamble}] = \frac{\mathbb{P}[\text{Preamble}|\text{Banque et Feministe}] \cdot \mathbb{P}[\text{Banque et Feministe}]}{\mathbb{P}[\text{Preamble}]}.$$

La *pure bayésienne* peut désormais conclure en comparant ces deux probabilités. De façon cruciale, peu importent ses estimations des quantités inconnues dans les expressions de droite, pourvu que ces quantités obéissent aux lois des probabilités, la *pure bayésienne* conclura *toujours* que la première probabilité $\mathbb{P}[1|\text{Preamble}]$ est supérieure à celle de $\mathbb{P}[2|\text{Preamble}]$. Elle donnera toujours la bonne réponse — contrairement au *pur fréquentiste*.

Mieux, on peut calculer à quel point la *pure bayésienne* trouvera l'hypothèse 1 plus probable que l'hypothèse 2, sachant le préambule. En effet, en jouant avec les lois des probabilités — ce que je vous invite à faire et refaire chez vous — on voit que son second calcul est équivalent au suivant :

$$\mathbb{P}[2|\text{Preamble}] = \mathbb{P}[\text{Feministe}|\text{Preamble et Banque}] \cdot \mathbb{P}[1|\text{Preamble}].$$

En d'autres termes, quelles que soient les hypothèses de la *pure bayésienne*, celles-ci se devront d'être conformes aux lois des probabilités. Dès lors, la *pure bayésienne* conclura inéluctablement que, sachant le préambule, la probabilité de l'hypothèse 2 est $\mathbb{P}[\text{Feministe}|\text{Preamble et Banque}]$ -fois celle de l'hypothèse 1. Toute probabilité étant inférieure à 1, on en déduit que, sachant le préambule et quel que soit le modèle (bayésien) envisagé, l'hypothèse 2 sera *toujours* moins probable que l'hypothèse 1. Et la comparaison des probabilités des deux hypothèses se réduit au calcul d'une expérience de pensée — à savoir déterminer la probabilité que Linda soit active dans le mouvement féministe, sachant le préambule et le fait qu'elle soit caissière dans une banque.

Les préjugés sont indispensables

Vous avez sans doute eu l'impression que le raisonnement de la *pure bayésienne* est ici beaucoup trop compliqué. Après tout, il suffisait de constater que la réponse 2 impliquait la réponse 1 pour déterminer la bonne réponse au problème de Linda. Vous auriez tout à fait raison. Si je vous ai présenté le raisonnement de la *pure bayésienne*, ce n'est pas pour que vous compreniez mieux le problème de Linda ; c'est pour vous montrer en quoi le raisonnement de la *pure bayésienne* diffère de celui du *pur fréquentiste*. La *pure bayésienne* ne se contente pas du calcul de la vraisemblance. Elle inclut aussi dans son analyse les *a priori* qu'elle a sur les réponses 1 et 2.

En termes plus formels, la quantité importante pour la *pure bayésienne* n'est pas la vraisemblance $\mathbb{P}[\text{Preamble}|1]$, mais la probabilité inverse $\mathbb{P}[1|\text{Preamble}]$. Celle-ci se déduit de la vraisemblance via la formule de Bayes :

$$\mathbb{P}[1|\text{Preamble}] = \frac{\mathbb{P}[\text{Preamble}|1]\mathbb{P}[1]}{\mathbb{P}[\text{Preamble}]}.$$

En particulier, la vraisemblance $\mathbb{P}[\text{Preamble}|1]$ doit s'accompagner de l'*a priori* $\mathbb{P}[1]$.

Désormais, dans ce livre, je vais prendre le risque de remplacer la terminologie technique qu'est l'*a priori* par un synonyme à connotation négative : le *préjugé*. Il s'agit là d'un effort d'honnêteté intellectuelle de ma part. Sachant que je suis victime d'un biais cognitif pro-bayésien, je vais chercher à défendre une version péjorative de la philosophie bayésienne, et à vous convaincre que même cette version est séduisante — par opposition avec l'approche *bullshit* qui consisterait à jouer uniquement sur la connotation des mots pour persuader. N'oubliez pas ce que signifie littéralement préjugé : il s'agit d'un jugement avant l'observation.

Les *préjugés*, donc, sont au cœur du débat qui oppose les *bayésiens* aux *fréquentistes*. Ces préjugés sont la principale cause du rejet de la formule de Bayes au cours des deux derniers siècles. Les sciences se voulaient objectives. Or les préjugés semblent nécessairement devoir être subjectifs. Pour les *fréquentistes* et pour la plupart des scientifiques, ces préjugés subjectifs sont la déficience fatale de la philosophie bayésienne.

Cependant, la *subjectivité n'a rien d'arbitraire*. Et s'ils sont subjectifs, les préjugés bayésiens ne sont absolument pas arbitraires ! Ils obéissent aux lois des probabilités et, dans l'idéal, découlent (en partie) de calculs de la formule de Bayes. À l'instar de leur rôle dans le problème de Linda, la *pure bayésienne* considère même que ces préjugés sont la force du raisonnement bayésien — pourvu que ces préjugés soient justement bayésiens. Pour elle, les préjugés sont indispensables pour raisonner juste. Les préjugés forment les fondations de la rationalité. En effet, *sans préjugé, aucune conclusion ne peut être tirée*. Telle est l'affirmation la plus controversée de la philosophie bayésienne.

Le soleil d'xkcd

Face à une affirmation aussi contre-intuitive et controversée, l'exemple de Linda ne suffit pas. Je vous propose donc d'étudier une formidable expérience de pensée proposée par Randall Munroe¹, que je vais réinterpréter à ma sauce.

Imaginez-vous à Paris. Votre stagiaire est à Hawaii. Juste avant minuit, il lancera deux dés. S'il tombe sur un double six, il vous dira que le soleil a disparu. Sinon, il vous dira *si* le soleil a disparu. Minuit sonne. Votre stagiaire appelle et dit que le soleil a disparu. Que pouvez-vous conclure ?

Souvenez-vous. Pour conclure, la *méthode scientifique* doit rejeter l'hypothèse alternative. Pour conclure que le soleil a disparu, il faut rejeter l'hypothèse alternative ☀️ selon laquelle le soleil est toujours là. Et pour rejeter ☀️, un *pur fréquentiste* devra calculer la *p-value* associée à ☀️. Autrement dit, il calculerait la probabilité $p = \mathbb{P}[\text{red}| \odot]$ de ce que vous savez² sachant que le soleil n'a pas disparu. Une *p-value* très faible reviendrait à dire que la donnée red est hautement improbable sous l'hypothèse ☀️. Ce qui justifierait de rejeter ☀️.

Or, sachant ☀️, pour que vous ayez reçu l'appel téléphonique red que vous avez reçu, il faut que votre stagiaire soit tombé sur un double six. La probabilité $p = \mathbb{P}[\text{red} | \odot]$ est donc égale à la probabilité d'un double six, qui est $p = 1/36 \approx 0,028$. En particulier, on a bien $p < 0,05$. On peut donc conclure. Le fait que votre stagiaire vous a dit ce qu'il a dit sachant ☀️ est hautement improbable. Il faut donc rejeter la théorie ☀️. Or, rejeter la théorie ☀️, c'est rejeter le fait que le soleil n'a pas disparu : on conclut donc que le soleil a disparu.

Incroyable ! En suivant les pas des *purs fréquentistes*, on en est arrivé à une conclusion absurde : juste parce que notre stagiaire nous a dit que le soleil a disparu, on doit conclure qu'il a en effet disparu ! Randall Munroe conclut avec une réplique d'une *pure bayésienne* amusée réagissant à la conclusion du *pur fréquentiste* : « Je te parie 50 \$ que non ». Étrangement, la conclusion de notre *pure bayésienne* semble beaucoup plus sensée que celle de celui qui a pourtant suivi la « méthode scientifique » prescrite par les *fréquentistes*.

Les préjugés au secours d'xkcd

Fort heureusement, qu'ils le sachent ou non, les scientifiques ont des élans bayésiens. Aucun de ceux à qui j'ai présenté cette expérience de pensée n'a réagi en disant que la *pure bayésienne* était ridiculement irrationnelle. Pour comprendre pourquoi, il faut s'intéresser à la manière dont elle détermine ses crédences en diverses théories. Pourquoi, même après le coup de fil du stagiaire, notre *pure bayésienne* croit-elle toujours davantage en l'existence du soleil ?

¹  Frequentists vs. Bayesians | xkcd | R. Munroe

² À savoir le fait red que votre stagiaire vous a dit que le soleil a disparu.

La réponse réside bien sûr dans le préjugé de la *pure bayésienne*. Ce préjugé n'est pas arbitraire. Il est en fait déterminé par l'ensemble de toutes ses observations passées. Appelons *Vecu* son vécu. La formule de Bayes conduit alors à

$$\mathbb{P}[\text{soleil et Vecu}] = \frac{\mathbb{P}[\text{soleil et Vecu}]\mathbb{P}[\text{soleil}|Vecu]}{\mathbb{P}[\text{soleil}|Vecu]}.$$

Les détails peuvent paraître effrayants. Vous pouvez les ignorer. Mais j'aimerais attirer votre attention sur le terme $\mathbb{P}[\text{soleil}|Vecu]$. Il s'agit du préjugé de la *pure bayésienne* avant l'appel du stagiaire. C'est un *a priori* conditionné par le vécu de la *pure bayésienne*. Ce vécu inclut l'observation d'un soleil qui s'est levé tous les jours jusque-là. Mais il inclut aussi les cours de physique. Ces cours affirment que le soleil est une boule de plasma, alimentée énergétiquement par de la fusion nucléaire de noyaux d'hydrogène, et dont la quantité d'hydrogène est largement suffisante pour que le soleil brille encore plusieurs milliards d'années³.

En particulier, *la subjectivité n'a rien d'arbitraire* ! En particulier, dans ce cas, elle repose entre autres sur tous les travaux scientifiques des siècles précédents. Autrement dit, tous les *bayésiens* avec des vécus différents, mais suffisamment importants pour avoir vu le soleil se lever des milliers de fois, seront d'accord sur un point : la réaction du *pur fréquentiste* férus de *p-values* est une mauvaise réaction. En effet, pour tous ces *bayésiens*, la probabilité *a priori* $\mathbb{P}[\text{non soleil}|Vecu]$ que le soleil a disparu avant de recevoir l'appel du stagiaire est ridiculement faible. C'est ce préjugé qui conduit à n'accorder qu'une infime crédence en la disparition du soleil malgré l'appel du stagiaire.

De façon générale, aucune donnée ne doit être analysée de manière isolée. *Une donnée est tel un caillou. Elle n'a de valeur que si on l'ajoute à un édifice.*

Les préjugés au secours de Sally Clark

Histoire d'enfoncer encore un peu plus le clou, revenons-en à Sally Clark, dont on a parlé au chapitre 2. Souvenez-vous qu'en deux ans, Sally Clark eut le malheur de perdre deux de ses nouveau-nés, ce qui a conduit à des suspicions de double meurtre. Cependant, on avait vu que ce qui nous intéresse, c'est la probabilité $\mathbb{P}[\text{deux morts}|Vecu]$ de son innocence 🙄 sachant la double mort 💀 de ses nouveau-nés. On avait appliqué la formule de Bayes à ce cas :

$$\mathbb{P}[\text{deux morts}|Vecu] = \frac{\mathbb{P}[\text{deux morts}|Vecu]\mathbb{P}[\text{deux morts}]}{\mathbb{P}[\text{deux morts}]}.$$

L'approche de la *p-value* ou de la *vraisemblance* attirerait notre attention sur le terme d'expérience de pensée $\mathbb{P}[\text{deux morts}|Vecu]$ qui mesure la difficulté à expliquer la

³  *La mort du Soleil* | Sense of Wonder | S. Carassou et E. Ledolley (2015)

double mort des nouveau-nés sachant l'innocence de Sally Clark. Le sophisme du procureur, celui du pédiatre Roy Meadow, est un argument profondément fréquentiste, puisqu'il insiste sur l'incroyablement faible valeur de ce terme d'*expérience* de pensée. Il n'y a pas de doute. La vraisemblance, ou probabilité $\mathbb{P}[\text{innocence}]$ qu'une personne innocente voit ses deux premiers nouveau-nés mourir, est ridiculement faible. Meadow l'a estimée à un sur 70 millions, soit environ 0,000001 %. Voilà qui est en dessous du seuil de 0,00003 % de la physique ! À l'instar du procureur, le *pur fréquentiste* qui ne jugerait que par la *p-value* serait contraint de rejeter l'hypothèse 🤪 de l'innocence de Sally Clark. Il serait constraint de la condamner.

Si le professeur de mathématiques Ray Hill a contesté cette décision, c'est en vertu d'un argument profondément bayésien. Pour Hill, il faut appliquer la formule de Bayes, c'est-à-dire, en particulier, prendre en compte la présomption d'innocence $\mathbb{P}[\text{innocence}]$. Ce n'est qu'après avoir inclus ce préjugé dans le raisonnement que l'on pourra mieux comprendre la situation de Sally Clark, et ainsi en venir à une meilleure décision à son sujet.

Contrairement à ce que soutiennent souvent certains tenants de la « méthode scientifique », la *pure bayésienne* prétend que les préjugés ne sont pas une faute dans nos raisonnements dont il faudrait nous débarrasser. Les préjugés sont *indispensables* pour raisonner correctement.

Les préjugés pour lutter contre les pseudo-sciences

Ce principe salvateur pour le problème de Linda, l'*expérience de pensée* d'xkcd et le procès de Sally Clark trouve de nombreuses applications dans la manière d'aborder des théories hautement probables ou hautement improbables.

La *pure bayésienne* ne prend ainsi même pas la peine d'écouter les discours de ceux qui prétendent pouvoir créer de l'énergie à partir de rien. Puisque le principe de conservation de l'énergie est lui-même un principe fondamental dans toutes les théories de la physique, elle a un préjugé très important sur l'impossibilité de créer de l'énergie à partir de rien. Qui plus est, sous cette hypothèse de conservation de l'énergie, il demeure néanmoins très probable que plusieurs personnes sur Terre soient convaincues pour des raisons erronées d'avoir réussi une telle expérience — on reviendra plus tard sur l'interprétation bayésienne de l'argument d'autorité. Nul besoin alors de regarder la vidéo. Les préjugés suffisent à rejeter l'*expérience*.

De la même façon, notre *pure bayésienne* a un préjugé très prononcé à l'encontre des théories pseudo-scientifiques sur le paranormal. La télékinésie qui serait capable de tordre des cuillères et les prémonitions capables d'anticiper le futur violent des principes fondamentaux de la physique. Or la *pure bayésienne* accorde une très grande crédence à ces principes fondamentaux. De plus, les nombreux cas avérés de fraude, et les nombreux biais cognitifs qui expliquent

pourquoi ceux qui ont assisté à ces expériences y croient, sont autant de facteurs qui amènent la *pure bayésienne* à ne pas modifier ses crédences en les pseudo-sciences, même lorsqu'un tenant tente de le convaincre. Être confronté à un tenant est en fait très probable même sous l'hypothèse selon laquelle ces pseudo-sciences sont erronées.

Ceci ne veut pas non plus dire que la *pure bayésienne* ne changera jamais d'avis. Mais pour cela, il va falloir une donnée D extraordinaire. Pour qu'il y ait un transfert de crédences, il faudra que $\mathbb{P}[D|A]$ soit très important pour une théorie alternative A , par opposition à des termes d'expérience de pensée $\mathbb{P}[D|T]$ pour toute théorie crédible T — or on verra au chapitre 10 que le biais de sélection fait que $\mathbb{P}[D|T]$ est en fait souvent très important, y compris pour des données D qui ont l'air mystérieuses⁴. Pour que la théorie alternative A devienne tout à coup aussi crédible que T , il faut en particulier que $\mathbb{P}[D|A]/\mathbb{P}[D|T]$ soit au moins aussi grand que le rapport inverse des crédences des théories *a priori*, à savoir⁵ $\mathbb{P}[T]/\mathbb{P}[A]$. Or, le cumul monumental des connaissances scientifiques au cours des derniers siècles fait que ce second rapport est souvent gigantesque. Comme le dirait Carl Sagan : « toute affirmation extraordinaire requiert des preuves extraordinaires. » Nous venons de démontrer la validité de ce principe à partir de fondements bayésiens !

Ou alors, pour amener un *bayésien* à changer radicalement d'avis, il va falloir lui proposer une théorie à laquelle ce *bayésien* n'a jamais pensé — cet argument ne tient pas pour convaincre la *pure bayésienne* puisque l'on verra au chapitre 7 qu'elle connaît *toutes* les théories (calculables). Les fondements de cette théorie alternative devront être aussi crédibles que les théories dominantes de la physique. De plus, cette théorie alternative devra expliquer aussi bien les observations physiques de l'Histoire des sciences. Enfin, elle devra mieux expliquer un phénomène en particulier. Comme ce fut le cas pour la trajectoire de Mercure et la relativité générale d'Einstein.

À l'inverse, la *pure bayésienne* peut avoir des crédences importantes en certaines affirmations, malgré un manque flagrant de données empiriques. Par exemple, si je dis à notre *pure bayésienne* que je me suis lancé dans l'ascension d'un sommet de l'Himalaya à plus de 6 000 mètres, et si je lui dis avoir énormément souffert pendant cette ascension, elle n'aura pas besoin de davantage de données empiriques pour croire que ce ne fut pas une balade de santé pour moi. Si la *pure bayésienne* a cru à ma fatigue à partir d'un simple témoignage, c'est parce qu'elle avait un gros préjugé sur mon incapacité à gravir des sommets sans cracher mes poumons.

⁴  *La sur-interprétation (overfitting)* | IA 11 | Science4All | C. Michel et L.N. Hoang (2018)

⁵ Je vous invite à prouver ceci à partir de la formule de Bayes !

Les préjugés au secours des sciences

Les préjugés ont aussi des conséquences très utiles sur l'approche des scientifiques confrontés à des anomalies empiriques. Ainsi, comme on en a déjà parlé, quand l'expérience OPERA a cru détecter des neutrinos voyageant à des vitesses supérieures à la vitesse de la lumière, l'annonce a été accueillie avec un scepticisme global. Par les physiciens théoriciens, mais aussi par les auteurs de l'expérience ! Ce n'est finalement à la surprise de personne que des failles dans l'expérience ont été découvertes par la suite. La crédence des physiciens en la relativité restreinte d'Einstein — qui affirme qu'aucune particule ne peut dépasser la vitesse de la lumière — est si énorme que l'hypothèse d'une erreur expérimentale est plus crédible que celle d'une violation de la théorie d'Einstein.

Ce qui est peut-être plus étrange, c'est que même les mathématiciens peuvent aussi avoir des crédences importantes en des théorèmes mathématiques non prouvés. Ainsi, aujourd'hui, la très grande majorité des théoriciens des nombres croient en la fameuse hypothèse de Riemann⁶, que beaucoup considèrent comme étant le plus prestigieux problème ouvert des mathématiques. À tel point que, de nos jours, un très grand nombre de théorèmes partent du postulat selon lequel l'hypothèse de Riemann est vraie pour en explorer les conséquences.

De la même façon, les informaticiens ont une crédence très favorable en la conjecture $P \neq NP$, que beaucoup considèrent comme le plus prestigieux problème ouvert de l'informatique théorique. Ces crédences, même si elles ne sont pas fondées sur des preuves mathématiques rigoureuses et indisputables, restent néanmoins justifiables et justifiées, notamment étant donné une approche bayésienne. Je ne peux que vous conseiller de lire l'excellent billet⁷ de blog de Scott Aaronson à ce sujet pour mieux comprendre l'origine des crédences de théoriciens vis-à-vis de théorèmes mathématiques non prouvés. En fait, les crédences d'Aaronson sont si importantes, que la publication d'articles mathématiques prouvant $P = NP$ ou des résultats similaires ne modifie que très peu la crédence d'Aaronson en $P \neq NP$. Pour Aaronson, il est plus probable que ces articles soient erronés. L'histoire lui a donné raison jusque-là — ce qui n'a fait sans doute que consolider sa crédence en $P \neq NP$.

Des raisonnements identiques s'appliquent à des sujets plus controversés. En 2016, le gouvernement américain a décidé d'autoriser la commercialisation des champignons OGM sans procédure de test⁸. Cette décision a soulevé de nombreuses réactions contestataires en Europe, où les OGM ont moins bonne presse. Autoriser des OGM, c'est déjà controversé. Le faire sans procédure de test semble carrément scandaleux tant il s'agit d'une menace pour la santé publique !

Pourtant, l'annonce a été reçue avec joie et espoir par les scientifiques. Caixa Gao, chercheuse en biologie, annonça ainsi : « la communauté de la recherche

⁶  Deux (deux ?) minutes pour l'hypothèse de Riemann | El jj | J. Cottanceau (2016)

⁷  The Scientific Case for $P \neq NP$ | Shtetl-Optimized | S. Aaronson (2014)

⁸  Gene-Edited CRISPR Mushroom Escapes US Regulation | Nature | E. Waltz (2016)

sera très contente de cette actualité.» Ces scientifiques ne seraient-ils pas en train de jouer aux savants fous ? Sont-ils vraiment conscients des risques potentiels sur la santé des populations ? Ne seraient-ils pas en train de créer le monstre de Frankenstein ?

Pour comprendre le point de vue de ces scientifiques, le plus simple est de raisonner en termes de préjugés (fondés). Pour commencer, il faut savoir que les organismes sont constamment génétiquement modifiés. En effet, à chaque reproduction, les gènes des deux sexes se combinent de sorte à former un nouveau brin d'ADN qui, de façon quasi certaine, n'avait encore jamais existé. À cela, s'ajoutent en plus les mutations de l'ADN qui s'accumulent pour toujours plus modifier génétiquement les organismes.

Dans la nature, il se produit ensuite un phénomène de sélection naturelle, qui favorise certains organismes génétiquement modifiés par rapport à d'autres. Dans l'agriculture, il se produit un phénomène similaire de sélection artificielle. Tout comme la sélection naturelle, ce phénomène favorise certains organismes par rapport à d'autres. Pendant des millénaires, c'est cette sélection artificielle qu'ont subi les espèces domestiquées, animales et végétales. Celle-ci a bouleversé les êtres vivants ainsi sélectionnés, transformant des loups agressifs en chihuahuas dociles et des petites bananes sauvages dégoûtantes pleines de gros noyaux en celles dont nous nous délectons quotidiennement. Tous les organismes qui nous entourent sont, par opposition à ceux que l'on trouvait il y a quelques siècles, des organismes génétiquement modifiés.

Mais ce n'est pas tout. À cela s'ajoutent d'autres modifications de la biodiversité locale dues, entre autres, à l'exploitation à grande échelle de la monoculture, l'importation de nombreuses espèces venant de l'autre bout du monde, l'utilisation accrue des pesticides et des insecticides, voire des techniques plus récentes d'irradiation par ultraviolets qui ont pour but d'accélérer les mutations génétiques.

Toutes ces modifications des génomes sont rapides et peu contrôlables. Il y a beaucoup d'incertitudes vis-à-vis des espèces mutées. Voilà qui conduit notre *pure bayésienne* à émettre des doutes *a priori* quant à leurs effets potentiellement nuisibles sur la santé — même si des méta-analyses sur plusieurs décennies et croisant de nombreuses recherches scientifiques suggèrent fortement que les OGM ne sont pas plus dangereux pour la santé publique que les agricultures traditionnelles.

Mais par ailleurs, depuis 2012, les chercheurs en biologie ont découvert une toute nouvelle façon d'éditer les génomes du vivant. Cette technologie appelée CRISPR Cas9 permet d'éditer lettre par lettre le génome. Autrement dit, elle permet de savoir exactement quelles modifications du génome ont été effectuées. Du coup, la *pure bayésienne* considère que l'OGM obtenu par CRISPR Cas9 avec un protocole contrôlé et justifié est, *a priori*, beaucoup plus fiable que celui obtenu par des méthodes où la modification du génome est beaucoup moins contrôlée. Voilà qui explique que l'autorisation de commercialisation sans test

d'OGM CRISPR Cas9 n'a soulevé que très peu d'inquiétude chez les scientifiques experts de la question. La *pure bayésienne* a un préjugé justifié très fort en le fait que de tels OGM sont beaucoup moins dangereux que les variétés obtenues via des méthodes traditionnelles.

Pour les OGM comme pour les diagnostics médicaux, pour Linda comme pour la justice, pour les sciences expérimentales comme pour les sciences théoriques, notre *pure bayésienne* ne peut pas bien raisonner sans utiliser ses préjugés. Ces préjugés sont sa botte secrète. Ils sont la raison du succès de ses prédictions.

Le bayésien a un préjugé sur tout

Considérons une hypothèse H peu étudiée au sujet de laquelle on ne dispose de presque aucune donnée⁹. Faut-il croire H ? Plutôt que de risquer de se tromper, certains scientifiques affirment qu'il est préférable de dire « je ne sais pas », voire « je n'en sais rien ». Certains bayésiens précisent qu'il est raisonnable de considérer un *a priori* dit *non-informatif*. Celui-ci prend souvent la forme d'une distribution *uniforme*, c'est-à-dire qui ne privilégie aucune hypothèse. Ainsi, si H est vraie ou fausse, il semble raisonnable de supposer qu'*a priori*, H a une probabilité 1/2 d'être vraie — ou plutôt, il faut lui accorder une crédence *a priori* de 1/2. Mais ceci me semble très problématique, pour plusieurs raisons.

La première raison est la plus fondamentale. Cette posture est généralement incompatible avec le bayésianisme. En effet, même si l'on n'a que très peu étudié H , bien souvent, H reste liée à des questions qu'on a étudiées, et au sujet desquelles de nombreuses données ont été collectées. C'est typiquement le cas de tous les exemples de ce chapitre. Qu'il s'agisse de Linda, du soleil, de procès, de complots, de théorèmes ou d'OGM, nous disposons tous de beaucoup de réflexions préalables et de données sur ces sujets. Pire encore, le paradoxe de Stein, dont on parlera au chapitre 13, démontre l'*inadmissibilité* (au sens statistique) du morcellement du savoir. Autrement dit, pour un bayésien (dont le raisonnement est *admissible*), il est *impossible* que des vécus sur des sujets d'apparence distants n'aient *aucun* effet sur les crédences en H . Dès lors, il est très improbable que l'on ait *exactement* $\mathbb{P}[H|\text{Vecu}] = 1/2$.

La deuxième raison est motivationnelle. Dire « je ne sais pas », c'est céder à la paresse. Comme le disait Poincaré, « douter de tout ou tout croire sont deux solutions également commodes, qui l'une et l'autre nous dispensent de réfléchir ». En particulier, déterminer l'*a priori* $\mathbb{P}[H|\text{Vecu}]$ est un calcul subtil et difficile. Mais ce n'est techniquement qu'un calcul. Et à défaut d'être une *pure bayésienne* capable de l'effectuer instantanément, il semble irrationnel de ne pas prendre le temps d'en trouver une bonne approximation. Bien sûr, il ne faudra pas oublier que nos calculs heuristiques ne sont qu'une grossière approximation.

⁹Pour fixer les idées, vous pouvez considérer l'hypothèse H qui dit que le Big Bang n'est qu'un rebond d'un univers plus ancien.

En particulier, la validité de l'approximation pourrait donc être remise en cause par des arguments purement théoriques — typiquement ceux qui montrent que certains calculs qui ont été négligés ont en fait une grande incidence sur la qualité de l'approximation. Mais de façon cruciale, pour trouver la motivation de se lancer dans ces calculs difficiles et piégeux, il faut absolument commencer par se convaincre que la réponse « je ne sais pas » n'est pas une réponse satisfaisante.

La troisième raison est pédagogique. Si l'on n'émet jamais de préjugé, alors on ne sera jamais exposé à l'erreurs. Et on ne se rendra alors pas compte de notre ignorance¹⁰ et des biais de nos préjugés inconscients¹¹. Pour combattre l'excès de confiance, il me semble bien plus préférable de ne pas écarter les occasions de contredire nos préjugés. Au contraire, il me semble souhaitable de finir des phrases comme « je me trompe peut-être, mais je parierais que », « c'est sans doute naïf, mais il me semble que », ou « avant de découvrir X , je pensais que ». En rendant nos préjugés explicites, on met ainsi plus facilement en évidence leur inadéquation avec les données empiriques. Il est alors plus facile de changer d'avis. Et on peut plus aisément prendre l'habitude d'avoir des croyances dynamiques. *L'apprentissage est une danse*. Dansons donc. C'est ainsi qu'on pourra plus facilement identifier les cas où on peut raisonnablement faire confiance à notre intuition, par opposition à ceux où notre intuition n'est vraiment pas fiable. Dans ces cas, il sera alors plus facile de déférer notre jugement à un modèle mathématique ou à une autorité reconnue.

La quatrième raison est ludique. Oui parce que se rendre compte qu'une prédictions tombe juste est plaisant — il n'y a qu'à voir les physiciens répéter des expériences dont ils connaissent pourtant l'issue ! Mais surtout, découvrir qu'une intuition très convaincante est erronée peut être d'une jouissance exquise¹². C'est ce qui est arrivé quand les physiciens ont cru avoir détecté des neutrinos plus rapides que la lumière¹³, ou quand les mathématiciens ont découvert que les nombres premiers successifs ne se comportent pas de manière aléatoire¹⁴ ! Comme le disait Isaac Asimov, « la phrase la plus excitante à entendre en science [...] n'est pas “eurêka”, mais “c'est bizarre” ». Celui qui n'a jamais vécu l'extase de la découverte d'un fait contre-intuitif ne comprendra jamais cette raison de vivre des scientifiques. Or la clé de cette extase n'est pas le fait contre-intuitif. La clé est le préjugé très crédible et pourtant rejeté.

En pratique, malheureusement, notre environnement social, éducatif et professionnel a tendance à stigmatiser les erreurs. Nous avons peur des erreurs. Et c'est pour cela que dire « je ne sais pas » ou $\mathbb{P}[H|\text{Vecu}] = 1/2$ est une échappatoire si populaire. Mais ceci a des conséquences très néfastes, notamment sur l'apprentissage des mathématiques. Parce que les mathématiques sont le domaine où il est le plus facile de reconnaître les erreurs, pour éviter toute

¹⁰ *La rationalisation | La Tronche en Biais | V. Tapas et T. Durand (2015)*

¹¹ *Tous racistes ? Les biais implicites | Science Étonnante | D. Louapre (2017)*

¹² *Top 8 des monstres mathématiques | Infini 11 | Science4All | L.N. Hoang (2016)*

¹³ *Neutrinos slower than light | Sixty Symbols | E. Copeland and T. Padilla (2012)*

¹⁴ On en reparlera au chapitre 14.

erreur, il est tentant de se taire et de ne jamais révéler ses préjugés sur les conjectures mathématiques. Pire, ceci conduit beaucoup à subir un « blocage » quand il s'agit de mathématiques, à tel point que le syndrome de « l'anxiété mathématique » a désormais sa propre page Wikipedia anglophone. Loin de n'affecter que les « mauvais » en mathématiques, ce syndrome cause une chute des aptitudes en mathématiques auprès de tous ceux qui craignent trop l'erreur.

Par opposition, mon professeur de mathématiques en première année de classes préparatoires, celui qui m'a fait découvrir toute la joie de faire des mathématiques, n'hésitait pas à dire « j'y crois » ou « je n'y crois pas » pour juger nos explications mathématiques — et il ne s'agissait généralement pas d'euphémismes ! Ce fut une délivrance pour moi de me rendre compte de cela. On *peut* prendre les mathématiques à la légère, et parier sur la vérité d'un théorème ou la validité d'une (idée de) preuve. Et ça veut aussi dire qu'on peut se planter ce faisant. Mais c'est souvent génial de se tromper ainsi¹⁵ ! Ce sont des telles erreurs de notre intuition qui la font progresser — et constater les progrès de l'intuition mathématique est également extrêmement jouissif ! Bref. En mathématiques, peut-être plus encore qu'ailleurs, la célébration des erreurs de l'intuition semble être une étape indispensable à un apprentissage sain et efficace.

Cette glorification des erreurs peut toutefois vous surprendre et vous embêter. Un médecin qui prescrirait un mauvais médicament ou découperait une veine par erreur ne devrait pas être célébré. De même, on a tendance à penser qu'un politicien qui fait son *mea culpa* ne devrait pas recevoir une promotion. Et plutôt que dire des bêtises ou de polluer un débat avec idioties, quand il s'agit d'une prise de parole publique avec des conséquences potentiellement majeures, affirmer « je ne sais pas » semble être une bonne stratégie rhétorique. Cela peut notamment permettre d'insister sur la difficulté des problèmes en jeu.

Il est vrai. Mais il est important de voir que dans tous ces exemples, il y a une dimension morale (ou stratégique) qui entre en jeu. Or le bayésianisme n'est pas une philosophie morale. Il n'a donc rien à dire sur ce qui est *moralement souhaitable* de faire, ou ce qu'il faut faire d'un point de vue égoïste. Le bayésianisme est une philosophie du savoir. Son but est d'organiser l'apprentissage et le savoir. Et c'est en ce sens qu'il me semble souhaitable de célébrer nos erreurs.

Plus généralement, le bayésianisme *impose* le calcul de prédictions (probabilistes) sur *tout* et n'importe quoi¹⁶. Le bayésien a *toujours* un préjugé¹⁷. Il ne peut pas dire « je ne sais pas ». La probabilité $\mathbb{P}[H|\text{Vecu}]$ a toujours une valeur exacte. Et il est improbable qu'elle soit égale à 1/2. Selon le bayésianisme, balayer ce préjugé sous le tapis serait très irrationnel¹⁸.

¹⁵  *Le bonheur de faire des erreurs* | My4Cents (Sceaux) | Science4All | L.N. Hoang (2016)

¹⁶ Thibaut Giraud parle de *pari épistémique*. C'est le pari qu'on ferait, pistolet sur la tempe. On peut hésiter et « flipper ». Mais selon Giraud, à moins d'estimer exactement $\mathbb{P}[H|\text{Vecu}] = 1/2$, il serait irrationnel de lancer une pièce plutôt que de parier *H* ou *non H*.

¹⁷ Bien sûr, il a aussi un préjugé sur la fiabilité de ce préjugé, notamment par opposition aux préjugés d'un expert ou d'un modèle qu'il ne connaît pas encore.

¹⁸  *Perception bayésienne* | Axiome 9 | T. Giraud et L.N. Hoang (2018)

Les préjugés erronés

Cette glorification des préjugés a de quoi surprendre. Après tout, dans le langage quotidien, le mot « préjugé » est connoté très négativement. Et pour cause. Les préjugés semblent inéluctablement conduire à de la discrimination, du racisme ou du sexism. Cependant, il y a une distinction essentielle à faire entre toute philosophie du savoir, et l'utilisation de ce savoir.

Avant d'en venir à ce problème crucial, il est bon de rappeler que beaucoup de préjugés sont des préjugés non-bayésiens, dans le sens où ils n'ont pas été obtenus par (une approximation de) la formule de Bayes. De nombreux préjugés sont même incohérents avec les lois des probabilités. On l'a vu. Même les plus grands mathématiciens sont incapables d'appliquer correctement la formule de Bayes, même dans des cas simples. Nos préjugés sont tous mal fondés, surtout si l'on n'a pas pris le temps de longuement méditer les origines de nos préjugés.

L'une des failles de raisonnement récurrentes est le manque de contextualisation des crédences bayésiennes. Par exemple, les films et séries télévisées présentent souvent le stéréotype selon lequel les Asiatiques ne savent pas conduire. De façon amusante, on peut avancer de bonnes raisons de croire en ce stéréotype. La très grande majorité des Asiatiques ont grandi dans des pays en voie de développement où apprendre à conduire n'est pas à la portée de tous. Au Vietnam, où je suis né, le traffic est de toute façon si chaotique qu'il est plus pratique de se mouvoir en mobylette ou en taxi. Il semble donc raisonnable d'affirmer que les Asiatiques conduisent moins bien que les Américains pour qui la voiture est reine. Cependant, si vous considérez un Asiatique qui a grandi dans un pays occidental (comme moi !), l'écart du niveau de conduite avec l'Américain est beaucoup moins net, sinon inexistant. *Sans conteste, sans contexte, c'est la mauvaise probabilité qu'on teste.*

De la même façon, André Kuhn, professeur de criminologie et de droit pénal aux universités de Lausanne, Neuchâtel et Genève, affirme que les étrangers sont associés à un plus haut taux de criminalité. Si je dis à la *pure bayésienne* de penser à un local et à un étranger, elle aura donc une crédence légèrement plus grande en la criminalité de l'étranger qu'en celle du local. Cependant, si je lui dis maintenant que ces deux personnes sont toutes deux jeunes, de sexe masculin et de milieu socio-économique modeste, alors André Kuhn affirme qu'elle aura maintenant des crédences similaires concernant les criminalités des deux individus. En fait, la raison pour laquelle la *pure bayésienne* avait une crédence légèrement supérieure en la criminalité de l'étranger, c'est surtout parce que l'étranger a une plus grande probabilité d'être jeune, de sexe masculin et de milieu socio-économique modeste. Ces trois facteurs sont les principaux facteurs de risque de criminalité. Cependant, une fois que l'on a contextualisé les deux individus, le fait que ces individus soient locaux ou étrangers n'a plus qu'un effet négligeable sur la crédence de la *pure bayésienne* en leur criminalité¹⁹.

¹⁹  Interview d'André Kuhn : les sciences criminelles | Podcast Science (2011)

De façon plus générale, on a tendance à aimer décrire, publier et lire des relations comme « A cause B ». En termes probabilistes, ceci revient à dire que $\mathbb{P}[B|A]$ est très supérieur à²⁰ $\mathbb{P}[B|\text{non } A]$. Cependant, ces phrases générales ne s'appliquent pas en général aux individus. En effet, chaque individu A possède toutes sortes de caractéristiques Z qui le distinguent de l'individu générique A . La quantité qui s'applique à cet individu est alors $\mathbb{P}[B|A, Z]$, et non pas $\mathbb{P}[B|A]$. Or, ces deux quantités peuvent être très différentes. Typiquement, l'indice de masse corporelle (IMC) est utile pour décrire des populations, même s'il n'a pas nécessairement un sens conclutif pour un individu en particulier²¹. Le problème, bien sûr, c'est que lister toutes les quantités $\mathbb{P}[B|A, Z]$ pour les différentes valeurs de Z a peu de chance de faire la une des journaux.

Pire, de tels liens de causalité peuvent n'avoir qu'un intérêt limité en pratique. En effet, la question intéressante est souvent davantage de savoir si un individu Z peut espérer obtenir B s'il se met à faire A . Or, se mettre à faire A , appelons cela A' , n'est pas la même chose que faire A . Typiquement, ceux qui font de la musculation peuvent soulever de grosses masses, mais une fois que vous aurez fait une de leurs séances de musculation, vous risquez de ne plus pouvoir soulever quoi que ce soit ! La quantité pertinente pour conseiller l'individu Z est alors $\mathbb{P}[B|Z, A']$, et non pas $\mathbb{P}[B|Z, A]$. Malheureusement, $\mathbb{P}[B|Z, A']$ est souvent beaucoup plus difficile à estimer que $\mathbb{P}[B|Z, A]$, et encore plus que $\mathbb{P}[B|A]$!

Une autre cause de nombreux préjugés inadéquats est le biais dans les observations avec lesquelles les crédences sont mises à jour. Il s'agit là encore essentiellement d'un problème de contextualisation. Il faut se rendre compte que ces observations ont été biaisées par un contexte, et qu'elles peuvent être sujettes à ne pas être représentatives d'observations similaires dans des contextes différents. Quand je dis à un guide indonésien que j'étais mathématicien, il répondit à ma grande surprise : « mais... tu n'es pas vieux ! » Bizarrement, l'image la plus répandue du mathématicien est celle d'un vieux sage barbu. Pourtant, comme je le lui rétorquai : « Tout vieux mathématicien a été jeune. » Je n'ai pas eu besoin d'en formaliser une preuve.

Les généralisations excessivement promptes, la sur-interprétation de caractéristiques non-pertinentes et les biais de représentations ne sont cependant que quelques-unes des très nombreuses causes de l'inexactitude de nos préjugés. À cela se rajoutent d'autres biais cognitifs bien connus, comme le biais de confirmation, la dissonance cognitive ou encore l'aisance cognitive. En particulier, une cause majeure de l'inexactitude de nos préjugés est le fait que ces préjugés reposent bien souvent davantage sur des on-dit que sur des données empiriques, et que nos crédences dépendent plus souvent de l'affection et de l'admiration que l'on a pour le charisme, la rhétorique et l'image de ceux qui cherchent à nous convaincre, que du calcul (même approximatif) de la formule de Bayes.

²⁰On verra au chapitre 13 que ce n'est pas tout à fait ce que Fisher entendait par « A cause B », mais la critique ci-dessus s'applique encore avec la définition de Fisher.

²¹  I.M.C. - "Être gros ?" | Risque Alpha | T. Le Magoarou (2018)

Il est ainsi souvent reproché aux bayésiens d'avoir le pouvoir de biaiser leurs conclusions via le choix des préjugés. En effet, pour toute conclusion, il semble exister un préjugé qui conduit à cette conclusion. Je le concède volontiers. L'approche bayésienne se prête tout simplement mal aux débats entre hooligans à l'avis tranché, et dont l'objectif annoncé est de gagner le débat. Si votre objectif est de persuader et de gagner les faveurs du plus grand nombre, je vous recommande davantage de vous tourner vers l'art de la rhétorique, du *putaclic* et de la provocation²², et d'oublier (presque) toute philosophie du savoir. Ce n'est pas la logique du premier ordre qui vous permettra de récolter des signatures et de gagner des voix.

Ce problème est également critique quand il s'agit de tester les produits d'entreprise avec des intérêts économiques majeurs, comme c'est le cas des tests critiques, des *crash* tests ou des labels de qualité. Dans tous ces cas, on a tendance à préférer une procédure simple et explicite. Or les préjugés adéquats sont souvent extrêmement complexes à modéliser, à décrire et à comprendre. L'interprétabilité des procédures de test est malheureusement souvent incompatible avec les fondements du bayésianisme.

Ceci étant dit, si vous et vos interlocuteurs cherchez avant tout à mieux comprendre le monde, quitte à utiliser des modèles que vous n'appréciez pas, alors il me semble indispensable de commencer par expliciter les préjugés des uns et des autres, et de chercher à expliquer les origines de ces préjugés. Ce n'est qu'une fois que ces préjugés seront clarifiés et rendus suffisamment bayésiens que vous pourrez aller sereinement de l'avant, et appliquer la formule de Bayes pour converger vers les meilleures théories.

Bref. Nos préjugés ne sont essentiellement jamais bayésiens. Mais la présence de préjugés inexacts n'est absolument pas un argument contre la nécessité des préjugés bayésiens. Ce serait comme rejeter la logique déductive sous prétexte que personne ne comprend la contraposition.

Les préjugés et la morale

Ceci étant dit, la *pure bayésienne* n'exclut pas la possibilité que les origines d'une personne affectent malgré tout nos crédences sur ses caractéristiques physiologiques ou culturelles. Le vidéaste Léo Grasset affirme ainsi que²³, « comme les conditions de sélection varient au niveau local, les populations humaines sont différentes génétiquement entre elles. » Il serait scientifiquement infondé d'ignorer ces différences. Il serait irrationnel, au sens bayésien, de penser que les origines génétiques et sociales des individus n'affectent pas les crédences que l'on doit avoir en leurs diverses caractéristiques, compétences et habitudes.

²²  *Chère conviction, mute-toi en infection VIRALE !!!* Démocratie 7 | Science4All | L.N. Hoang (2017)

²³  *Des races dans l'humanité ?* Dirty Biology | L. Grasset (2016)

Là où le bât blesse, ce n'est pas tant l'existence de différences, mais davantage les jugements moraux souvent attachés aux différentes origines génétiques. La capacité des blancs européens à mieux digérer le lactose que les Asiatiques adultes ne semble pas justifier une supériorité sociale ou morale. Quant au quotient intellectuel et au niveau mathématique, il semble qu'il soit bien plus une conséquence du niveau économique et de la qualité de l'éducation que de nos génotypes. Et même si cela n'était pas le cas, penser qu'avoir un meilleur quotient intellectuel est une preuve de supériorité est une construction sociale qui repose sur des jugements moraux. La philosophie du savoir qu'est le bayésianisme n'a pas son mot à dire quant à ces jugements moraux.

Puisque les jugements moraux et l'éthique sont des sujets si cruciaux, nous les aborderons malgré tout dans le dernier chapitre de ce livre. D'autant que pour les philosophies morales normatives dites *conséquentialistes*, le bayésianisme joue un rôle central et incontournable. Cependant, pour conclure ce chapitre, je peux d'ores et déjà souligner une raison pour laquelle il n'est pas forcément souhaitable de mettre en avant nos crédences bayésiennes, même si celles-ci ont été calculées par la formule de Bayes. L'exemple le plus évident est celui de la tarte peu appréciable dont votre ami semble si fier. Vous n'êtes pas obligé de lui dire ce que vous en pensez vraiment. Bien souvent, il semble moralement justifié de mentir à son interlocuteur si ce mensonge lui permet de mieux se sentir — quoique Kant et ses adeptes ne soient pas d'accord.

Dans le cas des préjugés, il y a un exemple peut-être plus subtil mais tout aussi irritant, voire blessant. Quand j'étais gamin, j'étais l'un des deux seuls Asiatiques d'une école de plusieurs centaines d'élèves — je ne connaissais même pas l'autre Asiatique ! Et j'étais le plus petit de ma classe. Chacune des questions, des remarques et des blagues sur mes origines Asiatiques ou sur ma taille n'était pas forcément blessante — même si l'appellation « chine-toc » ou « petit Chinois » n'étaient pas des compliments flatteurs. Le problème, ce qui fut exaspérant, c'est qu'à chaque fois que je rencontrais un nouvel enfant, celui-ci répétait les nombreuses crédences qu'il avait à mon sujet. J'avais un traitement de (dé)faveur. J'étais constamment confronté aux mêmes stéréotypes. C'est cette répétition qui fut insupportable. *Non, je ne fais pas de kung-fu.*

Corriger un préjugé par an, c'est facile. C'est ce qui arrive à ceux dont les traits visuels ne sortent pas de l'ordinaire. Mais corriger vingt fois le même préjugé chaque jour sans soutien d'un pair, c'est beaucoup plus difficile, fatigant et désagréable. C'est d'ailleurs non sans un sourire en coin que j'ai pu constater l'agacement de mon guide népalais quand, pendant 15 jours, il dut expliquer encore et encore à d'autres Népalais que, non, je ne suis pas Népalais.

Mais il y a pire. Je ne suis clairement pas le plus à plaindre. D'autres sont régulièrement désavantagés par les préjugés (possiblement justifiés) qui sont associés à leur apparence physique ou à leurs origines. Ce problème est particulièrement saillant quand il s'agit d'entretiens d'embauche, d'examens judiciaires ou d'aide à la personne. Comme on en reparlera au moment d'aborder la théorie des jeux, il peut alors y avoir un contraste saisissant entre les incitatifs des

personnes en charge et les conséquences néfastes sur les victimes de préjugés. C'est pour éviter de telles conséquences néfastes qu'il est indispensable de fonder une philosophie morale adéquate. On en reparlera dans le dernier chapitre de ce livre.

À l'instar des innovations technologiques, les préjugés peuvent être utilisés à bon ou à mauvais escient. Ceci étant dit, et je ne peux pas insister dessus suffisamment, les préjugés sont avant tout indispensables à la réflexion. Ils sont une condition nécessaire à la rationalité. Selon notre *pure bayésienne*, seuls ceux qui utilisent leurs préjugés peuvent espérer prétendre être rationnels.

Références en français

- ➲ *Sommes-nous des criminels ? Petite introduction à la criminologie* | Les Éditions de l'Hèbe | André Kuhn (2005)
- ➲ *Interview d'André Kuhn : les sciences criminelles* | Podcast Science (2011)

- ▶ *La mort du Soleil* | Sense of Wonder | S. Carassou et E. Ledolley (2015)
- ▶ *La rationalisation* | La Tronche en Biais | V. Tapas et T. Durand (2015)
- ▶ *Deux (deux ?) minutes pour l'hypothèse de Riemann* | El Jj | J. Cottanceau (2016)
- ▶ *Des races dans l'humanité ? Dirty Biology* | L. Grasset (2016)
- ▶ *Tous racistes ? Les biais implicites* | Science Étonnante | D. Louapre (2017)
- ▶ *I.M.C. - "Être gros ?"* | Risque Alpha | T. Le Magoarou (2018)
- ▶ *Le bonheur de faire des erreurs* | My4Cents (Sceaux) | Science4All | L.N. Hoang (2016)
- ▶ *Top 8 des monstres mathématiques* | Infini 11 | Science4All | L.N. Hoang (2016)

Références en anglais

- ➲ *Thinking Fast and Slow* | SpringerFarrar, Straus and Giroux | D. Kahneman (2013)

- ➲ *Gene-Edited CRISPR Mushroom Escapes US Regulation* | Nature | E. Waltz (2016)
- ➲ *Frequentists vs. Bayesians* | xkcd | R. Munroe
- ➲ *The Scientific Case for $P \neq NP$* | Shtetl-Optimized | S. Aaronson (2014)

- ▶ *Neutrinos slower than light* | Sixty Symbols | E. Copeland and T. Padilla (2012)

[La formule de Bayes] n'avait aucune application de son vivant, mais aujourd'hui, grâce aux ordinateurs, elle est quotidiennement utilisée en modélisation du changement climatique, en astrophysique et en analyse des marchés financiers.

Bill Bryson (1951-)

Dans chaque non-bayésien, il y a un bayésien qui lutte pour se libérer.

Dennis Lindley (1923-2013)

6

Les prophètes du bayésianisme

Une histoire mouvementée

Les fréquentistes et leur gourou Ronald Fisher ont longtemps persécuté ceux qui semblaient à leurs yeux n'être qu'une secte obscure. Pendant deux siècles, les quelques fidèles du bayésianisme durent œuvrer dans le secret et n'osèrent pas avouer leurs convictions hérétiques en public. Interdit par le fréquentisme, le bayésianisme a même frôlé l'extinction à plusieurs reprises.

Mais en s'appuyant sur les textes anciens de Price et Laplace, une poignée de fervents apôtres réussirent à maintenir la flamme de la foi bayésienne. Ces bayésiens prophétiques ont su adapter le dogme bayésien au monde moderne, que ce soit pour la finance, l'ingénierie ou les sciences. De nos jours, les grandes universités se sont même mises à proposer des messes hebdomadaires pour inviter les croyants et les apprentis à lire et relire les préceptes de Bayes. À tel point que, désormais, il n'est plus déraisonnable de se déclarer bayésien, y compris dans la sphère académique — même si on me regarde encore régulièrement de travers quand je prends fait et cause pour le bayésianisme.

Si je prends plaisir à évoquer cette métaphore, c'est parce que l'histoire du bayésianisme est passionnante en elle-même. Elle est aussi caractéristique de l'histoire des sciences. Contrairement à la manière dont elle est parfois racontée, la science n'est pas qu'une succession d'éclairs de génie et un triomphe de la rationalité. L'abus de pouvoir, la jalousie et les rivalités ont joué un rôle tout aussi important dans l'évolution des idées. Et ce qui a pu être rejeté par les

plus grands savants pendant des siècles peut finir par être accepté par toute la communauté scientifique.

Selon l'écrivain Sharon McGrady, c'est précisément le cas du bayésianisme. McGrady a même fini par dédier tout un livre à l'histoire rocambolesque du bayésianisme, et n'a pas hésité à l'intituler *The Theory that would not Die: How Bayes' Rule Cracked the Enigma Code, Hunted down Russian Submarines, and Emerged Triumphant from two Centuries of Controversy*.

Dans ce chapitre, je vous propose de brièvement explorer les fascinants méandres de l'histoire du bayésianisme. Pour ce faire, il est utile de faire un détour par le XVII^e siècle, au moment où Blaise Pascal et Pierre de Fermat cherchèrent enfin à mathématiser la notion de probabilité.

Les origines de la théorie des probabilités

Pascal et Fermat se demandèrent notamment comment répartir les gains lors d'un jeu de pur hasard inachevé, en fonction des points marqués jusque-là. Imaginez typiquement que deux joueurs misent 10 euros dans un jeu de 11 lancers de pile ou face avec une pièce non biaisée. Le joueur qui gagne le plus de lancers, c'est-à-dire celui qui en gagne 6 ou plus, remporte alors les 20 euros en jeu. Supposez que le score soit de 4-0 au moment où quelque chose vient interrompre la partie. Quelle est alors la façon juste de répartir les gains ?

Intuitivement, celui qui mène 4-0 devrait récupérer plus d'argent du pot commun, puisque sa probabilité de victoire finale est plus grande. Mais quelle fraction du pot mérite-t-il ? Pour déterminer une réponse rigoureuse, Pascal et Fermat durent établir comment propager l'incertitude de chaque lancer. Autrement dit, ils connaissaient la cause de l'incertitude du jeu (à savoir l'incertitude de chaque lancer), et devaient en déterminer les conséquences sur les probabilités de victoires finales des deux joueurs. Ils devaient construire une logique deductive des probabilités. Ceci les conduisit à jeter les bases de la théorie des probabilités, et à introduire des notions comme l'espérance et la loi binomiale.

Mais la théorie de Pascal et Fermat était encore très incomplète. L'homme qui donna vraiment corps et âme à la théorie des probabilités est sans doute plutôt Abraham de Moivre. Les persécutions religieuses l'amenaient à fuir la France à la fin du XVII^e siècle, de Moivre trouva refuge dans le foisonnant environnement intellectuel qu'était la *Royal Society*, où il put côtoyer, entre autres, Isaac Newton, John Wallis et John Locke. Là-bas, il publia un ouvrage séminal intitulé *The Doctrine of Chances*. Ce livre esquissa notamment les premières lignes de l'un des plus beaux théorèmes des mathématiques, le théorème central-limite. Ce théorème déduit la loi de probabilité que suit une variable aléatoire qui est obtenue en ajoutant une infinité de petites perturbations aléatoires et indépendantes.

Le mystérieux Thomas Bayes

Cependant, il y avait un problème de la théorie des probabilités que de Moivre ne sut pas résoudre, et qui faisait écho aux réflexions philosophiques de David Hume, un Écossais postérieur à de Moivre dont on a déjà parlé dans le chapitre 4. Cet autre problème fondamental fut alors intitulé problème de la probabilité inverse. Mais il n'est autre que le problème de l'induction. Il s'agissait d'établir la probabilité des causes sachant les conséquences.

C'est là qu'intervint le pasteur Thomas Bayes de l'Église presbytérienne. Tel un bon mathématicien confronté à une question difficile, Bayes s'intéressa à un exemple simple. Il imagina ainsi une table sur laquelle une boule blanche est posée (uniformément) aléatoirement. Bayes, le dos tourné à la table, n'a aucune idée de l'endroit où se trouve la boule blanche. Bayes va devoir déterminer la position de la boule blanche, ou du moins les positions probables de la boule blanche, à partir de conséquences de cette position.

Plus spécifiquement, l'assistant de Bayes pose ensuite une boule noire sur la table, toujours (uniformément) aléatoirement. Bayes a toujours le dos tourné et ne sait pas non plus où se trouve la boule noire. Il demande alors à son assistant si la boule blanche se trouve à gauche ou à droite de la boule noire. Son assistant lui répond. L'assistant répète ensuite le même procédé avec une seconde boule noire, annonçant à Bayes si cette seconde boule noire a fini à gauche ou à droite de la boule blanche. Et ainsi de suite avec d'autres boules noires.

S'il connaissait la position de la boule blanche, Bayes aurait pu calculer les probabilités des réponses de son assistant. La position de la boule blanche aurait ainsi été (en partie) la cause des réponses de l'assistant aux questions de Bayes. Le problème de la probabilité inverse consiste alors à déterminer la cause, à savoir les positions probables de la boule blanche, sachant les conséquences, c'est-à-dire les réponses de l'assistant. Vous l'aurez compris. C'est en intuitant la formule qui porte son nom que Thomas Bayes résolut le problème de la localisation (probable) de la boule blanche.

On pourrait croire que ceci mit fin au problème de la probabilité inverse. Il n'en est rien. Bayes, à l'instar de nombreux statisticiens dont on parlera dans ce chapitre, eut un comportement énigmatique. Il ne publia pas sa formule magique. Avait-il peur d'une controverse ? Ceci semble peu probable. De son vivant, il avait pris position contre les critiques de Georges Berkeley envers les mathématiques de Newton¹. Avait-il peur de remettre en doute ses croyances religieuses ? Certainement pas, puisque sa théorie de la probabilité inverse avait notamment pour but d'insister sur la notion de cause, qui en remontant jusqu'aux causes premières, devait finir par prouver l'existence de Dieu.

L'une des hypothèses les plus crédibles pour expliquer la non-publication de Bayes est tout simplement que Bayes ne voyait pas toute la beauté de sa for-

¹  *La quête mathématique de l'infiniment petit* | Science4All | L.N. Hoang (2016)

mule, voire qu'il n'y croyait pas lui-même. Quoi qu'il en soit, plusieurs experts s'accordent à affirmer avec grande crédence que Bayes n'était pas bayésien.

Du reste, sa formule ne fut publiée qu'à titre posthume, deux ans après sa mort, en 1763, grâce aux travaux monumentaux de Richard Price. En fait, des deux savants, Price semblait le plus bayésien. Mais lui non plus n'était pas si bayésien. D'ailleurs, ses efforts consentis à publier le travail de Bayes semblent motivés par une volonté de prouver l'existence de Dieu. Ainsi Price affirmait-il : « L'objectif que j'ai, c'est d'établir quelles raisons nous avons pour croire qu'il y a dans la constitution des choses des lois fixes selon lesquelles les choses se produisent, et que, par conséquent, le cadre du monde doit être l'effet de la sagesse et puissance d'une cause intelligente ; et donc [le but est] de confirmer l'argument des causes finales pour l'existence d'une déité. »

Laplace, le père du bayésianisme

En fait, le premier dont on peut dire qu'il était bayésien n'est pas Anglais. Il s'agit du Français Pierre-Simon Laplace. Laplace est l'un des plus grands mathématiciens de l'histoire, et peut-être le plus grand de mes héros. Pendant longtemps, il fut davantage reconnu pour ses travaux en analyse et leurs applications à l'astronomie qu'il publia dans son ouvrage en cinq volumes intitulé *Traité de mécanique céleste*. Cet ouvrage offrit notamment de nouvelles réponses à la question de la stabilité du système solaire. Newton avait déjà prouvé que si la Terre et le Soleil étaient seuls dans l'univers, alors ils formeraient un système stable jusqu'à la nuit des temps. Cependant, lorsqu'il fallut inclure Jupiter dans le modèle, les équations devinrent insolubles. Newton finit par jeter ses bras en l'air et par conclure que seule une intervention divine pouvait mettre de l'ordre dans ce système complexe et stabiliser ainsi les trajectoires des planètes.

Armé de nouveaux outils d'analyse mathématique, dont la transformée qui porte son nom, Laplace réussit à fournir de très bonnes raisons de croire que le système solaire est en fait stable sans intervention divine. Après avoir lu le *Traité* de Laplace, le général Napoléon Bonaparte aurait alors demandé : « Newton a parlé de Dieu dans son livre. J'ai déjà parcouru le vôtre et je n'y ai pas trouvé ce nom une seule fois. » Ce à quoi Laplace aurait répondu : « Je n'ai pas eu besoin de cette hypothèse. »

Cependant, Laplace n'avait pas tout à fait résolu rigoureusement la question de la stabilité du système solaire. Et on ne peut absolument pas lui en vouloir. Des générations de mathématiciens se sont ensuite cassées les dents sur ce problème incroyablement difficile, dont Carl Friedrich Gauss, Henri Poincaré, Andrey Kolmogorov, Jacques Laskar ou encore Cédric Villani. À l'instar de Poincaré qui a trouvé une erreur dans son mémoire qui devait prouver la stabilité du système solaire, la communauté astrophysique et mathématique a maintes fois alterné ses crédences en cette stabilité. De nos jours, les simulations de Jacques Laskar

semblent avoir remporté les faveurs de la communauté scientifique. Ces simulations prédisent l'instabilité du système solaire sur le très, très long terme. Rassurez-vous, on a le temps de voir venir.

L'une des difficultés auxquelles Laplace fut confronté était l'imprécision des données observationnelles dont il disposait. Il faut dire que ces données avaient été collectées par des Arabes vers l'an 1000, des Romains vers l'an 100, des Grecs en 200 avant Jésus-Christ, voire même des Chinois 1100 ans avant Jésus-Christ. Or les instruments de mesure de l'époque étaient malencontreusement imprécis. Laplace avait des données fausses. Laplace pouvait-il tout de même exploiter ces données malgré leurs erreurs ?

Laplace attaqua le problème sous un angle typiquement bayésien. Il connaissait les mesures prises par les astronomes des siècles précédents, et devait en déterminer les causes — à savoir les vraies positions des astres dans le ciel. Conscient de la structure de ce problème, et n'ayant, semble-t-il, pas eu vent de la découverte de Bayes, Laplace se confronta au problème de la probabilité inverse. En 1774, Laplace publia le *Mémoire sur la probabilité des causes par les événements*. Quel document fabuleux ! Il y combina les travaux préalables de Moivre, les outils analytiques de Lagrange et son propre génie, pour établir la formule de Bayes dans toute sa généralité et sa splendeur.

Laplace ne s'intéressa d'ailleurs pas qu'à l'astronomie. Plus tard dans sa vie, il publia ses réflexions dans deux livres, où il sortit les mathématiques de leur champ d'application habituel. Laplace proposa notamment d'appliquer sa théorie des probabilités aux sciences naturelles comme l'astronomie, mais aussi aux sciences sociales, aux témoignages, aux tests médicaux, aux tribunaux judiciaires, au recensement des populations, et à bien d'autres problèmes encore. Laplace exploita d'ailleurs lui-même sa nouvelle philosophie pour l'étude des sexes des nouveau-nés, ce qui l'amena à conclure avec une très grande crédence qu'un nouveau-né est plus probablement un garçon qu'une fille.

Pour Laplace, les raisonnements probabilistes n'étaient qu'une simple mathématisation du bon sens. Laplace voyait certainement la formule de Bayes comme étant la *bonne* façon de réfléchir. Cependant, il était aussi conscient des erreurs récurrentes dans l'application de cette formule par ses contemporains. Le « bon sens » de ses contemporains était miné de sophismes. Une partie de ses livres peut ainsi être interprétée comme des germes de sciences cognitives.

Vers la fin de sa vie, Laplace développa aussi des méthodes statistiques non-bayésiennes, qui s'appuient notamment sur le théorème central-limite qu'il avait prouvé. Laplace comprit alors que cette approche fréquentiste était équivalente à l'approche bayésienne pour des ensembles de données suffisamment grands. Étant plus pratique pour traiter les grandes quantités de données, Laplace finit par préférer cette approche pour de nombreux cas pratiques. Laplace était un bayésien pragmatique.

La loi de succession de Laplace

Mais revenons-en à l'un des calculs les plus intrigants du *Mémoire* de 1774 de Laplace. Pour illustrer sa théorie de la probabilité inverse, Laplace introduisit l'exemple d'une urne possédant un grand nombre de billets blancs et noirs. Cet exemple est en fait très similaire au problème que s'était posé Bayes — un matheux dirait que les deux problèmes sont *isomorphes*. La proportion de billets blancs dans l'urne est supposée inconnue. Laplace tire au hasard un billet de l'urne. Le billet est blanc. Que penser de la proportion de billets blancs dans l'urne ? Comment expliquer que le billet tiré soit blanc ? Quelle est la cause du fait qu'il a tiré un billet blanc ?

Fisher, ce *fréquentiste* acharné, aurait sans doute jeté ses bras en l'air et répondu que cette question n'avait aucun sens. Pour Fisher, c'était là une question non-statistique, voire non-scientifique. Il s'agissait là d'une question *vide de sens*.

Pas pour Laplace. Laplace eut la brillante idée de commencer par *préjuger* la proportion de billets blancs avant de tirer le billet. Il supposa *a priori* que la proportion de billets blancs était un nombre (uniformément) aléatoire entre 0 et 1. Notez bien que l'aléatoire de Laplace n'est pas un incertain réel. Il s'agit d'une représentation de son ignorance (subjective).

Quoi qu'il en soit, Laplace fit ensuite une inférence bayésienne, pour mettre à jour son préjugé au vu de la couleur du billet tiré. Après avoir appliqué la formule de Bayes (qui était en fait sa propre formule !), Laplace conclut alors que, même *a posteriori*, la proportion de billets blancs restait un nombre aléatoire entre 0 et 1. Mais s'il devait maintenant prédire la couleur d'un second billet tiré de l'urne, il attribuerait une probabilité de $2/3$ à la couleur blanche.

Plus généralement, s'il avait tiré p billets blancs et q billets noirs, Laplace aurait attribué une probabilité $(p+1)/(p+q+2)$ au fait qu'un nouveau billet tiré de l'urne soit blanc. Telle est la loi dite de *succession de Laplace*, laquelle a été déduite de la formule de Bayes.

Malheureusement les calculs bayésiens de Laplace requièrent des outils d'analyse sur lesquels je ne peux pas m'étendre dans ce livre. Mais je suggère fortement aux plus motivés d'aller jeter un œil au Problème I à la page 30 de son Mémoire de 1774 en accès libre sur Internet !

En particulier, le génie de Laplace fut de combiner deux sortes d'aléatoire : l'aléatoire du tirage de billet, et celui qui modélise l'ignorance de Laplace sur la proportion de billets blancs. Si les contemporains de Laplace avaient pris le temps de comprendre le génie de la solution de Laplace à ce problème, l'histoire des sciences, et celle de la philosophie des sciences, auraient peut-être pris un autre tournant.

Par exemple, la loi de succession de Laplace lui permit d'enfin fournir une réponse à la question de Hume. Sachant que cela fait désormais j jours de suite que le soleil se lève, peut-on croire que le soleil se lèvera encore demain ?

Si chaque jour était un billet, si le billet était noir lorsque le soleil se lève, et blanc sinon, alors on aurait $p = 0$ et $q = j$. Par conséquent, en appliquant la théorie bayésienne de Laplace, on en vient à prédire que la probabilité que le soleil ne se lève pas demain est égale à $1/(j + 2)$.

Laplace s'appuyant sur la Bible, il choisit de poser j égal au nombre de jours correspondant à 5 000 ans, ce qui lui donnait une probabilité d'environ un sur un million que le soleil ne se lève pas demain. Confronté à l'absurdité de son résultat, Laplace s'empessa d'ajouter que « ce nombre [un million] est beaucoup plus grand pour qui, considérant les principes qui règlent les jours et les saisons dans la totalité des événements, réalise que nul dans l'instant actuel peut arrêter son cours ». Un *bayésien* doit prendre en compte l'ensemble de ses connaissances pour affiner ses prédictions.

Malgré cette remarque, la prédiction de Laplace attisa malencontreusement les foudres des critiques. Elle fut moquée, encore et encore, ce qui conduisit beaucoup à dénigrer toute la théorie des probabilités de Laplace. La prédiction malheureuse de Laplace est sans doute une cause majeure du déclin de la philosophie bayésienne pendant les deux siècles qui suivirent. Pourtant, étrangement, la formule de Laplace est en fait, au vu des connaissances actuelles, étonnamment juste !

Pour commencer, il nous faut corriger la valeur que Laplace a attribué à j . On sait aujourd'hui que cela fait environ 5 milliards d'années que le soleil se lève tous les jours. Du coup, la formule de Laplace nous dit que la probabilité qu'il ne se lèvera pas demain est d'un sur deux mille milliards environ. En particulier, on en vient à prédire que le soleil finira par ne pas se lever d'ici quelques milliards d'années. Or, curieusement, les astrophysiciens nous disent aujourd'hui que dans 5 milliards d'années, le soleil deviendra une géante rouge qui grossira tellement qu'elle englobera la Terre. Et quand bien même cette transformation du soleil en géante rouge n'en serait pas la cause, les simulations de Laskar suggèrent que notre planète bleue finira par quitter son orbite dans quelques milliards d'années. Incroyable ! La physique moderne nous donne donc deux raisons différentes de croire que Laplace avait vu juste !

On pourrait croire qu'il s'agit là d'une coïncidence cosmique inexplicable. À n'en pas douter, il s'agit d'un coup de chance, puisque la prédiction de Laplace est fondamentalement probabiliste — tout aurait pu arriver ! D'ailleurs, ce même raisonnement appliqué à la disparition à venir de l'univers semble échouer². Qui plus est, l'interprétation que je présente là n'est pas entièrement bayésienne³.

²Quoiqu'un article de Caldwell, Kamionkowski et Weinberg (2003) intitulé *Phantom Energy and Cosmic Doomsday* prédit justement un *Big Rip* de notre univers dans 22 milliards d'années, ce qui colle à nouveau avec une prédiction à la Laplace !

³Il s'agit là d'un *posterior-mean*, i.e. toutes les prédictions futures se fondent sur la probabilité *a posteriori* moyenne. Par opposition, la *pure bayésienne* intégrerait tous les *a posteriori* crédibles de la probabilité que le soleil disparaîsse demain, ce qui l'amènerait à conclure que l'espérance de vie du soleil est infinie. Cependant, l'espérance résume mal la crédence *a posteriori* de la *pure bayésienne*. Ainsi, la médiane de l'*a posteriori* purement bayésien correspond bel et bien aux mêmes ordres de grandeurs que l'espérance du *posterior-mean*.

Cependant, il s'agit bien moins d'une coïncidence cosmique que ce que l'on pourrait croire naïvement.

Imaginons que nous cherchions à déterminer la durée de vie d'un être humain en fonction de son âge. La méthode de Laplace⁴ revient à prédire qu'il vivra environ encore autant que son âge. Bien entendu, cette prédition sera erronée si l'humain en question est un nouveau-né ou une personne très âgée. Cependant, ceci est peu probable. La plupart du temps, on tombera sur un humain qui a entre 20 et 60 ans, ce qui conduira à une prédition de la durée de vie des hommes entre 40 et 120 ans.

Mieux encore, supposons que la durée de vie d'un homme est égale à 100 ans, et supposons que toutes les tranches d'âge sont aussi présentes les unes que les autres dans la population. Alors, un calcul probabiliste montre que la prédition moyenne de l'espérance de vie des humains sera exactement 100 ans⁵ !

Ce phénomène mystérieux a été appelé *effet de Lindy* par l'écrivain Albert Goldman, puis par le mathématicien Benoît Mandelbrot et le statisticien Nassim Taleb, en référence à un Deli appelé Lindy's où des comédiens se demandaient comment survivre longtemps dans le *show-business*. Goldman remarqua que le nombre d'apparitions d'un comédien à venir était proportionnel au nombre de ses apparitions passées. Mandelbrot rajouta : « Peu importe la quantité des travaux produits par une personne, celle-ci augmentera en moyenne d'une même quantité additionnelle. » Taleb justifiera ensuite cette observation empirique par l'omniprésence des lois dites *de puissance*, à l'instar de l'étonnante *loi de Zipf* qui prédit que la n -ième lettre la plus fréquente est environ n fois moins fréquente que la plus fréquente⁶.

Aussi étonnant que cela puisse paraître, les conséquences de la loi de succession de Laplace ont de nombreuses applications pratiques, et ont permis, entre autres, de déterminer le nombre de tanks nazis pendant la seconde guerre mondiale à partir des numéros de séries des tanks capturés⁷.

Formellement, en notant $VIE(p)$ le temps de vie du soleil en supposant que sa probabilité de disparaître un jour donné est p , on a calculé $\mathbb{E}_{vie}[\mathbb{VIE}(\mathbb{E}_p[p|D])]$. L'espérance vraiment bayésienne est $\mathbb{E}_{vie,p}E[\mathbb{VIE}(p)|D]$, tandis que la médiane bayésienne est la valeur de x telle que $\mathbb{P}[\mathbb{VIE}(p) \leq x|D] \leq 1/2$.

⁴En fait, Laplace s'est contenté de donner la probabilité que le soleil ne se lève pas demain, et ne s'est pas intéressé à l'espérance de vie du soleil.

⁵En effet, posons $X = 100$, et soit x l'âge de l'humain pris au hasard. La prédition de l'espérance de vie est alors $2x$. La prédition moyenne s'obtient ensuite en intégrant sur toutes les valeurs de x . On obtient $\mathbb{E}[2x] = \int_0^X 2x \frac{dx}{X} = \frac{1}{X} [x^2]_0^X = X$.

⁶  *The Zipf Mystery* | VSauce | M. Stevens (2015)

⁷La loi de succession de Laplace apparaît également dans le paradoxe de l'apocalypse (*doomsday argument*), dont Mr Phi et moi avons parlé dans cet épisode d'Axiome :

 *Jouvence conflictuelle* | Axiome 5 | T. Giraud et L.N. Hoang (2017)

Le grand hiver du bayésianisme

Malheureusement, les sciences n'en étaient pas là. Plutôt que d'y voir une illustration de la stupéfiante efficacité du raisonnement bayésien, les grands savants du XIX^e siècle rejetèrent presque unanimement les probabilités inverses de Laplace. Le mathématicien George Chrystal affirma : « [ces probabilités] étant mortes, on doit les enterrer hors de vue de manière décente, et non pas les présenter dans des livres de cours et dans des exercices d'examens... On se doit d'autoriser l'oubli silencieux des indiscretions des grands hommes. »

D'autres eurent des réactions plus virulentes envers l'approche de Laplace, et la présence des crédences subjectives dans sa théorie. Le philosophe John Stuart Mill critiqua Laplace en qualifiant sa philosophie « d'aberration de l'esprit », voire « d'ignorance [...] prétendant être science ».

En dehors peut-être d'une utilisation par Joseph Bertrand pour prendre des décisions malgré les incertitudes en temps de guerre, et celle d'Henri Poincaré pour rejeter la pertinence des éléments à conviction dans l'affaire Dreyfus, les crédences de Laplace et la formule de Bayes semblèrent avoir disparu du champ des sciences.

Ce fut encore un peu plus le cas au début du XX^e siècle, alors qu'émergèrent les statistiques fréquentistes d'Egon Pearson, Jerzy Neyman et Ronald Fisher. Même si ces génies ne s'entendaient pas entre eux, tous étaient d'accord sur un point : il fallait mettre fin à la subjectivité des théories de Bayes et Laplace. Fisher insulta violemment ces théories avec des noms d'oiseaux, comme « ordures fallacieuses », tandis que Neyman enleva toute notion bayésienne de sa théorie des intervalles de confiance car « toute la théorie serait plus jolie si elle était bâtie dès le début sans aucune référence au bayésianisme et aux *a priori* ». Dès lors, et pendant presque tout le 20ème siècle, les mots « subjectif », « *a priori* » et « bayésien » furent bannis des départements de statistiques.

Le bayésianisme n'était pas tout à fait mort pour autant. Quelques irréductibles comme Émile Borel, Frank Ramsey et Bruno de Finetti se dirent que les probabilités subjectives étaient des outils mathématiques indispensables pour comprendre les paris d'argent. Cependant, ils furent relativement ignorés de leur vivant.

Le grand rival bayésien de Fisher fut plutôt le géologue Harold Jeffreys. Alors que Fisher appliquait à merveille ses théories fréquentistes à ses expériences de génétique, Jeffreys y voyait de sévères limitations quand il s'agissait d'appliquer le fréquentisme à la sismologie. Il est en effet bien difficile de répéter les tremblements de terre pour étudier les propagations des ondes sismiques... Les mesures sismographiques étaient rares et imprécises. Néanmoins, armé de méthodes bayésiennes, Jeffreys sut interpréter ses données pour localiser les épicentres des séismes, et même pour deviner, à juste titre, que le centre de la Terre était liquide. Néanmoins, la véhémence de Fisher, qui rejeta la scientificité des méthodes bayésiennes, prit le pas sur la placidité de Jeffreys.

Bayes au secours des alliés

Alors qu'éclata la seconde guerre mondiale, dans le monde académique, les statistiques étaient anti-bayésiennes. Mais en dehors du monde académique, les statistiques étaient surtout dénigrées. Quand il se rendit compte du rôle clé que pouvait jouer le déchiffrement des codes nazis, le gouvernement anglais se mit à embaucher en priorité des littéraires, des artistes et des historiens. Heureusement, les mathématiciens anglais l'avaient anticipé. Ils s'étaient déclarés physiciens pour que le gouvernement pense à eux. Les statisticiens, en revanche, furent ignorés. Et ce fut sans doute une bonne chose, car la formule de Bayes, alors rejetée par les « vrais » statisticiens, allait être un élément clé.

La cryptographie de la seconde guerre mondiale était d'un nouveau type. Elle était mécanisée. Les nazis utilisaient notamment une machine de codage appelée *Enigma*. *Enigma* était une sorte de machine à écrire, dont la spécificité était d'écrire le code chiffré de ce qui était tapé. Mieux encore, pour déchiffrer un code, il suffisait de taper le code chiffré⁸.

Enfin, pas tout à fait. La manière dont la machine chiffrait et déchiffrait les codes dépendait de la configuration de la machine. Chaque jour, les nazis utilisaient une configuration différente de la machine. Or, *Enigma* était vendue avec des millions de telles configurations. Pire encore, les militaires nazis avaient à leur disposition une fonctionnalité additionnelle qui démultiplia le nombre de configurations d'*Enigma*. Il y en avait désormais des centaines de trillions. Il était illusoire d'espérer toutes les tester.

Petit à petit, sous l'impulsion notamment de Winston Churchill, les autorités anglaises compriront que les mathématiques allaient être la clé pour déchiffrer les codes ennemis. Une *dream team* fut ainsi rassemblée à Bletchley Park. Cette équipe réunissait notamment Peter Twinn, Gordon Welchmann, Derek Taut, Bill Tute, Max Newman, Jack Good, mais aussi et surtout, le très grand Alan Turing.

Fort de ses découvertes sur la théorie du calcul de 1936 dont on reparlera dans le prochain chapitre, Turing comprit rapidement comment automatiser bon nombre d'étapes pour casser *Enigma*, comme le met en scène le film *Imitation Game*. Ceci lui permit de construire la machine *Bombe*, qui, tous les jours, déchiffrera les codes des armées nazies de terre et de l'air. Cependant, la *Kriegsmarine* utilisait une version encore plus sophistiquée d'*Enigma*. *Bombe* n'était pas suffisamment rapide pour la *Kriegsmarine*. Pire encore, les autorités nazies utilisaient un code encore plus complexe qui ne reposait pas sur *Enigma*, mais sur la machine de Lorenz.

Le premier défi de Turing fut alors de convaincre les autorités anglaises que l'*Enigma* de la *Kriegsmarine* et la machine de Lorenz pourraient être craquées. Il fallait les convaincre qu'investir dans le décodage de ces codes ne serait pas

⁸  Le décryptage d'*Enigma* | Science4All | R. Barbulescu et L.N. Hoang (2017)

vain. Pendant longtemps, les autorités ne furent pas convaincues. Ces codes semblaient trop compliqués, et leur déchiffrement trop coûteux en temps, en ressources humaines et en matériel. Cependant, argua Turing, le jeu en valait la chandelle.

Churchill finit par être convaincu. Il affirmera plus tard que « la seule chose qui [l']effrayait vraiment pendant la guerre était le danger des sous-marins [de la *Kriegsmarine*] ». Ces sous-marins avaient déjà coulé un grand nombre de navires de ravitaillement venant d'Outre-Atlantique. Le capitaine Jerry Roberts ajoute que si la situation avait perduré, « il est tout à fait possible, même probable, que la Grande-Bretagne aurait été affamée et aurait perdu la guerre. » Quant à craquer la machine de Lorenz, ceci permettrait de connaître directement les intentions et stratégies d'Adolf Hitler — et découvrir notamment qu'il s'attendait à un débarquement à Calais plutôt qu'en Normandie.

Turing eut le feu vert pour lancer ses recherches. Il restait à trouver des bonnes idées. Comme vous vous y attendez sans doute, la solution était la formule de Bayes. Turing détermina notamment une manière heuristique mais quantitative d'appliquer cette formule. L'unité de mesure de Turing était alors le *banburismus* ou *ban*, du nom de la ville qui lui fournissait du matériel pour automatiser autant que possible le calcul des *bans*. Au lendemain de la guerre, le mathématicien Claude Shannon, que Turing rencontra aux États-Unis pendant la guerre, formalisera une variante de ces *bans* et leur donna un nom qui nous est aujourd'hui très familier : le *bit*. On y reviendra.

Pour l'heure, revenons-en à Turing et à la seconde guerre mondiale. Quand une configuration d'Enigma semblait partiellement décoder un message, cette configuration gagnait des *bans*, ou, dit autrement, de la crédence bayésienne. En combinant les *bans* de différentes configurations, Turing put orienter ses recherches pour tester en priorité les configurations les plus prometteuses. Ce procédé, que j'ai grotesquement simplifié, permit de grandement accélérer le déchiffrement. À terme, Turing, ses collègues et ses machines devinrent capables de lire une grande fraction des messages nazis.

L'historien Harry Hinsley affirme que les travaux des mathématiciens anglais « ont raccourci la guerre d'au moins deux ans et probablement de quatre ans ». D'autres suggèrent même que l'issue aurait été incertaine. Ce qui est moins discutable, c'est que les mathématiques de Turing et de ses collègues, et l'utilisation opportune de la formule de Bayes, auront sauvé des millions de vie.

Cependant, au lendemain de la guerre, tout ceci était confidentiel, et Winston Churchill fit tout pour s'assurer que cela le resterait. Il ordonna la destruction de tout document qui puisse suggérer que les codes nazis avaient été déchiffrés, et enterra la formule de Bayes (et les machines de Turing) six pieds sous terre.

Des îlots bayésiens dans un océan fréquentiste

Au lendemain de la guerre, le terme « bayésien » demeurait une insulte. Pendant le Maccarthysme des années 1950, alors que les communistes étaient chassés des institutions américaines telles des sorcières, un statisticien américain dit d'un de ses collègues, en ne plaisantant qu'à moitié, que ce collègue était « anti-américain car [il] était bayésien, ce qui décrédibilisait le gouvernement américain. » Un autre statisticien rajouta : « les statisticiens bayésiens ne se restreignent pas suffisamment à Bayes même. Si seulement ils suivaient les pas de Bayes et publiaient uniquement à titre posthume, nous serions à l'abri de beaucoup de problèmes. » Les départements de statistiques des universités, en particulier, étaient profondément anti-bayésiens. Jack Good, qui avait exploité la formule de Bayes aux côtés de Turing pendant la guerre, avait bien essayé de louer les mérites de méthodes bayésiennes. Mais ses discours tombèrent encore et encore dans l'oreille d'un sourd.

C'est loin des bancs de classe que la flamme bayésienne fut ravivée, notamment par le charismatique actuaire américain Arthur Bailey. Les estimations des aléas de la vie étaient cruciaux pour déterminer le prix des assurances. Plus un risque est associé à une grande probabilité, plus le coût de la couverture de ce risque sera élevé, et plus le prix doit l'être aussi. Cependant, ces probabilités n'étaient pas fondées sur la *p-value* de Fisher. Elles étaient calculées à l'aide de formules obscures. Rares étaient les actuaires qui en connaissaient l'origine, mais tous constataient que ces formules donnaient des résultats cohérents. Les calculs d'actuariat marchaient bien, mais personne ne savait pourquoi ! Bailey, qui avait été formé à l'école fréquentiste, en fut horrifié.

Cependant, Bailey finit par découvrir que les étranges formules de l'actuariat, à l'instar d'autres formules magiques qu'on découvrira dans de futurs chapitres, étaient mystérieusement similaires à la formule de Bayes. Après un an de scepticisme, Bailey finit alors par embrasser les inférences quasi-bayésiennes sur lesquelles reposaient les prix des assurances. Il en vint même à rejeter son éducation fréquentiste et à faire campagne contre les méthodes de Fisher. En 1950, il publia un article reliant la théorie de la crédibilité sur laquelle reposait l'actuariat, aux travaux de Laplace, Price et Bayes. Il y vanta le mérite des probabilités subjectives et annonça la fin de la tyrannie fréquentiste. Malheureusement, Bailey mourut d'une crise cardiaque peu de temps après avoir pris position contre Fisher.

Il y a bien eu deux théoriciens et demi aux réflexions bayésiennes. Commençons par le demi-bayésien. Avant la guerre, en 1933, Andrey Kolmogorov proposa enfin des axiomes sur lesquels pouvait reposer la théorie des probabilités. Pour Kolmogorov, ce qui importait n'était pas l'interprétation de ces probabilités, mais les règles de manipulations des probabilités. Quand, dos au mur, on lui demanda d'appliquer sa théorie à la stratégie militaire, Kolmogorov développa le même raisonnement que Bertrand un siècle plus tôt. Ce raisonnement était bayésien. Même si Kolmogorov se disait plutôt fréquentiste.

Après la guerre, la formalisation mathématique de la théorie des probabilités conduisit Dennis Lindley et Leonard Savage à rejeter les statistiques fréquentistes de Fisher. Par opposition, la formule de Bayes était une conséquence directe des axiomes de Kolmogorov, et était donc mathématiquement parfaitement fondée. Pire, en 1958, Lindley publia un article prouvant l'incohérence d'une méthode d'induction de Fisher appelée induction fiduciaire. Lindley avait osé tenir tête à Fisher ; et il avait eu raison. Victorieux, Lindley devint un activiste militant pour le bayésianisme, déclara que toute statistique était un cas particulier ou une approximation de la formule de Bayes et ouvrit jusqu'à 10 départements de statistiques bayésiennes en Angleterre.

En 1954, Savage, quant à lui, publia *Foundations of Statistics*, où il prit fait et cause pour l'interprétation subjective des probabilités. Là où Savage se distingua des autres, c'est dans son acceptation messianique de la formule de Bayes. Savage, plus que tout autre, ne voyait pas cette formule comme un outil de raisonnement parmi d'autres. Pour Savage, c'est le *seul* outil de raisonnement. Un raisonnement juste *est* un calcul de la formule de Bayes. Et tout compromis est irrationnel (mais possiblement justifiable par pragmatisme). Savage était religieusement bayésien.

Quand on lui demanda si cela remettait en cause l'objectivité des sciences, Savage répondit que l'objectivité est l'émergence d'un consensus dans la communauté scientifique, qui apparaît lorsque suffisamment de données s'accumulent. Cependant, Savage rajouta que c'était la seule façon de définir l'objectivité. Pour Savage, les méthodes fréquentistes ne sont pas objectives, puisqu'elles requièrent constamment une interprétation des résultats statistiques, voire un choix de la méthode fréquentiste. Qui plus est, les tentatives d'objectivisation de Fisher, via notamment son induction fiduciaire, revenaient « à une tentative intrépide de faire des omelettes bayésiennes sans casser d'œufs bayésiens. » Hélas, à l'instar de Bailey, Savage mourut d'une crise cardiaque en pleine croisade pro-bayésienne.

Bayes secouru par les praticiens

Loin des inquiétudes des théoriciens, les statistiques bayésiennes connurent aussi un renouveau dans divers domaines dans lesquels les méthodes fréquentistes semblaient insuffisantes. En particulier, Robert Schlaifer et Howard Raiffa s'appuyèrent notamment sur la théorie des jeux de von Neumann et Morgenstern, et combinèrent la théorie de l'utilité aux probabilités subjectives pour développer une théorie de la décision en présence d'incertitude. Ce faisant, Schlaifer et Raiffa transformèrent la *Harvard Business School* en un *hot house* bayésien. Bientôt, suite à la publication de leur livre, toutes les écoles de commerce se mirent aux statistiques bayésiennes et de nombreux prix Nobel d'économie finirent par être décernés à des chercheurs bayésiens comme John Harsanyi ou Roger Myerson, dont on reparlera dans un prochain chapitre.

Le génie des statistiques bayésiennes fut de pouvoir s'adresser à des cas où les données se font rares. En 1950, un économiste demanda au statisticien David Blackwell comment déterminer la probabilité d'une autre guerre mondiale dans les 5 années à venir. Blackwell, en bon élève fréquentiste, répondit alors : « Oh ! Cette question n'a aucun sens. Les probabilités s'appliquent à de longues séquences d'événements répétables. Or il s'agit clairement là d'un cas unique. La probabilité est soit 0, soit 1. Mais nous ne saurons pas avant 5 ans. » L'économiste rétorqua : « j'avais peur que vous disiez cela. J'ai parlé à plusieurs autres statisticiens, et ils m'ont tous dit la même chose. » Plus tard, quand il comprit la faiblesse du pouvoir prédictif des statistiques fréquentistes, Blackwell se convertit au bayésianisme.

Une autre application majeure des statistiques bayésiennes fut l'étude des effets nocifs du tabac sur le cancer du poumon. Le héros de cette étude épidémiologique est Jerome Cornfield. Cornfield fut confronté aux vives critiques des anti-bayésiens Neyman et Fisher. Fisher, en particulier, contra les arguments de Cornfield en leur reprochant l'absence d'expériences contrôlées et répétées comme le requéraient ses méthodes fréquentistes. Fisher, qui recevait notamment des financements de l'industrie du tabac et niait les effets nocifs du tabac, en vint même à proposer l'hypothèse selon laquelle le cancer du poumon révélait une prédisposition à fumer le tabac ! Là encore, à l'instar de Lindley, avec le temps, Cornfield finit par avoir le dernier mot. La communauté scientifique est désormais unanime. Le tabac est un facteur de risque majeur du cancer du poumon.

De son côté, John Tukey appliqua les statistiques bayésiennes à la prédiction des résultats des élections présidentielles. En 1960, la course entre Nixon et Kennedy était remarquablement serrée et incertaine, si bien qu'aucune chaîne de télévision n'osait annoncer le résultat final. À deux heures du matin, Tukey donna enfin le feu vert à la chaîne de télévision américaine NBC pour annoncer la victoire de Kennedy. Mais ce ne sera finalement qu'à huit heures du matin, que la chaîne eut le courage de le faire. Les méthodes de Tukey restèrent longtemps secrètes. En particulier, étant professeur de statistiques, Tukey refusa d'en admettre l'aspect bayésien.

Plus récemment, les méthodes bayésiennes ont retrouvé le vent en poupe, notamment en 2008, lorsque Nate Silver fut le premier de l'histoire à prédire correctement les résultats des 50 États américains. La prédiction de Silver de 2016 aura été moins brillante, mais on y reviendra.

De la même façon, nombreux sont ceux qui, se frottant à l'incertain des événements rares, se tournèrent irrémédiablement vers la formule de Bayes pour trouver des solutions pratiques à leurs problèmes. Ainsi, Norman Rasmussen s'arma de crédences bayésiennes pour estimer la probabilité d'incidents majeurs dans les centrales nucléaires, tandis que la NASA embaucha une agence dont les outils bayésiens prédiront une probabilité d'incidents majeurs d'un sur 35 dans le lancement des fusées. Voilà qui est beaucoup plus significatif — et réaliste ! — que la probabilité d'un sur 100 000 qui fut prédite par la NASA.

Cependant, jusqu'aux années 1990, ces succès bayésiens étaient rares et disparates. Et il y avait une bonne raison à cela. Les calculs bayésiens sont longs, difficiles et rapidement hors de portée des équations mathématiques. Ils nécessitaient souvent des calculs d'intégrales sans forme close. Le bayésianisme semblait prometteur. Mais il n'était pas toujours pratique. L'avènement d'une théorie plus générale et plus applicable du calcul allait changer la donne.

Le triomphe de Bayes, enfin !

Dans les années 1960, Ray Solomonoff, dont on reparlera longuement dans le prochain chapitre, combina la théorie du calcul de Turing et la formule de Bayes pour anticiper une formulation générale de l'intelligence artificielle. À l'instar d'autres avant lui, Solomonoff fut très virulent envers le fréquentisme et son gourou : « La subjectivité en science est souvent considérée comme le Mal [...] si elle survient, alors les résultats ne sont pas du tout des "sciences". Le grand statisticien, R.A. Fisher, était de cet avis. Il voulait faire des statistiques "une vraie science" exempte de la subjectivité qui fut si présente dans son passé. Je pense que Fisher s'était sérieusement trompé à ce sujet, et que son travail dans ce domaine a profondément endommagé la compréhension des statistiques par la communauté scientifique — un dommage dont elle se soigne beaucoup trop lentement. » Malheureusement, les idées de Solomonoff sont longtemps restées purement théoriques, celui-ci ne disposant pas des machines nécessaires pour les expérimenter.

Quand les machines à calculer virent le jour, cependant, le bayésianisme put enfin vivre sa renaissance messianique. Frederick Mosteller fut ainsi l'un des premiers à exploiter ces nouveaux outils pour résoudre des problèmes bayésiens difficiles. Puis, à partir des années 1980 notamment, l'émergence de techniques dites de *Monte-Carlo*, et surtout de *Markov-Chain-Monte-Carlo* (MCMC), révolutionna l'utilisation pratique de la formule de Bayes. Au lieu de calculer exactement les intégrales qui échappaient aux équations mathématiques, les méthodes de Monte Carlo effectuent des échantillonnages qui permettent le calcul d'approximations des intégrales. En particulier, le programme *Bayesian inference Using Gibbs Sampling* (BUGs) annonça le triomphe final du bayésianisme, tandis que, plus récemment, le *deep learning* et d'autres méthodes de *machine learning* tirent profit des *a priori* bayésiens pour sans doute annoncer les changements sociétaux les plus spectaculaires de l'histoire⁹.

Enfin, au cours des dernières décennies, la formule de Bayes et le cadre bayésien semblent révolutionner notre compréhension de l'intelligence, que celle-ci soit artificielle ou humaine. Des informaticiens comme Judea Pearl, Geoffrey Hinton et Michael Jordan, et des neuroscientifiques comme Josh Tenenbaum, Karl Friston et Stanislas Dehaene, ont fait du bayésianisme le pilier incontournable de toute forme de cognition. On y reviendra.

⁹  *Humains versus machines* | IA 1 | Science4All | L.N. Hoang (2017)

Bayes est partout

Pour finir ce chapitre, il est bon de revenir sur l'incroyable étendue du champ d'application des statistiques bayésiennes à travers l'histoire. Pêle-mêle, on peut dresser la liste suivante : diagnostic médical, génétique, épidémiologie, astrophysique, biologie, politique, guerre, cryptographie, géologie, théologie, jeux, assurance, paris, prise de décision, économie, ingénierie aérospatiale, intelligence artificielle, neurosciences... .

Ça, c'est ce dont on a pris le temps de parler dans ce chapitre. Mais les applications transcendent de loin cette liste. On peut ainsi ajouter, toujours pêle-mêle et de façon non exhaustive : le sport, la psychologie, l'archéologie, la paléontologie, l'éducation, les réseaux sociaux, la traduction automatique, le traitement du signal, le séquençage du génome, l'étude des protéines, l'allocation des ressources, la communication, l'analyse d'images, la publicité, la finance, la planification, la logistique et plein d'autres domaines encore... .

Malheureusement, ce chapitre est trop court pour vraiment explorer les méandres de l'histoire mouvementée de la théorie bayésienne. Fort heureusement, il y a d'excellentes ressources disponibles sur Wikipedia ou sur Less Wrong¹⁰ pour aller un peu plus loin. Mais surtout, je vous recommande fortement l'excellent livre de Sharon McGayne dont j'ai parlé en début de chapitre. Ce livre montre que la marche des sciences est tout sauf un long fleuve tranquille. *L'apprentissage est une danse*. Une danse pleine de remous. Mais cette danse semble inéluctablement tendre vers le progrès ; et ce progrès semble être l'acceptation des méthodes bayésiennes.

Références en français

 *Mémoire sur la probabilité des causes par les événements* | Imprimerie Royale | P.S. Laplace (1774)

 *Théorie analytique des probabilités* | V. Courcier | P.S. Laplace (1812)

 *Essai philosophique sur les probabilités* | Bachelier | P.S. Laplace (1840)

 *Peut-on mathématiquement prédire l'avenir du système solaire ?* Espace des sciences | C. Villani (2014)

 *La mort du Soleil* | Sense of Wonder | S. Carassou et E. Ledolley (2015)

 *Le théorème central limite* | La statistique expliquée à mon chat | L. Maugeri, G. Grisi et N. Uyttendaele (2017)

 *Le décryptage d'Enigma* | Science4All | R. Barbulescu et L.N. Hoang (2017)

¹⁰  *A History of Bayes' Theorem* | Less Wrong | lukeprog (2011)

- ▶ *La quête mathématique de l'infiniment petit* | Infini 7 | Science4All | L.N. Hoang (2016)
- ▶ *Humains versus machines* | IA 1 | Science4All | L.N. Hoang (2017)

⌚ *Jouvence conflictuelle* | Axiome 5 | T. Giraud et L.N. Hoang (2017)

Références en anglais

- 📚 *The Doctrine of Chances: or, A Method of Calculating the Probability of Events in Play* | W. Pearson | A. de Moivre (1718)
- 📚 *The Foundations of Statistics* | Wiley Publications in Statistics | L. Savage (1950)
- 📚 *Game Theory and Economic Behavior* | Princeton University Press | J. von Neumann et O. Morgenstern (1944)
- 📚 *Applied Statistical Decision Theory* | MIT Press | H. Raiffa et R. Schlaifer (1944)
- 📚 *The Unfinished Game: Pascal, Fermat, and the Seventeenth-Century Letter that Made the World Modern* | Basic Books; First Trade Paper Edition | K. Devlin (2010)
- 📚 *The Theory that would not Die: How Bayes' Rule Cracked the Enigma Code, Hunted down Russian Submarines, and Emerged Triumphant from two Centuries of Controversy* | Yale University Press | S. McGrawne (2011)

- 📖 *The Influence of Ultra in the Second World War* | Cambridge Security Group Seminar | H. Hinsley (1993)
- 📖 *Credibility Procedures: Laplace's Generalization of Bayes' Rule and the Combination of Collateral Knowledge with Observed Data* | New York State Insurance Department | A. Bailey (1950)
- 📖 *Smoking and Lung Cancer: Recent Evidence and a Discussion of some Questions* | International Journal of Epidemiology | J. Cornfield, W. Haenszel, C. Hammond, A. Lilienfeld, M. Shimkin, and E.L Wynder (1959)
- 📖 *Algorithmic probability: Theory and Applications* | Information Theory and Statistical Learning | Springer | R. Solomonoff (2009)

- ▶ *Mathematicians: Blaise Pascal* | Singingbanana | J. Grime (2009)
- ▶ *The Zipf Mystery* | VSauce | M. Stevens (2015)
- ▶ *Confidence Interval for the Mean* | Wandida | J.Y. Le Boudec (2016)

- 🌐 *A History of Bayes' Theorem* | Less Wrong | lukeprog (2011)

Mon intérêt précoce pour ce domaine est né de ma fascination pour les sciences et les mathématiques. Cependant, au moment d'étudier la géométrie, mon intérêt portait davantage sur la manière dont les preuves étaient découvertes que sur les théorèmes eux-mêmes. De même, en science, mon intérêt concernait davantage la manière dont les choses étaient découvertes que les contenus de ces découvertes. L'œuf doré n'était pas aussi excitant que l'oie qui le pondait.

Ray Solomonoff (1926-2009)



Le démon de Solomonoff

Ni homme ni machine

J'ai eu le bonheur de méditer le *pur bayésianisme* par moi-même. Lentement, mais sûrement, j'ai senti que *la* bonne philosophie du savoir se devait d'astucieusement combiner la formule de Bayes et l'informatique théorique. Mais pendant longtemps, je ne savais pas comment combiner ces deux morceaux du puzzle. Bien que j'eusse déjà entamé l'écriture de ce livre, j'étais alors fort ignorant.

Et puis, j'ai lu Ray Solomonoff.

Ce fut l'un des plus grands moments de ma vie. J'ai été saisi. J'ai eu l'impression que toutes les pièces d'un gigantesque puzzle venaient de parfaitement s'emboîter sous mes yeux ébahis, pour laisser place au Saint-Graal de la philosophie du savoir dont j'avais suivi la trace depuis de longues années. Incroyable ! Pour accéder à la connaissance, il suffit d'effectuer les calculs bayésiens de Solomonoff — et toute alternative a de bonnes chances d'être vouée à l'échec. Ce sont ces calculs que la *pure bayésienne* effectue à longueur de journée.

Malheureusement, le *pur bayésianisme* de Solomonoff requiert des calculs très complexes. Notre *pure bayésienne* n'est pas humaine. Pire, les calculs requis transcendent même les capacités de toute machine à calculer. Ils ne peuvent donc pas être exécutés par une machine non plus. La *pure bayésienne* n'est ni humaine, ni machine ! Si l'on admet la thèse de Church-Turing, elle ne peut être qu'un démon qui échappe aux lois de la physique.

En référence aux travaux génialissimes de Solomonoff qui datent des années 1960, dans ce chapitre, je l'appellerai le *démon de Solomonoff*¹. Mais avant d'en venir au démon de Solomonoff, il nous faut revenir quelques décennies en arrière, et introduire l'un des plus merveilleux concepts de l'Histoire des idées.

Les fondements de l'algorithmique

Au début du XX^e siècle, les mathématiques sont en crise. Bertrand Russell vient de publier le paradoxe qui porte son nom. Ce paradoxe dévastateur² montre l'extrême difficulté de faire reposer les mathématiques sur des fondations solides. En l'absence de telles fondations, les mathématiques ressemblent alors à un château de cartes qui peut s'effondrer à la première brise. David Hilbert en est conscient. Solidifier ce château de cartes doit devenir la priorité des plus grands logiciens et mathématiciens. Frege, Cantor, Peano, Russell, Whitehead, Lebesgue, Zermelo, Fraenkel et Tarski ne sont que quelques-uns des grands génies qui s'affairent à cette tâche ardue. Des décennies de travaux s'accumulent en des ouvrages de plus en plus opaques aux profanes. Mais Hilbert ne désespère pas. « Nous devons savoir, nous allons savoir », déclare-t-il à la radio.

Sauf qu'en 1931, un jeune logicien de 25 ans anéantit les espoirs d'Hilbert. Ce logicien est souvent reconnu comme étant le plus grand logicien de tous les temps. Il s'agit de Kurt Gödel. Gödel montra que tous les efforts des logiciens resteraient à jamais vains : toute fondation mathématique est vouée à n'être qu'un château de cartes. On ne pourra jamais prouver que ces fondations sont inébranlables. Tel est le (second) théorème d'incomplétude de Gödel³.

À défaut de réconforter Hilbert et la communauté mathématique, les travaux de Gödel et la logique formelle construite par les autres logiciens ont toutefois le bon goût de mettre le doigt sur l'importance des règles de manipulations de symboles dont on a parlé au chapitre 3. De façon très formelle, les mathématiques se réduisent dès lors à un langage très précis, avec une syntaxe et une grammaire très rigides. Les phrases de ce langage (si on le suppose cohérent) se divisent dès lors en quatre catégories : démontrable, réfutable, syntaxiquement invalide et indécidable. Qui plus est, déterminer à quelle catégorie une phrase donnée appartient revient à se poser la question de l'existence d'une suite d'opérations de symboles, qui part des symboles composant des phrases admises vraies, appelées axiomes, et aboutit à la phrase donnée (ou à sa négation).

¹J'ai beaucoup hésité sur le genre du démon. Mais il me tenait à cœur de faire un clin d'œil à l'Histoire, et aux démon de Descartes, de Laplace et de Maxwell. Après tout, tous ces démons ont la particularité de flirter avec ou de transcender les limites algorithmiques de l'univers ! J'ai finalement décidé de diffuser la responsabilité du choix de la terminologie, en proposant un vote par scrutin de Condorcet randomisé :

https://twitter.com/science_4_all/status/963327420441952256

votation.ovh/#resultat/1090630306

² La diagonale dévastatrice de Cantor | Infini 16 | Science4All | L.N. Hoang (2016)

³ Les théorèmes d'incomplétude de Gödel | Infini 18 | Science4All | L.N. Hoang (2016)

C'est peut-être cette étude des opérations sur les symboles de la logique formelle qui aura conduit Kurt Gödel, Alonzo Church et Alan Turing à découvrir indépendamment trois définitions distinctes de ce qu'est une suite d'opérations « physiquement valides » sur des symboles. Gödel définit la classe des fonctions généralement récursives, Church introduit le λ -calcul et Turing inventa la machine à calculer qui porte aujourd'hui son nom. De façon stupéfiante, Church et Turing découvrirent que toutes ces définitions étaient en fait équivalentes⁴ !

Cette découverte semblait si profonde que Church et Turing postulèrent la thèse dite *de Church-Turing*. Cette thèse affirme que toute notion de « suite d'opérations physiquement valide », ou de « manipulations de symboles purement mécaniques », ou de « calcul réalisé par une machine » ou « d'algorithme », était en fait équivalente à celles de Gödel, Church et Turing. Comme le suggère Scott Aaronson, « si vous y réfléchissez suffisamment longtemps, vous finirez par conclure que tout calcul peut être réalisé par une machine de Turing ».

Au fondement de l'informatique théorique, on trouve le théorème fondamental d'Alan Turing qui prouve l'existence de machines de Turing dites *universelles*. Ces machines universelles peuvent simuler n'importe quelle autre machine de Turing. Elles peuvent donc calculer tout ce que les fonctions généralement récursives de Gödel et le λ -calcul de Church peuvent calculer. Autrement dit, il existe une machine qui, pourvu qu'on lui fournit le bon code à exécuter, sera capable de faire tous les calculs (physiquement) imaginables.

De prime abord, on pourrait croire que cette notion de calcul n'intéresse que des logiciens, des théoriciens, et peut-être quelques scientifiques réalisant des simulations. Cependant, la thèse de Church-Turing peut être interprétée comme une loi physique de notre univers. En effet, elle postule l'absence de machine à calculer dans notre univers qui puisse résoudre un problème qu'une machine de Turing ne peut pas résoudre. Ce postulat porte sur l'univers tout entier ! Ainsi, si la thèse de Church-Turing est vraie, alors l'univers tout entier ne peut pas faire quelque chose qu'une machine de Turing universelle ne saurait pas calculer. Ou dit autrement, tout dans notre univers peut être simulé par une machine de Turing universelle⁵. En particulier, si la thèse de Church-Turing est vérifiée⁶, alors nos cerveaux aussi ne sont autres que des machines de Turing.

La thèse de Church-Turing aura grandement influencé l'industrie des nouvelles technologies. Puisque, en termes de calculs, on ne saura jamais faire mieux qu'une machine de Turing universelle, autant investir lourdement dans la production de machines de Turing universelles. Ces machines de Turing universelles ont envahi notre quotidien sous diverses appellations. On les appelle aujourd'hui des ordinateurs, des tablettes ou encore des téléphones intelligents⁷.

⁴  *Making a computer Turing complete* | B. Eater (2018)

⁵  *La machine de Turing* | IA 4 | Science4All | L.N. Hoang (2017)

⁶ Même la mécanique quantique est simulable par une machine de Turing (surtout si on adopte les interprétations d'Everett ou de De Broglie-Bohm). Cette simulation par une machine classique peut toutefois être exponentiellement plus longue que par ordinateur quantique.

⁷ Il ne s'agit techniquement pas de machines de Turing, car leur mémoire est finie.

Qu'est-ce qu'un *pattern* ?

Les machines de Turing ont de très nombreuses conséquences, que ce soit en mathématiques, en physique et en technologie. Mais c'est pour une autre raison encore que je vous en parle. Cette raison est philosophique. En effet, en termes épistémologiques, ces machines de Turing semblent être l'objet idéal pour formaliser la notion de *pattern* ou de régularité, dont tant de mathématiciens parlent de façon informelle.

Considérons la suite suivante : 1, 2, 4, 8, 16. Sauriez-vous compléter cette suite ? Il y a de bonnes chances que vous deviniez que le nombre qui vient après 16 est 32. Il est même très probable que vous soyez particulièrement confiant en votre prédiction. Mais pourquoi donc ? Qu'est-ce qui fait que le futur de cette suite soit si prévisible, alors que je ne vous en ai donné qu'un échantillon restreint composé que de 5 données ? Et votre crédence en votre prédiction, est-elle justifiée ?

Il est probable que le raisonnement que vous avez eu en tête soit le suivant : pour passer de 1 à 2, on multiplie par 2. Pour passer de 2 à 4, on multiplie aussi par 2. Idem pour passer de 4 à 8, puis de 8 à 16. La suite en question est donc obtenue en multipliant le dernier chiffre de la suite par 2. Il s'agit là d'un *pattern* si régulier et si simple qu'il semble voué à se prolonger. Ou dit autrement, il existe un algorithme très simple, c'est-à-dire une règle de calcul, qui permet de produire tous les éléments de la suite. Et il semble bien que, en vertu notamment du rasoir d'Ockham sur lequel on reviendra, cette simplicité de l'algorithme soit un argument quasi conclusif.

Et pourtant, il existe une toute autre façon d'expliquer la suite qui nous est donnée. Prenez un disque, et placez deux points sur le bord du disque. Tracez la droite liant ces deux points. Le disque est désormais divisé en 2 parties. Rajoutez un troisième point sur le bord du disque, et tracez les deux droites liant ce point aux deux autres points. On voit alors se dessiner un triangle intérieur et trois morceaux du disque à l'extérieur du triangle. Le disque est maintenant divisé en 4 parties. Et bien, si vous rajoutez un quatrième point et si vous le reliez aux trois autres points, le disque sera alors divisé en 8 parties. L'ajout du cinquième point le divise ensuite en 16 parties !

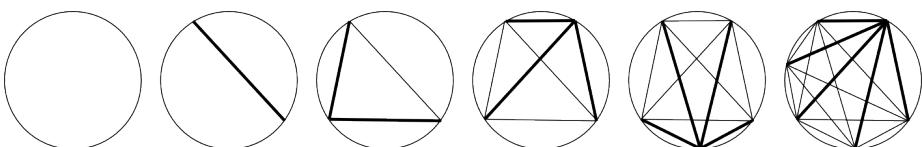


Figure 7.1. L'ajout d'un point et des segments de droite le reliant aux points précédemment ajoutés produit 1, 2, 4, 8, 16 puis 31 morceaux de disque.

La suite 1, 2, 4, 8, 16 correspond donc au nombre de parties d'un disque découpé par l'ajout d'un nouveau point sur son bord et le traçage des droites reliant ce

nouveau point aux autres. On peut donc être tentés de compléter la suite en comptant le nombre de parties du disque suite à l'ajout d'un sixième point. Surprise ! Ce nombre de parties est alors 31, et non pas 32 ! On a donc là une autre façon de compléter la suite⁸.

Voilà qui remet complètement en cause la crédence qu'on devait avoir en notre complétion initiale de la suite. Faut-il compléter la suite par 31 ou par 32 ? Y a-t-il une bonne réponse ? Et quelle devrait être notre crédence en les diverses réponses possibles ? Se pourrait-il que la suite puisse être naturellement complétée par une troisième autre possibilité ?

On sent tout de même que, s'il y a une bonne raison de compléter la suite par 31, et s'il y a sans doute encore d'autres façons de compléter la suite, 32 demeure la complétion la plus crédible de la suite. Y aurait-il une bonne façon de formaliser cela ?

La complexité de Solomonoff

En 1963, le mathématicien Andrey Kolmogorov répondit par l'affirmative. Kolmogorov s'appuya sur les notions de calcul de Gödel, Church et Turing, pour définir une mesure de la complexité des suites numériques comme 1, 2, 4, 8, 16, 32 et 1, 2, 4, 8, 16, 31. Cette complexité, aujourd'hui connue sous le nom de complexité de Kolmogorov, est depuis devenue une notion fondamentale de la théorie du calcul.

Kolmogorov n'est toutefois pas le premier à avoir pensé à cette notion de complexité. Solomonoff l'avait précédé, en publiant un rapport préliminaire de ses recherches dès 1960. Il serait donc plus juste de parler de complexité de Solomonoff ! Cependant, sans doute parce que Solomonoff a eu le malheur de combiner sa mesure de complexité à un bayesianisme considéré hérétique aux États-Unis, c'est le Russe Kolmogorov qui fut davantage lu et cité. L'ironie du sort, c'est que Solomonoff recevra le prix Kolmogorov en 2003, pour sa découverte de la complexité de Kolmogorov !

Puisque, d'une certaine manière, ce chapitre est aussi un hommage à Solomonoff, je vais m'autoriser l'affront de violer la terminologie communément acceptée, et parler de complexité de Solomonoff au lieu de complexité de Kolmogorov.

De façon grossière, la complexité de Solomonoff, donc, est la longueur du plus court code source qui, une fois lancé, génère la suite numérique en question. Cependant, comme tout programmeur le sait bien, la longueur de ce plus court code source dépend du langage de programmation utilisé. Un code source en Java sera presque toujours plus long qu'un code source en Matlab. La complexité de Solomonoff est donc mal définie. Elle dépend du langage utilisé, ou si l'on

⁸  A Curious Pattern Indeed | 3Blue1Brown | Grant Sanderson (2015)

considère des codes sources écrits directement en langage machine, de la machine de Turing considérée⁹.

Fort heureusement, cette dépendance n'est pas si grande. En effet, il existe des programmes informatiques, appelés compilateurs, qui sont capables de traduire les codes sources d'un langage en codes machines, voire de traduire des codes sources d'un langage de programmation en codes sources d'un autre langage. Ces compilateurs peuvent être très longs à décrire. Cependant, ils ont une longueur finie qui, de façon cruciale, ne dépend pas du code source à traduire.

Détaillons. Considérons deux machines de Turing universelles M et N . Appelons C un compilateur de M vers N , c'est-à-dire un code source de la machine N capable de traduire les codes sources de la machine M en codes exécutables par la machine N . Ce compilateur a alors une longueur de code source compilateur(N, M) indépendante de tout code à traduire.

Supposons maintenant que l'on ait un code source S pour la machine M qui génère une suite numérique. Alors on peut obtenir un code source pour la machine N qui génère la même suite numérique comme suit. On écrit d'abord le compilateur C , puis le code source S , et on demande ensuite à la machine de Turing N d'appliquer le code source S , interprété via le compilateur C . La machine de Turing N va alors se mettre à effectuer les mêmes calculs que ceux de la machine M lancés sur le code source S . De façon grossière, on peut écrire $M(S) = N(C, S)$. En particulier, la machine N va alors générer la bonne suite numérique¹⁰.

Mieux encore, la longueur du code source obtenu pour la machine de Turing N ne sera pas beaucoup plus longue que pour la machine de Turing M : elle sera égale à la longueur de S plus la longueur de C . En particulier, la complexité de Solomonoff de la suite numérique selon N sera au plus sa complexité de Solomonoff selon M plus une constante indépendante de la suite numérique. On écrit $K_N(\text{suite}) \leq K_M(\text{suite}) + \text{compilateur}(N, M)$. Le même raisonnement appliqué à un compilateur de N vers M nous fournit l'inégalité $K_M(\text{suite}) \leq K_N(\text{suite}) + \text{compilateur}(M, N)$. On en vient alors à la conclusion qu'à une constante près, la complexité de Solomonoff selon M est la même que selon N .

Tout ceci peut sembler obscur. Voici ce qu'il faut en retenir : certes la complexité de Solomonoff d'une suite n'est pas une quantité *objective*, mais elle l'est « presque » — surtout si l'on considère des machines universelles « raisonnables ». En tout cas, les informaticiens ont appris à vivre avec. Et bien, un peu plus loin, on verra que cette subjectivité de la complexité de Solomonoff est précisément la cause de la subjectivité des probabilités du démon de Solomonoff. Et à l'instar des informaticiens, les bayésiens ont appris à vivre avec.

⁹Il est bon de signaler que le code source minimal en Java ne ressemblera probablement pas du tout à un code « bien écrit » en Java. Typiquement, il contiendra en lui une procédure de décompression de fichier, et le gros du code source sera le résultat d'une compression. En particulier, ce code source sera très certainement très illisible pour un humain.

¹⁰  *La machine de Turing* | Passe-Science | T. Cabaret (2015)

Revenons-en à notre suite 1, 2, 4, 8, 16. Pour savoir qui de 31 ou 32 est le successeur le plus « évident » de cette suite, on peut étudier la complexité de Solomonoff des suites 1, 2, 4, 8, 16, 31 et 1, 2, 4, 8, 16, 32. En vertu du rasoir d'Ockham, la suite dont la complexité de Solomonoff est la plus petite sera plus crédible. Bien entendu, parce que la complexité de Solomonoff est subjective, la réponse à cette question l'est aussi ! Cependant, tout informaticien sentira que la seconde suite, celle complétée par 32, est « plus facile à coder » que la première, dans tout langage de programmation « raisonnable¹¹ ».

En fait, notre exemple est encore trop simpliste pour que la complexité de Solomonoff s'applique de manière univoque. En effet, dans ce cas, il semble que, dans de nombreux langages, la manière la plus succincte de décrire ces suites est de les décrire explicitement, termes après termes. Là où la complexité de Solomonoff deviendra beaucoup plus pertinente, c'est à partir du moment où les suites numériques en question sont beaucoup plus longues.

Typiquement, si au lieu de considérer les 5 premières puissances de 2, on considérait maintenant la suite des 100 premières puissances de 2, dès lors, décrire chaque terme de la suite un à un devient laborieux. Pour tout langage « raisonnable », l'informaticien saura faire beaucoup mieux, et écrire un court programme qui calcule ces puissances successives de 2. On aurait là un cas où la complexité de Solomonoff de la suite est nettement inférieure à la longueur de la suite, quel que soit le langage de programmation en question — pourvu que ce langage soit « raisonnable ». On pourra alors raisonnablement dire que le 101-ième élément de la suite est très probablement 2^{100} .

Le mariage de l'algorithme et des probabilités

La complexité de Solomonoff est un concept fantastique pour étudier les suites numériques qui découlent de règles de calculs précises et sans erreur. Cependant, dès que l'on se tourne vers les sciences empiriques, on est inéluctablement envahi par les imprécisions et les erreurs. Si la suite numérique correspond à des mesures physiques ou des données des sciences sociales, on doit s'attendre à ce que les éléments de la suite ne coïncident pas tout à fait avec les puissances de 2. Supposons donc que la suite qui nous est fournie est en fait : 0,9 ; 2 ; 4,1 ; 7,9 ; 15,8. On aurait maintenant envie de prédire une valeur autour de 31 ou 32, mais avec une erreur de mesure de l'ordre de 0,2.

Le langage des probabilités devient alors inéluctable. En particulier, au lieu d'effectuer une prédiction déterministe — comme dire que le prochain élément de la suite sera exactement 32 — on aurait alors envie de faire une prédiction probabiliste. Typiquement, on pourrait vouloir dire que 32 est très probable,

¹¹Pour toute suite, on peut toujours concevoir un langage qui génère cette suite avec peu d'instructions. Cependant, si la suite est sophistiquée, le langage qui permet de décrire cette suite succinctement ne semblera pas « raisonnable », ni « naturel ». Il s'agira d'un langage optimisé pour la suite en question.

mais que des nombres comme 31,9 ou 32,2 le sont aussi. En particulier, une prédiction probabiliste bien définie doit être capable de calculer, pour toute complétion possible de la suite, la probabilité de cette complétion.

En combinant les théories du calcul de Gödel, Church et Turing aux probabilités de Bayes, Price et Laplace, on en vient alors à une nouvelle définition de ce qu'est une théorie. Cette définition est le fruit du génie de Solomonoff. Pour Solomonoff, une théorie est alors un code source d'une machine de Turing qui, étant donné une suite numérique, calcule la probabilité de cette suite numérique.

Jusque-là, par souci pédagogique, je me suis restreint aux suites numériques. Cependant, notamment suite aux travaux de Shannon dont on reparlera, toute suite de données peut être traduite en une suite de 0 et de 1. On parle de *suites binaires*. Ainsi, sans perte de généralité, pour l'informaticien, toutes les données auxquelles vos sens vous confrontent, des images que vos yeux regardent aux sons que vos oreilles écoutent en passant par les odeurs que votre nez sent ou le sens de l'équilibre que vos oreilles internes mesurent, tout peut se décrire par une (très longue) suite¹² de 0 et de 1. Une théorie selon Solomonoff est donc un algorithme qui, étant donné une suite binaire finie, en calcule la probabilité.

Une telle définition de ce qu'est une théorie peut paraître restrictive. Nombre de théories communément considérées scientifiques, comme la théorie de l'évolution ou les lois de Newton, ne semblent pas du tout correspondre à un algorithme de calcul de probabilités de suites binaires. Il est vrai. D'une certaine manière, on peut considérer que ces théories sont tout simplement trop ambiguës, imprécises ou incomplètes vis-à-vis des standards de Solomonoff.

Cependant, il est peut-être plus intéressant de voir que les vrais calculateurs de probabilités d'aujourd'hui sont surtout les cerveaux de ceux qui pensent selon ces théories. À ces cerveaux, il est possible de raconter une histoire de l'évolution du vivant, et ils pourront alors rétorquer une estimation de la crédibilité de cette histoire. Or, si l'on en croit la thèse de Church-Turing, ces cerveaux ne font qu'effectuer des calculs qu'une machine de Turing avec le code source adéquat pourrait effectuer. Comme les histoires racontées sont encodables en une suite de 0 et de 1, il semble que les scientifiques en général, et les évolutionnistes en particulier, ne sont finalement pas si distants du formalisme de Solomonoff.

En particulier, dès lors, la théorie de l'évolution ou les lois de Newton correspondent à un ensemble de sous-procédures auxquelles les algorithmes prédictifs font appel. En fait, les théories scientifiques sont souvent davantage un ensemble d'équations, qu'un informaticien modélisera par des structures algorithmiques. L'informaticien combinera ensuite ces structures à des algorithmes de résolutions d'équations que l'on a tous appris à l'école. Voilà qui lui fournira des bibliothèques de codes qui seront très certainement utiles à un algorithme prédictif crédible.

¹²Le débit d'information capté par vos sens est estimé à quelque chose entre 10^7 et 10^{10} bits d'information par seconde, même s'il est très vite très compressé. Voir par exemple :

 Two views of brain function | Trends in Cognitive Sciences | M. Raichle (2010)

On alors pourrait considérer que les bibliothèques incontournables correspondent à des sortes de lois fondamentales. Dans le langage usuel des scientifiques, il est en effet fréquent de distinguer les théories invariantes à travers le temps, des données additionnelles utiles à la prédiction. Typiquement, le physicien va distinguer les *lois* de la physique des *états* physiques de l'univers. L'informaticien va souvent distinguer les *instructions* d'un algorithme et les *données* sur lesquelles l'algorithme est lancé. Et on a tendance à penser qu'une *loi* physique peut avoir une vérité (ou une validité) indépendante de tout *état* physique.

Cependant, Turing a justement montré que cette distinction entre *loi* et *état* n'avait rien de fondamental. Après tout, toute suite d'instructions (ou toute loi physique) peut être encodée comme étant des données d'une machine de Turing universelle. Comme les données, toute instruction (ou loi physique) n'est donc qu'une information. Tel fut d'ailleurs le principe fondamental qui permit à John von Neumann de concevoir l'architecture des ordinateurs modernes, et d'ainsi garantir le fait qu'ils soient bel et bien des machines de Turing universelles. Dans cette architecture, *instructions* et *données* sont toutes des informations enregistrées dans la mémoire des ordinateurs. Et il n'y a aucune différence fondamentale entre elles¹³.

Dès lors, selon Solomonoff, les seules informations dignes d'intérêts sont les descriptions d'algorithmes *prédictifs*. Insister sur ce caractère prédictif permet alors de clarifier la notion de complexité des théories si cruciale à l'application du rasoir d'Ockham. En effet, une loi de la physique, seule, peut paraître très simple. L'équation $\vec{F} = m\vec{a}$ tient en une poignée de symboles. Tout comme dire « c'est la faute aux aliens ». Mais c'est parce que de telles lois, seules, ne sont pas prédictives. Pour devenir prédictives, elles doivent être combinées à des descriptions de l'état physique de l'univers plus ou moins détaillées.

Voilà qui permet justement de distinguer la complexité de diverses lois. Les lois qui nécessitent des descriptions ultra-détaillées de l'état physique de l'univers pour produire des prédictions constitueront *in fine* une machinerie d'une énorme complexité de Solomonoff. En termes algorithmiques, la combinaison d'une telle loi et de la description de l'état physique constituera alors un code source extrêmement long. Par opposition, des théories qui parviennent à effectuer des prédictions avec une description très partielle de l'état physique de l'univers seront formidablement plus simples. Et donc plus crédibles *a priori*.

Bref. Selon Solomonoff, une théorie digne d'intérêt doit contenir en elle la description (partielle) de l'état physique de l'univers, parce qu'elle doit avant tout être prédictive. Idéalement, elle devrait aussi tenir compte de son incertitude sur cet état physique de l'univers, voire de son incertitude sur ses incertitudes. Mais surtout, *in fine*, elle doit effectuer des prédictions (probabilistes). Enfin, pour être digne d'intérêt, les prédictions de cette théorie doivent être *calculables*. Car, à en croire la thèse de Church-Turing, seules de telles prédictions pourront être faites à l'intérieur de notre univers.

¹³  *How do computers work? The Von Neumann Architecture* | Solid State Tech (2017)

Si l'on suit les pas de Solomonoff, il faut alors croire que l'ensemble des théories dignes d'intérêt est exactement l'ensemble des algorithmes de calcul de probabilités de suites binaires. De la même manière qu'Aaronson prétend que « si vous y réfléchissez suffisamment longtemps, vous finirez par conclure que tout calcul peut être réalisé par une machine de Turing », j'irai jusqu'à suggérer que, si vous y réfléchissez suffisamment longtemps, vous finirez par conclure que toute théorie prédictive est une théorie à la Solomonoff.

Le préjugé de Solomonoff*

En bon bayésien, Solomonoff propose ensuite de mettre en compétition les différentes théories prédictives et de juger des crédences respectives de ces théories. Et bien sûr, pour ce faire, il lui faut d'abord déterminer un *a priori* sur l'ensemble des algorithmes de calcul des probabilités de suites binaires. Pour que ce préjugé de Solomonoff soit conforme au bayésianisme, il faut en particulier que la somme des crédences *a priori* en les différentes théories prédictives soit égale à 1.

Bien entendu, il y a de nombreuses manières de garantir cela. Mais, puisque le nombre de codes sources à n caractères croît exponentiellement avec n , les codes sources longs seront intuitivement exponentiellement moins crédibles que les codes sources courts. Voilà qui implémente nécessairement une version très forte du rasoir d'Ockham, sur laquelle on reviendra plus tard.

De façon plus concrète, étant donné un langage de programmation ou une machine de Turing, une manière assez canonique de choisir une loi de probabilité *a priori* sur l'ensemble des théories prédictives calculables est la suivante¹⁴. Appelons $c_n \geq 0$ le nombre de code sources de longueur n qui correspondent à des théories prédictives. On peut considérer que chaque code source de la sorte a une probabilité *a priori* égale à¹⁵ $1/(c_n 2^n)$. Autrement dit, la probabilité *a priori* d'une théorie T est déterminée par la longueur $K(T)$ de sa description dans le langage choisi, via l'équation¹⁶ $\mathbb{P}[T] = 1/(c_{K(T)} 2^{K(T)})$.

En particulier, cette probabilité *a priori* est donc intimement liée à la complexité de Solomonoff, puisqu'elle repose sur la longueur des descriptions des théories prédictives. Or ces longueurs de description dépendent du choix du langage utilisé. Elles sont donc *subjectives*. Autrement dit, on vient de lier la subjectivité de l'*a priori* bayésien et la subjectivité de la complexité de Solomonoff. Par universalité de la machine de Turing et existence de compilateurs, on conclut que cette subjectivité est arbitraire... mais pas si arbitraire que cela !

¹⁴Il ne s'agit bien sûr pas de la *seule* loi de probabilité imaginable. Ce choix est subjectif. Mais l'argument important, c'est que, parce que le nombre c_n de théories de longueur n croît exponentiellement en n , selon tout préjugé « raisonnable », la crédence *a priori* en une théorie décroît exponentiellement vite avec la longueur de sa description.

¹⁵Techniquement, il faudrait traiter spécifiquement les cas $c_n = 0$.

¹⁶Je cache sous le tapis de nombreuses difficultés techniques dues notamment aux limites calculatoires des machines de Turing... On reviendra sur certaines.

Comme l'explique si majestueusement Solomonoff : « pendant bien longtemps, je sentais que le fait que ma théorie algorithmique des probabilités dépendait d'une machine référence était un défaut sérieux de mon concept, et j'ai essayé de trouver une machine universelle “objective”. Quand j'ai cru avoir enfin trouvé une telle machine, je me rendis compte que je n'en voulais pas — je n'en trouvais pas d'utilité ! [...] Il est possible de faire des prédictions sans données, mais il n'est pas possible de faire de prédition sans *a priori*. »

Bayes au secours du démon de Solomonoff*

Maintenant que le préjugé de Solomonoff est construit, il ne reste plus qu'à appliquer la formule de Bayes pour déterminer les théories à la Solomonoff les plus crédibles ! Supposons que l'on ait observé jusque-là la suite binaire¹⁷ a_1, a_2, \dots, a_n . La crédence bayésienne en une théorie prédictive T s'obtient alors via la formule du savoir suivante :

$$\mathbb{P}[T|a_1, \dots, a_n] = \frac{\mathbb{P}[a_1, \dots, a_n|T]\mathbb{P}[T]}{\mathbb{P}[a_1, \dots, a_n|T]\mathbb{P}[T] + \sum_{A \neq T} \mathbb{P}[a_1, \dots, a_n|A]\mathbb{P}[A]},$$

où les théories prédictives A sont toutes les alternatives à T .

Cependant, là n'est pas tant l'objectif du démon de Solomonoff. Son but est la prédition, pas le calcul des crédences — même si celles-ci sont utiles. Pour en arriver là, à l'aide des probabilités conditionnelles, on peut calculer la prédition d'une théorie T donnée. Selon la théorie prédictive T , la probabilité que la suite a_1, \dots, a_n soit suivie par un 1 est

$$\mathbb{P}[a_{n+1} = 1|a_1, \dots, a_n, T] = \frac{\mathbb{P}[a_1, \dots, a_n, 1|T]}{\mathbb{P}[a_1, \dots, a_n|T]}.$$

En combinant ces deux équations, on en vient enfin à la prédition du démon de Solomonoff qui, comme on l'a vu au chapitre 3, s'obtient en prenant la moyenne des prédictions des différentes théories, pondérées par leurs crédences bayésiennes. Après quelques détails de calcul, que je vous épargne mais vous invite à refaire, on obtient la prédition suivante :

$$\mathbb{P}[a_{n+1} = 1 | a_1, \dots, a_n] = \frac{\sum_T \mathbb{P}[a_1, \dots, a_n, 1|T]\mathbb{P}[T]}{\sum_T \mathbb{P}[a_1, \dots, a_n|T]\mathbb{P}[T]}.$$

¹⁷Notez que contrairement aux statistiques fréquentistes, on ne fait là aucune hypothèse *iid*, c'est-à-dire qu'on n'a jamais à supposer que l'on dispose de variables indépendantes, et encore moins de variables identiquement distribuées.

C'est cette formule magique, appelée induction de Solomonoff, que le démon de Solomonoff applique à longueur de journée pour effectuer des prédictions ! Quelle formule incroyable ! Cette formule magique est la formule de Bayes dans sa forme la plus pure et la plus idéale. Autrement dit, être *pur bayésien*, c'est ni plus ni moins effectuer le calcul ci-dessus !

En particulier, une fois la machine de Turing ou le langage de programmation fixé, toute ambiguïté disparaît. Être rationnel, réfléchir et prédire n'est pas une combinaison d'une multitudes de règles imprécises, arbitraires et parfois incompatibles, dont seuls des génies incompris sont capables. Le savoir se déduit uniquement d'un calcul. Et c'est ce calcul qu'effectue le démon de Solomonoff.

La complétude de Solomonoff

Le théorème fondamental de l'induction de Solomonoff est ce que Solomonoff a appelé la complétude de sa formule. De manière grossière, la complétude de l'induction de Solomonoff signifie que s'il y a un *pattern* calculable dans les données, alors le démon de Solomonoff finira par trouver ce *pattern*, en un temps proportionnel à la complexité de Solomonoff du *pattern*¹⁸.

Plus précisément, plus les données sont sophistiquées, plus le démon de Solomonoff aura besoin de données pour apprendre leur structure cachée. Mais la complétude de Solomonoff montre que cette quantité de données nécessaires n'excèdera jamais la sophistication des données¹⁹ — y compris lorsque cette structure cachée est « bruitée » par des fluctuations aléatoires !

Ce théorème fondamental, et le fait qu'aucune autre approche ne semble faire aussi bien ou mieux, est, il me semble, un argument essentiellement conclusif pour le *bayesianisme* — même s'il faut d'abord se convaincre que toute théorie prédictive est une théorie à la Solomonoff, et même si l'idéal serait de prouver que toute approche ayant cette propriété est nécessairement une forme d'induction à la Solomonoff.

¹⁸Plus précisément, le théorème de complétude de Solomonoff dit que s'il existe une théorie sous-jacente T^* à découvrir, alors l'espérance de la somme cumulée de toutes les erreurs de prédictions du démon de Solomonoff est bornée par la complexité de Solomonoff de T^* . En mesurant ces erreurs via la divergence KL dont on parlera au chapitre 15, on obtient

$$\mathbb{E}_a \left[\sum_{n=0}^{\infty} D_{KL} (\mathbb{P}[\cdot | a_{1:n}, T^*] || \mathbb{P}[\cdot | a_{1:n}]) \middle| T^* \right] \leq 2K(T^*),$$

pour un langage de programmation dont les caractères sont binaires. D'ailleurs, on a là une justification du principe d'uniformité de Hume. Celui-ci revient à postuler l'existence d'une théorie sous-jacente T^* . Or, c'est précisément ce que garantit la thèse de Church-Turing. Par ailleurs, au début de l'épisode 7 du podcast Axiome, je prétends que la complétude de Solomonoff résout le paradoxe logique du *grue*.

¹⁹Le mot « sophistication » correspond d'ailleurs ici à une définition très précise, qui sera introduite au chapitre 18.

L'incalculabilité de Solomonoff

Devant tout cet enthousiasme, vous pourriez vous demander si, par conséquent, l'induction de Solomonoff ne mettrait pas fin à la quête d'une philosophie du savoir. En grande partie, à mon sens, la réponse est oui. Mais il y a tout de même une énorme faille dans l'induction de Solomonoff qui me force à devoir écrire le reste de ce livre. L'induction de Solomonoff est en fait trop difficile à appliquer. Non seulement par nous autres et nos capacités cognitives limitées — souvenez-vous que même Erdős luttait pour comprendre une version simpliste de la formule de Bayes. Mais aussi par les ordinateurs.

L'induction de Solomonoff est *incalculable*. Qu'est-ce que cela veut dire ? Cela veut dire qu'aucune machine de Turing ne pourra l'appliquer rigoureusement. Il y a une raison très simple à cela, mais peu pertinente, et surtout, une autre raison subtile et dévastatrice.

La raison simple est que, pour être appliquée rigoureusement, l'induction de Solomonoff requiert l'étude simultanée d'une infinité de théories prédictives, tout simplement parce qu'il existe une infinité d'algorithmes. Cependant, aucun ordinateur, ni aucun réseau d'ordinateurs, ne peut effectuer ce calcul infini.

Ceci étant dit, on pourrait alors rétorquer que, par construction de l'*a priori*, les théories trop complexes sont, de toutes manières, associées à des probabilités exponentiellement négligeables. Si l'on néglige ces théories dont on sait qu'elles n'auront qu'un impact limité sur la prédiction de l'induction de Solomonoff, ne pourrait-on pas alors calculer une très bonne approximation de l'induction exacte de Solomonoff ? Malheureusement, non.

La vraie difficulté de l'induction de Solomonoff n'est pas le nombre de théories à considérer, mais les calculs que requièrent ces théories. De nos jours, les algorithmes qui tournent sur nos ordinateurs ont tendance à tourner vite et bien. C'est parce que les ingénieurs en programmation ont bien travaillé ! Mais en général, il est en fait très difficile de savoir si un algorithme va tourner vite. Il est même tout aussi difficile de savoir s'il va finir par terminer, ou si son calcul ne fait que se complexifier sans jamais terminer.

L'un des exemples les plus frappants d'un tel algorithme est celui de la conjecture de Syracuse, aussi appelée conjecture de Collatz, d'Ulam, tchèque ou $3n + 1$. On donne un entier à cet algorithme. Si cet entier est 1, il s'arrête là ; si l'entier est pair, il le divise par 2 ; s'il est impair, il le multiplie par 3 et ajoute 1. Puis, l'algorithme répète l'opération au résultat qu'il a obtenu. La conjecture de Syracuse pose la question suivante : cet algorithme terminera-t-il toujours, quel que soit l'entier qu'on lui propose²⁰ ?

Aussi étrange que cela puisse paraître, on ne connaît pas la réponse à cette question toute simple. Pire, on n'a aucune idée de comment approcher ce pro-

²⁰  Top 5 des problèmes de maths simples mais non résolus | Micmaths | M. Launay (2016)

blème, ni même s'il existe une façon de le résoudre. Le grand mathématicien Paul Erdős affirmait : « les mathématiques ne sont pas encore prêtes pour de tels problèmes. »

Ce cas est peut-être une instance de l'indécidabilité découverte par Turing. Juste après avoir défini la notion de calcul via la machine de Turing, Turing s'est empressé de se demander si l'on pouvait anticiper la terminaison d'un calcul avant de se lancer dedans. Existe-t-il une façon de déterminer si un calcul termine ? Y a-t-il un super-calcul qui détermine si un calcul termine ? Ou dit encore autrement, peut-on fabriquer une machine de Turing qui prédit en temps fini si d'autres machines de Turing terminent en temps fini ?

Si vous avez l'impression que la question de Turing ressemble à un serpent qui se mord la queue, ce n'est pas un hasard. En s'inspirant de l'argument de la diagonale de Cantor, du paradoxe de Russell ou encore du théorème de Gödel, Turing montra via un argument d'auto-référence que la réponse à ces questions était non. On ne peut pas toujours prédire l'arrêt d'un calcul. On dit que le problème de l'arrêt est *incalculable* (ou indécidable)²¹.

Et ça, c'est un énorme problème pour l'induction de Solomonoff. En effet, pour effectuer l'induction de Solomonoff, il faut calculer les prédictions $\mathbb{P}[a_1, \dots, a_n | T]$ de diverses théories T . Dans l'idéal, doivent être prises en compte toutes les quantités qui correspondent à un calcul qui termine. Cependant, si l'on en croit la thèse de Church-Turing, il est physiquement impossible de déterminer la terminaison de ces calculs. En particulier, la conséquence dramatique de ce raisonnement, c'est qu'après un temps de calcul fini, il est impossible d'exclure le fait que les théories prédictives dont les calculs sont en cours termineront.

Pire encore, il est aussi impossible de prédire si leurs éventuelles prédictions modifieront drastiquement le résultat calculé jusque-là. En général, il est donc impossible de mesurer le degré de validité d'un résultat partiel obtenu après un temps fini. L'induction de Solomonoff est non seulement incalculable. Toute approximation de l'induction l'est aussi !

L'incomplétude de Solomonoff

Cette conclusion embarrassante donne envie de rejeter le bayésianisme. Certes, le démon de Solomonoff, qui transcende les lois de la physique et la thèse de Church-Turing, saura détecter toute régularité dans un jeu de données. Mais ce démon de Solomonoff n'a pas d'équivalent physique. Nous et nos ordinateurs, qui sommes vraisemblablement contraints par la thèse de Church-Turing, semblons dans l'impossibilité d'effectuer une approximation de l'induction de Solomonoff. De façon générale, il nous est et nous sera toujours physiquement impossible de calculer la crédence adéquate en nos approximations de la formule de Bayes. À quoi bon vénérer cette formule du savoir ?

²¹  *Incomplétude* | Passe-Science | T. Cabaret (2015)

La réponse de Solomonoff réside dans un autre de ses théorèmes. Cet autre théorème affirme que tout algorithme est nécessairement incomplet. Plus précisément, toute philosophie du savoir calculable sera incapable de détecter toute régularité d'un jeu de données. Tel est le stupéfiant théorème d'incomplétude de Solomonoff — qui me fascine encore plus que celui de Gödel !

Dit autrement, quelle que soit votre philosophie du savoir, pourvu qu'il s'agisse d'une philosophie *calculable*, il existera des univers qui vous piègeront, et vous conduiront à *constamment* effectuer des prédictions très erronées ! La *calculabilité* et la *complétude* sont deux propriétés incompatibles²².

Ainsi, l'incalculabilité malheureuse de l'induction de Solomonoff est précisément ce qui lui permet d'échapper au théorème d'incomplétude des philosophies du savoir calculables. Pour Solomonoff, cette incalculabilité n'est pas une pathologie ; c'est une propriété nécessaire à toute philosophie du savoir désirable.

En quête de pragmatisme

J'ai été à la fois émerveillé et abasourdi par Solomonoff. Sa théorie est solidement fondée sur les notions les plus fondamentales de l'informatique et des probabilités. La construction de Solomonoff est incroyablement naturelle, dans le sens où c'est exactement celle que j'ai commencé à faire au moment de longuement méditer le bayésianisme, et que j'aurais faite si j'y avais réfléchi suffisamment longtemps et si j'avais les capacités cognitives nécessaires. Et dans le même temps, les conclusions de Solomonoff sont univoques, brutales et inattendues — même si, plus je méditais la formule de Bayes, plus je les sentais venir...

Le savoir et la rationalité nous sont hors de portée. Nous sommes contraints de nous contenter d'approximations dont on ne peut même pas mesurer le degré de validité. Pire encore, parce que nos ressources en puissance de calcul et en temps sont toujours limitées (notamment par les lois de la physique), nous sommes contraints de nous restreindre à une version extrêmement heuristique de l'induction de Solomonoff.

Cette restriction sera d'ailleurs d'autant plus contraignante que la taille des données observées est grande, comme c'est le cas du *Big Data*. De nos jours, nos données, numériques ou sensorielles, se comptent en giga, téra, peta, exa, voire zettaoctets. Autrement dit, la suite a_1, \dots, a_n que l'on étudie en pratique contient des milliards, voire des milliards de milliards d'éléments ! Il est même illusoire de stocker en mémoire de telles données, et plus illusoire encore d'espérer approximer l'induction de Solomonoff dans ces conditions. Voilà qui devrait forcer l'humilité et la prudence²³.

²²L'idée de la preuve est de supposer que le *pattern* des données piège constamment l'algorithme prédictif. De façon cruciale, contrairement au cas de l'induction de Solomonoff, ce *pattern* est là calculable.

²³  *La logique ne suffit pas* | IA 6 | Science4All | L.N. Hoang (2018)

Face à ce constat dévastateur, le reste de ce livre, comme le monde des sciences, des statistiques et de l'intelligence artificielle, est contraint de se contenter de philosophies du savoir dites *heuristiques*. On ne pourra pas *bien* savoir. Mais peut-être peut-on tout de même *assez bien* savoir. Pour ce faire, sachant que l'on sait comment *bien* savoir, on cherchera dans ce livre à s'inspirer du démon de Solomonoff et à coller autant que possible à ses conclusions.

Je vous propose ainsi d'introduire désormais une seconde philosophie, imprécise, et un second personnage fictif. Je les appellerai respectivement le *bayésianisme pragmatique* et le *bayésien pragmatique*. Contrairement à la *pure bayésienne*, le *bayésien pragmatique* sera contraint par des ressources limitées en puissance de calculs et en espace mémoire. Dès lors, il lui sera préférable de favoriser les calculs de nombreux algorithmes rapides et efficaces, plutôt que de ne prendre le temps de tester qu'une poignée d'algorithmes dont les temps de calcul sont élevés. Voilà qui exigera du *bayésien pragmatique* une connaissance intime des algorithmes.

En fait, le *bayésianisme pragmatique*, bien plus que sa version pure, requiert une théorie sophistiquée et avancée du calcul et de l'information, que l'on appelle plus généralement l'*informatique théorique*, et dont la compréhension peut être complémentée par des expériences de calculs qui appartiennent au domaine de l'informatique expérimentale ou empirique. Ces disciplines modernes des sciences, initiées dès les années 1930 par Gödel, Church et Turing, forment désormais l'un des domaines les plus incompris, les plus fascinants et les plus prometteurs de la recherche d'aujourd'hui.

La science informatique n'est pas uniquement l'art d'utiliser nos technologies modernes. En fait, pour le bayésien compétent que j'aspire à être, les concepts de l'informatique théorique forment avant tout l'arsenal d'outils le plus important à maîtriser pour déterminer une philosophie pragmatique optimale du savoir — en admettant que nous venons de déterminer *la* bonne philosophie idéalisée du savoir. Dans son livre *Quantum Computing since Democritus*, l'informaticien Scott Aaronson va même jusqu'à proposer de renommer l'informatique théorique en *épistémologie quantitative*. Il insiste en particulier sur l'importance de la théorie de la complexité algorithmique pour toute philosophie du savoir²⁴.

Cependant, je vous propose de laisser la quête de cette philosophie du savoir de côté pour l'instant. On y reviendra longuement à partir du chapitre 14. D'ici là, je vous propose d'observer l'omniprésence des principes bayésiens, que ce soit en cryptographie, en sociologie, en biologie ou encore dans l'émergence du consensus scientifique. Dans les prochains chapitres, on va s'éloigner du bayésianisme pour explorer des domaines qui peuvent sembler distants, avant de constater, encore et encore, que derrière tant de phénomènes divers et variés, se cachent en fait des principes bayésiens.

²⁴  *Why Philosophers Should Care About Computational Complexity* | S. Aaronson (2011)

Références en français

 *Les Métamorphoses du calcul : Une étonnante histoire des mathématiques* | Le Pommier | G. Dowek (2007)

 *Introduction à la calculabilité - la machine de Turing* | Wandida | R. Guerraoui (2013)

 *Alan Turing - Enigma, ordinateur et pomme empoisonnée* | e-penser | B. Benamran (2015)

 *La machine de Turing* | Passe-Science | T. Cabaret (2015)

 *Incomplétude* | Passe-Science | T. Cabaret (2015)

 *Les machines de Turing* | Math&Magique (2016)

 *Top 5 des problèmes de maths simples mais non résolus* | Micmaths | M. Launay (2016)

 *La diagonale dévastatrice de Cantor* | Infini 16 | Science4All | L.N. Hoang (2016)

 *Les théorèmes d'incomplétude de Gödel* | Infini 18 | Science4All | L.N. Hoang (2016)

 *La machine de Turing* | IA 4 | Science4All | L.N. Hoang (2017)

 *La logique ne suffit pas* | IA 6 | Science4All | L.N. Hoang (2018)

 *Linguistique causale* | Axiome 7 | T. Giraud et L.N. Hoang (2018)

Références en anglais

 *A preliminary report on a general theory of inductive inference (Report ZTB-138)* | Zator Co | R. Solomonoff (1960)

 *A formal theory of inductive inference. Part I* | Information and Control | R. Solomonoff (1964)

 *A formal theory of inductive inference. Part II* | Information and control | R. Solomonoff (1964)

 *The discovery of algorithmic probability* | Journal of Computer and System Sciences | R. Solomonoff (1997)

 *Algorithmic probability: Theory and applications* | Information theory and statistical learning | R. Solomonoff (2009)

 *Why Philosophers Should Care About Computational Complexity* | S. Aaronson (2011)

 *Two views of brain function* | Trends in Cognitive Sciences | M. Raichle (2010)

 *Quantum Computing since Democritus* | Cambridge University Press | S. Aaronson (2013)

- ▶ *The Universal Turing Machine* | ZettaBytes | R. Guerraoui (2016)
- ▶ *Turing and the Halting Problem* | Computerphile | M. Jago (2014)
- ▶ *A Curious Pattern Indeed* | 3Blue1Brown | Grant Sanderson (2015)
- ▶ *Circle Division Solution* | 3Blue1Brown | Grant Sanderson (2015)
- ▶ *Making a computer Turing complete* | B. Eater (2018)
- ▶ *How do computers work? The Von Neumann Architecture* | Solid State Tech (2017)

Affirmer que vous vous moquez du droit à la confidentialité parce que vous n'avez rien à cacher, c'est comme affirmer que vous vous moquez de la liberté d'expression parce que vous n'avez rien à dire.

Edward Snowden (1983-)

L'ennemi connaît le système.

Claude Shannon (1916-2001)

8

Garder le secret

Classé confidentiel

Pendant la guerre du Vietnam, les généraux américains voulaient déterminer la proportion de soldats qui fumaient de la marijuana. Le problème, c'est que si un supérieur demande à un soldat s'il fume, pour éviter toute punition, le soldat répondra probablement non. L'armée américaine avait besoin d'un sondage qui n'incrimine aucun de ceux qui y répondraient. Elle devait mettre en place un sondage qui garantisse mathématiquement la confidentialité des réponses. Seriez-vous capable de deviner comment l'armée américaine a pu y arriver ?

Garder le secret est un enjeu récurrent pour les militaires. On raconte que Jules César codait les messages qu'il envoyait et qu'il recevait, en décalant les lettres de l'alphabet. La lettre A était remplacée par la lettre D, la lettre B par la lettre E, le C par le F, et ainsi de suite. Plus tard, d'autres dirigeants militaires utilisèrent des encodages cryptographiques plus sophistiqués, en remplaçant le A par n'importe quelle lettre de l'alphabet, le B par une toute autre, et ainsi de suite. On parle de *codage par substitution*, car toute lettre est systématiquement substituée par une autre lettre.

L'avantage du codage par substitution est que, contrairement au codage par décalage de César, le nombre de codages possibles devient astronomique. En effet, dans le cas du codage de César, le nombre de codages possibles est le nombre de décalages possibles des lettres de l'alphabet. Il y a 26 lettres, donc 26 décalages possibles (dont un qui consiste à ne pas décaler les lettres). Et

le problème avec cela, c'est qu'un hacker n'aura que 26 décalages à tester pour déchiffrer un code.

Cependant, si on s'autorise désormais à effectuer n'importe quelle permutation des lettres, le nombre de possibilités est radicalement plus grand. En effet, on peut remplacer le A par n'importe laquelle des 26 lettres, puis le B par n'importe laquelle des 25 lettres restantes, puis le C par n'importe laquelle des 24 lettres restantes, et ainsi de suite. On voit que le nombre de codages par substitution est alors $26 \times 25 \times 24 \times \dots \times 2 \times 1$, que l'on note $26!$ (et que l'on prononce 26 « factorielle »). Ce nombre est gigantesque. Il est de l'ordre de 10^{26} , soit environ le nombre d'étoiles dans l'univers ! Même un ordinateur ne tourne pas suffisamment vite pour tester toutes ces combinaisons en moins de temps que l'âge de l'univers.

La seconde guerre mondiale et les progrès technologiques ont conduit à la mécanisation de la cryptographie. Ainsi, parmi les plus importantes machines de guerre nazies, on trouvait alors plusieurs machines dont le rôle n'était pas de tuer, mais de coder et décoder des messages secrets. Ces machines étaient *Enigma* et la machine de Lorenz, comme on en a déjà parlé au chapitre 6. Le nombre de combinaisons de ces machines était là encore gargantuesque. Il était illusoire de tester toutes les combinaisons.

La cryptographie d'aujourd'hui

Depuis, le contexte technologique a évolué et nous disposons tous de supercalculateurs interconnectés. Nous nous envoyons quotidiennement des milliards de messages à travers Internet. Pour sécuriser la confidentialité de nos messages, la cryptographie est plus indispensable que jamais.

Il nous faut recevoir les emails de nos amis, sans que quiconque puisse les surveiller. Il nous faut nous connecter à des réseaux sociaux, sans que personne puisse s'y connecter à notre place. Il nous faut demander à nos banques d'effectuer des transactions financières, et les banques doivent pouvoir certifier que ces demandes viennent bien de clients et non pas d'usurpateurs.

La solution à tous ces problèmes qui ont envahi notre quotidien est la cryptographie. Grâce à la cryptographie, les vendeurs et les acheteurs ont pu profiter d'un nouveau canal d'information pour faciliter les échanges commerciaux, sans lequel nombre d'entreprises comme PayPal, Amazon, Netflix, Uber ou autres Airbnb n'auraient jamais pu voir le jour¹.

Cette cryptographie repose essentiellement sur deux grandes découvertes de 1976 et 1977 : le protocole de Whitfield Diffie et Martin Hellman d'un côté, et le protocole de Ron Rivest, Adi Shamir et Leonard Adleman (RSA) de l'autre. Le protocole de Diffie-Hellman est une astucieuse façon pour Alice et Bob de

¹  10 prouesses de la cryptographie | Crypto | String Theory | L.N. Hoang (2018)

créer un secret partagé en communiquant à voix haute à travers Internet. À l'aide de ce secret partagé, Alice et Bob pourront déterminer un codage commun, et pourront ainsi communiquer à travers Internet de manière sécurisée, même s'ils ne se sont jamais rencontrés physiquement !

Le protocole de Diffie-Hellman est utilisé depuis par de nombreuses applications comme WhatsApp. Ce protocole garantit que même les entreprises possédant ces applications seront mathématiquement incapables de lire nos communications sécurisées². Aujourd'hui, on a tendance à prendre ceci pour acquis, mais avant la découverte de Diffie-Hellman, ce n'était pas clair que cela puisse se faire un jour.

Le protocole RSA, quant à lui, a permis une cryptographie asymétrique. Pour utiliser RSA, Alice doit créer une paire de clés. Elle révèle la clé publique, mais garde secrète sa clé privée. Bob, comme n'importe qui d'autre, pourra alors envoyer des messages encodés à Alice, en chiffrant son message avec la clé publique d'Alice. Seule Alice pourra alors déchiffrer ce message, puisque la clé privée est nécessaire pour ce faire.

Mieux encore, Alice peut alors signer des messages qu'elle enverrait à Bob, en chiffrant (un hash de) ces messages avec sa clé privée. Quand Bob les aura déchiffrés avec la clé publique, il saura alors que seule Alice a pu envoyer ces messages. En effet, la clé privée est indispensable pour chiffrer les (hashs des) messages que la clé publique pourra déchiffrer. En « signant » ainsi ces messages, Alice peut s'authentifier et prouver à sa banque ou son réseau social que les messages qu'ils reçoivent viennent bien d'elle et pas d'usurpateurs.

Dans tous les exemples de cryptographie que l'on vient rapidement de voir, la sécurité est garantie par l'immensité du nombre de codages possibles, et par l'hypothèse selon laquelle tout hacker devra tester une grosse portion de ces codages pour réussir son coup. Ainsi, s'il y a 10^{20} possibilités, même si le hacker arrive à écarter 99% des possibilités, il lui restera encore 10^{18} possibilités à tester, ce qui prendra encore beaucoup de temps, même avec les ordinateurs modernes.

Cependant, tout ce raisonnement ignore les deux outils de prédilection du bayésien : le préjugé et la formule de Bayes.

Bayes à l'assaut des codes cryptés

Le 2 mai 1568, après de nombreuses péripéties malheureuses, Marie Stuart I d'Écosse fuit l'Écosse pour trouver refuge chez sa cousine en Angleterre. Mais, parce que les catholiques considéraient que Marie était l'héritière légitime du

²En fait, cette affirmation repose sur la conjecture non démontrée selon laquelle le logarithme discret ne peut pas être résolu en temps polynomial, et sur l'absence d'ordinateurs quantiques.

trône dans un pays devenu protestant, la reine Élisabeth I d'Angleterre la traita comme une rivale, et la fit enfermer pendant les 19 ans qui suivirent.

C'est en prison que Marie se mit à user de cryptographie. Ses messages étant interceptés par la cour de la reine, Marie s'assura d'utiliser le codage par substitution pour communiquer secrètement, notamment avec un certain Anthony Babington. Cependant, il semblerait que ces messages finirent par être décodés, révélant ainsi un complot pour assassiner la reine. La mise à jour de ce complot, connu sous le nom de conspiration de Babington, conduisit ensuite à l'exécution de Marie, le 8 février 1587.

Mais, en l'absence d'ordinateurs, comment le codage par substitution a-t-il pu être cassé ? On l'a vu, le nombre de codages par substitution est plus grand que le nombre d'étoiles dans l'univers. Ce n'est donc certainement pas en les testant tous que la cour de la reine Élisabeth I a pu décoder les messages de Marie. Pour casser le codage par substitution, la cour s'appuya sur ses préjugés. En particulier, le préjugé qui cassa le code de Marie fut celui sur la langue anglaise.

Dans la langue anglaise, les lettres E, T, A, O et I sont beaucoup plus fréquentes que toute autre. Par conséquent, la lettre la plus fréquente du message encrypté est sans doute le substitut de la lettre E, la seconde plus fréquente est sans doute le substitut du T, et ainsi de suite. Mais ce n'est pas tout ! Les mots de la langue anglaise sont très rigides, puisque rares sont les combinaisons de lettres qui forment un mot. Ainsi, une fois qu'on aura décodé T*E et qu'il faut déterminer la lettre * manquante, on devinera que * est très probablement un H, un I ou un O. Mieux encore, s'il s'agit du premier mot d'une phrase, on pourra être très confiant en le fait qu'il s'agit d'un H.

Vous voyez que surgit tout à coup naturellement le langage des probabilités. En fait, derrière le raisonnement intuitif que je viens de mettre en avant, se cache la formule de Bayes. Il s'agit de déterminer nos crédences en le message original et le codage utilisé, étant donné le message codé. Ou dit autrement, il s'agit de déterminer les causes, étant donné les conséquences. Voilà précisément le contexte d'application de la formule de Bayes.

C'est pour des raisons similaires qu'un mot de passe généré aléatoirement est beaucoup plus sécurisé que le mot de passe « 123456 ». Les hackers ont des préjugés justifiés sur les mots de passe des utilisateurs, car ils savent que certains mots de passe sont plus fréquents que d'autres. Un hacker malin lancera ainsi un algorithme qui testera d'abord les mots de passe les plus probables. Tel fut aussi le principe des machines à casser les codes de Turing pendant la seconde guerre mondiale.

Bien entendu, dans le cas de Marie et du codage par substitution, l'ensemble des causes est très grand, ce qui rend le calcul bayésien humainement impossible. Cependant, le fait que le codage se fasse lettre par lettre permet de segmenter l'ensemble des codages pour en simplifier l'analyse. Ou dit plus simplement, on peut d'abord se demander quel est le substitut du E, puis quel est celui du T, et ainsi de suite.

Il s'agit là d'un luxe dont Alan Turing ne disposait pas lorsqu'il s'attaqua à *Enigma* ou à la machine de Lorenz. Pour casser ces codes plus sophistiqués, Turing dut davantage formaliser le calcul bayésien, en utilisant notamment l'échelle logarithmique dont on parlera dans un futur chapitre. Les calculs de Turing auront ensuite été formalisés par le grand Claude Shannon, qui développa ainsi une théorie mathématique de la communication et de la cryptographie.

Depuis, les informaticiens font bien attention à ce que les messages encryptés n'aient pas de propriétés spécifiques qui permettraient des décodages comme ceux dont on vient de parler. De façon amusante, comme Shannon le découvrit, une bonne manière de ce faire consiste à d'abord compresser autant que possible le message initial. En effet, en compressant un message, on détruit ce faisant toute sa rigidité, comme le fait que, dans le mot T^*E , le signe * ne peut correspondre qu'à quelques lettres.

De nos jours, les systèmes cryptographiques construits à partir des mathématiques de Shannon, Diffie-Hellman et Rivest-Shamir-Adleman, sont considérés mathématiquement sécurisés, pourvu que les ordinateurs quantiques ne voient pas le jour. D'autres systèmes cryptographiques encore plus solides, dits *post-quantiques*, ont été proposés. Toutefois, on n'a pas encore réussi à prouver qu'aucun algorithme, classique ou quantique, ne sera capable de les casser — ce problème est d'ailleurs intimement relié au prestigieux problème P versus NP.

Le sondage randomisé

Cependant, toute cette cryptographie ne résout pas le problème du sondage de l'armée américaine sur la consommation de marijuana. Certes, les soldats pourraient chiffrer leur réponse. Mais si quelqu'un déchiffre ces réponses, alors la confidentialité des réponses sera violée et les soldats refuseront de révéler les réponses de manière honnête³. D'un autre côté, si personne ne déchiffre les réponses, alors l'armée américaine n'aura pas avancé. Elle n'aura rien appris.

D'une certaine manière, on veut autoriser l'armée à apprendre sur ses soldats, sans toutefois ne rien apprendre sur aucun de ses soldats.

L'astuce géniale fut de *randomiser* les réponses. Plus précisément, avant de répondre, chaque soldat lance une pièce. Si la pièce tombe sur pile, le soldat répond honnêtement. Si la pièce tombe sur face, alors il répond oui. De façon cruciale, lorsque le chef militaire demande au soldat s'il a fumé de la marijuana, ce chef militaire ne sait pas, et ne saura jamais, si la pièce est tombée sur pile ou face. Ainsi, si le soldat répond oui, le chef militaire ne peut pas savoir si le soldat

³On peut imaginer une urne dans laquelle les soldats déposeraient leurs réponses de manière anonymisée. Cependant, comme on le verra, l'anonymisation ne garantit (en général) pas la confidentialité. Typiquement, on peut imaginer que l'armée truque l'urne en la remplissant uniquement de « non », sauf pour le bulletin d'un soldat. L'armée pourra alors déterminer la réponse de ce soldat en fonction de la présence ou de l'absence de « oui » dans l'urne.

répond oui parce qu'il a fumé, ou s'il répond oui parce que sa pièce est tombée sur face. Autrement dit, le soldat dispose d'un déni plausible. Néanmoins, en recoupant les réponses des différents soldats, l'armée américaine peut désormais déterminer la proportion de ses soldats qui ont fumé.

En effet, supposons que l'on a reçu 200 réponses, 160 étant des oui, les 40 restantes étant des non. On sait alors que la moitié de toutes les réponses étaient des cas où la pièce était tombée sur face, et où le soldat a donc dit oui à cause de la pièce. Par conséquent, environ 100 des oui sont en fait des oui forcés par la pièce. Il reste alors 100 réponses honnêtes, 60 oui et 40 non. Ces réponses ont été des réponses honnêtes. Par conséquent, on en déduit qu'environ 60 % des soldats américains fument de la marijuana⁴. On a réussi à sonder les soldats sans compromettre aucun des soldats⁵ !

En fait, pas tout à fait. Si l'on raisonne en termes bayésiens, on se rend compte que notre *a priori* sur le fait qu'un soldat pris au hasard fume de la marijuana a évolué. Ainsi, avant le sondage, le chef militaire pensait peut-être que 20 % de ses soldats fumaient. Le sondage aura alors drastiquement modifié le préjugé du chef militaire sur la consommation de marijuana d'un de ses soldats pris au hasard. Ce préjugé est passé d'une probabilité de 20 % à une probabilité d'environ 60 %. Bien entendu, ceci n'a rien d'étonnant : c'était, après tout, l'objectif du sondage.

Cependant, là où le bât blesse, c'est la suspicion du chef militaire concernant les soldats sondés qui ont répondu oui au sondage. En effet, si un soldat a répondu oui plutôt que non, c'est qu'il a tout de même plus de chance d'être un consommateur de marijuana que s'il avait répondu non, et même que ceux qui n'ont pas répondu du tout. Pour déterminer la crédence d'un chef militaire bayésien en la consommation d'un soldat ayant répondu oui, il nous faut appliquer là encore la formule de Bayes. Écrivons-la :

$$\mathbb{P}[\text{Marijuana}|\text{oui}] = \frac{\mathbb{P}[\text{oui}|\text{Marijuana}] \mathbb{P}[\text{Marijuana}]}{\mathbb{P}[\text{oui}]}.$$

Je vous épargne les calculs (mais je vous invite à les faire chez vous !). Avec les données ci-dessus, le chef militaire devrait conclure à une crédence de 75 % en la consommation de marijuana d'un soldat ayant répondu oui. Ce résultat contraste notamment avec la crédence de 60 % pour un soldat qui n'a pas participé au sondage. Autrement dit, en acceptant de répondre au sondage, le soldat s'est légèrement incriminé. Sa confidentialité n'a pas été entièrement épargnée — même si elle n'a pas été entièrement violée non plus.

⁴En bons bayésiens, on devrait en fait plutôt appliquer la loi de succession de Laplace que l'on a vue au chapitre 6, ce qui donne 61/102, voire partir d'un préjugé plus informé que Laplace. De plus, il ne faut pas oublier de calculer également l'incertitude sur le pourcentage obtenu !

⁵  *Il donne du cannabis à son chat, ça tourne mal | La statistique expliquée à mon chat* (2017)

Il y a, à l'inverse, un cas où la confidentialité du soldat est entièrement violée : le cas où sa réponse est non. En effet, dès lors, en supposant que le soldat n'a pas menti, le chef militaire peut être sûr que le soldat ne fume pas. Or ceci peut en fait être un énorme problème. En effet, il se pourrait alors qu'à l'avenir, pour une raison mystérieuse, des recherches découvrent que les soldats américains envoyés au Vietnam qui ne fumaient pas de marijuana ont de très grandes chances de contracter le cancer du côlon. Les assurances pourraient alors vouloir augmenter les frais des soldats dont elles sont sûrs qu'ils ne fument pas de marijuana.

Pour n'enfreindre la confidentialité de personne, il nous faut modifier le mécanisme du sondage. Pour ce faire, lorsque la pièce tombe sur face, le soldat tire désormais une seconde pièce qui déterminera sa réponse. Autrement dit, dans cette variante, tout soldat a une chance sur deux de répondre honnêtement, une chance sur quatre de répondre oui à cause des lancers de pièces, et une chance sur quatre de répondre non à cause des lancers de pièces. Comme précédemment, on voit que le chef peut déterminer la proportion de fumeurs.

De plus, en appliquant la même formule de Bayes que précédemment, le chef militaire aura une crédence *a posteriori* de 82 % en le fait qu'un soldat ayant répondu oui au sondage est effectivement un fumeur. Cette fois, à l'inverse, le chef militaire attribuera une crédence non nulle en le fait qu'un soldat ayant répondu non au sondage fume malgré tout. Cette crédence sera de 33 %. Ces quantités de 82 % et 33 % sont bien entendu à comparer avec la crédence de 60 % en le fait qu'un soldat n'ayant pas participé au sondage fume. En particulier, on a pris l'habitude de mesurer le degré de confidentialité par les rapports 82/60 et 60/33. Ainsi, dans ce cas, la perte de confidentialité n'est jamais pire qu'un facteur 2.

La confidentialité du sondage randomisé

Ces rapports que l'on a calculés ici dépendent toutefois de la proportion de 60 % de fumeurs. Or, les chefs militaires ne pouvaient pas l'anticiper avant le sondage. Forte de cette remarque, la chercheuse en informatique théorique Cynthia Dwork a inventé une nouvelle théorie pour étudier mathématiquement la confidentialité. Au cœur de cette théorie se trouve le concept de *confidentialité différentielle*, ou *differential privacy* en anglais.

Contrairement à l'analyse que l'on vient de faire, la confidentialité différentielle veut garantir un niveau de confidentialité avant que les soldats ne répondent au sondage. Autrement dit, elle détermine la pire perte de confidentialité que les chefs militaires risquaient de devoir imposer aux soldats sondés, quelle que soit la proportion de fumeurs⁶.

⁶  *What is Privacy?* Wandida | L.N. Hoang (2017)

Supposons que 1 % des soldats fument, et considérons un soldat qui a répondu oui. La crédence *a posteriori* qu'il fume se déduit de la formule de Bayes :

$$\begin{aligned}\mathbb{P}[\text{fumeur}|\text{oui}] &= \frac{\mathbb{P}[\text{oui}|\text{fumeur}] \mathbb{P}[\text{fumeur}]}{\mathbb{P}[\text{oui}|\text{fumeur}] \mathbb{P}[\text{fumeur}] + \mathbb{P}[\text{oui}|\text{non-fumeur}] \mathbb{P}[\text{non-fumeur}]} \\ &= \frac{\frac{3}{4} \cdot 0,01}{\frac{3}{4} \cdot 0,01 + \frac{1}{4} \cdot 0,99} \approx 0,029.\end{aligned}$$

Cette crédence est presque trois fois plus que si le soldat n'avait pas répondu ! En fait, dans le cas limite où la proportion de soldats fumeurs tend vers zéro, le facteur multiplicatif sera exactement 3. Il s'agit d'ailleurs du pire cas. Le sondage randomisé avec deux pièces est dit ($\ln 3$)-différentiellement confidentiel.

La confidentialité différentielle de Cynthia Dwork contraste en particulier avec les méthodes naïves de pseudonymisation des données encore largement répandues, notamment en épidémiologie, où les noms des personnes sondées sont effacés. Cependant, si l'on connaît l'âge, le sexe, l'adresse, le niveau socio-économique, la consommation alimentaire ou encore le niveau d'étude d'une personne, alors il est en général possible de croiser cette information avec d'autres données accessibles sur le web, et déterminer ainsi l'identité de la personne en question. *La pseudonymisation n'offre aucune garantie de confidentialité.*

La définition de la confidentialité différentielle*

La confidentialité différentielle de Cynthia Dwork est un critère très général pour distinguer les méthodes qui offrent une garantie démontrable de confidentialité. Comme pour le cas du sondage randomisé, imaginez que vous souhaitiez extraire une information utile en analysant les données de différents individus. Ce faisant, votre *a posteriori*, une fois l'information utile apprise, sera généralement distinct de votre préjugé — c'était, après tout, le but de l'extraction d'information utile ! Cependant, intuitivement, la confidentialité exige que vos crédences *a posteriori* ne discriminent pas les individus analysés, ni entre eux, ni par opposition à ceux dont les données n'ont jamais été analysées.

Et bien justement. Un mécanisme d'extraction d'information sera dit différemmentiellement confidentiel si, *a posteriori*, les crédences sur un individu analysé sont toujours presque les mêmes que celles d'un individu non-analysé. Par transitivité, ceci impliquera aussi la non-discrimination entre individus analysés. En effet, supposons qu'Alice et Charlie ont été analysés, mais que Bob ne l'a pas été. Alors, par définition de la confidentialité différentielle, les crédences *a posteriori* entre Alice et Bob sont nécessairement similaires. Il en va de même de celles entre Bob et Charlie aussi. Du coup les crédences *a posteriori* entre Alice et Charlie seront elles aussi similaires.

De façon plus formelle et sans perte de généralité, Dwork suppose que les données des personnes analysées se trouvent dans ce que l'on appelle une *base de données*. Pour Dwork, c'est la confidentialité différentielle de cette base de données qu'il faut garantir. L'astuce pour ce faire c'est d'interdire toute lecture de cette base de données autre que via l'un des rares mécanismes dont on a prouvé les propriétés de confidentialité différentielle.

Le cas limite est bien entendu le cas où aucun mécanisme d'extraction n'existe, qui est d'ailleurs équivalent au cas où tous les mécanismes d'extraction retournent des informations complètement indépendantes de la base de données. Dans ces cas, aucune information sur le contenu de la base de données ne peut être extraite. Tout se passe alors exactement comme si la base de données n'existe pas, ou comme si elle était chiffrée mais que personne ne pouvait la déchiffrer. À ce moment-là, il est clair que la base de données sera entièrement confidentielle. Mais elle sera aussi inutile.

En fait, de manière générale, il est impossible d'extraire une quelconque information utile, sans au moins partiellement violer la confidentialité des données. L'approche de Dwork consiste justement à quantifier le compromis entre extraction d'information utile et confidentialité. Ainsi, de manière grossière, la quantité de confidentialité différentielle perdue par un mécanisme d'extraction d'information sera mesurée par deux paramètres ϵ et δ . Un mécanisme ($\epsilon = 0, \delta = 0$)-différentiellement confidentiel sera parfaitement confidentiel — mais il n'extraira aucune information.

Reste à définir ϵ et δ . Pour ce faire, revenons-en à la définition intuitive. On a vu qu'un mécanisme différentiellement confidentiel ne devait pas conduire à une discrimination *a posteriori* entre une Alice dans la base de données et un Bob à l'extérieur de la base de données. L'astuce pour ce faire est de garantir que la réponse retournée par le mécanisme demeure essentiellement la même réponse, une fois Alice retirée de la base de données (et donc impossible à discriminer de Bob).

De façon plus précise, appelons X la base de données originale avec Alice, et Y la base de données obtenue après suppression des données d'Alice. Appelons R le résultat retourné par le mécanisme. Un mécanisme confidentiellement différentiel devra retourner des résultats R similaires pour les bases de données X et Y . Et ce sont les paramètres ϵ et δ qui mesureront cette similarité.

Plus formellement, un mécanisme de requête dans une base de données sera (ϵ, δ) -différentiellement confidentiel si, quelle que soit la base de donnée étudiée, avec grande probabilité $1 - \delta$, les crédences *a posteriori* qu'un individu dans la base de données ait une caractéristique donnée ne sont jamais e^ϵ fois plus grande que pour un individu extérieur à la base de données⁷. Autrement dit, *a posteriori*, tout se passe presque comme si Alice n'était pas dans la base de données.

⁷Dans le cas du sondage randomisé, on a $\delta = 0$, et $\epsilon = \ln 3$.

De façon purement formelle, la confidentialité différentielle peut s'écrire en une inégalité plus facile à manipuler pour les mathématiciens. Formellement, un mécanisme qui retourne des réponses R est (ε, δ) -différentiellement confidentiel si, pour toute paire de base de données X et Y qui ne diffère que d'un ajout ou d'une suppression des données d'un individu, on a l'inégalité

$$\mathbb{P}[R|X] \leq e^\varepsilon \mathbb{P}[R|Y] + \delta.$$

En particulier, on peut alors démontrer que, pour garantir la confidentialité différentielle, la réponse R devra être une fonction aléatoire de la base de données. Autrement dit, une même requête différentiellement confidentielle lancée deux fois de suite sur une même base de données aura nécessairement une probabilité non nulle de retourner deux résultats différents.

Le mécanisme laplacien

La confidentialité différentielle peut sembler bien encombrante. Cependant, dans certaines applications, elle peut ne pas être gênante. Typiquement, le mécanisme dit *laplacien* nous permet d'effectuer des sondages différentiellement confidentiels tout en fournissant une réponse parfaitement convenable.

Par exemple, imaginez que vous soyez un hôpital avec des données sur des patients. Vous souhaitez déterminer la proportion de cancers du poumon sans compromettre la confidentialité des patients. Pour ce faire, au lieu de révéler le nombre exact de patients avec ce cancer dans votre base de données, perturbez ce nombre en rajoutant une quantité tirée au hasard selon la loi de Laplace, et révélez le résultat randomisé.

Je ne vais pas détailler la loi de Laplace⁸. Mais sachez qu'elle dépend d'un paramètre qui va être l'intensité typique de la perturbation du résultat. Pour une confidentialité différentielle $(\varepsilon, \delta = 0)$, cette perturbation doit être de l'ordre de $1/\varepsilon$. Ainsi, si un mécanisme laplacien dit qu'il y a 243 cas de cancers du poumon, vous savez que ce nombre n'est pas le vrai nombre. Le vrai nombre est quelque chose comme $243 \pm 1/\varepsilon$.

Ceci peut paraître insatisfaisant. Cependant, il faut contraster cette incertitude avec les fluctuations statistiques de tout sondage. Supposons que l'on sonde 500 personnes au hasard. Cet échantillon ne sera qu'approximativement représentatif. En fait, s'il y a une fraction $n/500$ de cancers du poumon dans la population, alors, pour chaque sondage de 500 personnes, on s'attend à en obtenir environ⁹ $n \pm \sqrt{n}$. En particulier, l'incertitude concernant le sondage est de l'ordre de \sqrt{n} .

⁸Sa fonction de densité est $f(x) = \frac{1}{2b} \exp(-b|x|)$, dont la variance est $2b^2$.

⁹Ceci se déduit du théorème centrale-limite ou, pour les puristes, des inégalités de concentration comme les inégalités de Chernoff.

Ainsi, si on s'attend à environ n cas, on peut ajuster l'incertitude nécessaire à la confidentialité différentielle à la valeur de n , en posant $1/\varepsilon = \sqrt{n}$, ou, dit autrement, $\varepsilon = 1/\sqrt{n}$. Ce faisant, on garantit une $(1/\sqrt{n}, 0)$ -confidentialité différentielle, essentiellement sans détériorer la précision du sondage. En particulier, en sondant un très grand nombre de personnes, la confidentialité du sondage est alors quasi totale.

Robustesse à la composition

Au cours de la dernière décennie, la confidentialité (ε, δ) -différentielle est devenue l'un des concepts les plus étudiés et les plus excitants de la science informatique. Au-delà de la pertinence intuitive du concept, on peut retracer cette popularité grandissante à deux propriétés fondamentales des mécanismes différentiellement confidentiels : la robustesse à la composition avec des calculs ultérieurs et l'addition des pertes de confidentialité successives.

Commençons par la robustesse à la composition. On l'a vu, la défaillance d'une simple pseudonymisation des données était la possibilité de recouper les métadonnées, c'est-à-dire les informations associées aux données comme l'âge, le sexe ou l'adresse, avec d'autres jeux de données pour dé-anonymiser ces données. De telles techniques ont par exemple été utilisées pour retracer les propriétaires des comptes Bitcoin malveillants, dont les adresses étaient des pseudonymes.

En effet, Bitcoin est une technologie moderne qui permet les transactions financières sans intervention d'autorité centrale. Contrairement à des virements classiques qui requièrent l'accord de banques, n'importe quel ordinateur du web peut valider des transactions Bitcoin annoncées par les comptes Bitcoin¹⁰. La possibilité d'effectuer rapidement de telles transactions décentralisées, pseudonymisées et sans contrôle des autorités étatiques a d'ailleurs longtemps profité aux trafiquants d'armes ou de drogues. Dès lors, la dé-anonymisation de leurs comptes devint une question de sécurité nationale. Dans de nombreux cas, elle fut réalisée avec succès. L'astuce pour ce faire était de constamment croiser les métadonnées des transactions Bitcoin (comme l'heure de la transaction et son destinataire) avec d'autres jeux de données.

Ce qu'il faut retenir de ceci, c'est que la pseudonymisation des données, même intelligemment pensée comme cela a sans doute été le cas par les trafiquants d'armes et de drogues, n'est absolument pas une garantie de confidentialité. En particulier, si cette pseudonymisation peut paraître suffisante au moment de la publication des données, elle n'a aucune garantie de robustesse vis-à-vis de la composition de ces données avec d'autres informations.

¹⁰En fait, pour avoir le pouvoir de valider ces transactions, il faut et il suffit de résoudre un problème de maths très difficile, ce qui fait que, de nos jours, ce sont des fermes de grilles de calculs qui prennent soin de valider ces transactions. Ces grilles de calculs sont directement rémunérées en Bitcoin pour leur travail.

Cette faiblesse de la pseudonymisation est la force de la confidentialité différentielle. Que ce soit au moment de la publication des résultats ou des siècles plus tard après avoir croisé les résultats de cette publication avec divers autres jeux de données, la garantie de confidentialité des données auxquelles le mécanisme différentiellement confidentiel a été appliqué demeurera la même.

Cette garantie est préservée même dans des cas extrêmes. Imaginez qu’Alice ait accepté de donner ses données à une étude, mais que tous les autres individus de l’étude sont en fait fictifs et bien connus. Malgré le fait que toute la base de données sauf les données d’Alice soit connue, pourvu que le seul accès à cette base de données ait été (ε, δ) -difféntiellement confidentiel, les données d’Alice demeureront (ε, δ) -difféntiellement confidentielles à jamais !

Avec probabilité $1 - \delta$, la crédence sur Alice ne sera jamais multipliée par plus de e^ε , entre le cas où Alice est dans la base de données, et le cas où elle n’y est pas. Que ce soient juste après la publication des résultat du mécanisme, ou des siècles plus tard, suite à un croisement avec d’autres jeux de données.

L’additivité des pertes de confidentialité

Bien entendu, cette remarque n’est valable que si la base de données en question n’a été sondée que par le mécanisme (ε, δ) -difféntiellement confidentiel en question. En pratique, on pourrait vouloir sonder la base de données plusieurs fois, à l’aide de divers mécanismes (ε, δ) -difféntiellement confidentiels.

L’autre propriété fondamentale de la confidentialité différentielle est le fait que les pertes de confidentialité différentielle vont alors s’additionner. Autrement dit, si on applique un premier mécanisme $(\varepsilon_1, \delta_1)$ -difféntiellement confidentiel, suivi d’un second mécanisme $(\varepsilon_2, \delta_2)$ -difféntiellement confidentiel, alors la perte de confidentialité différentielle totale sera au plus $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$.

Ce théorème remarquable se comprend bien à l’aide d’un raisonnement approximatif en termes de crédence¹¹. À la suite de l’application du premier mécanisme, avec probabilité $1 - \delta_1$, la crédence sur un individu ne sera pas multipliée par plus que e^{ε_1} par rapport au cas où il n’est pas sondé. À la suite de l’application du second mécanisme, avec probabilité maintenant $(1 - \delta_1)(1 - \delta_2)$, la crédence ne sera alors pas multipliée par plus que $e^{\varepsilon_1} e^{\varepsilon_2}$. Or, un petit calcul algébrique nous dit que $(1 - \delta_1)(1 - \delta_2) \geq 1 - (\delta_1 + \delta_2)$ et $e^{\varepsilon_1} e^{\varepsilon_2} = e^{\varepsilon_1 + \varepsilon_2}$. En combinant tout ceci, on en déduit qu’après application successive des deux mécanismes, avec probabilité au moins $1 - (\delta_1 + \delta_2)$, la crédence sur l’individu ne sera jamais multipliée par plus de $e^{\varepsilon_1 + \varepsilon_2}$. Voilà qui correspond exactement à dire que la confidentialité différentielle de l’application successive des deux mécanismes est au pire $(\varepsilon_1 + \varepsilon_2, \delta_1 + \delta_2)$. La succession de mécanismes conduit à un cumul des pertes de confidentialité différentielle !

¹¹  *Interpretation of ϵ and δ 's of Differential Privacy (Proof)* | Wandida | L.N. Hoang (2017)

Cette bonne nouvelle révèle également une difficulté majeure de la confidentialité. Plus on fait de requêtes à la base de données, plus on viole la confidentialité des données qui s'y trouvent. En fait, quand on conçoit un système confidentiel, il est important de contrôler tout son cycle de vie, et d'anticiper sa suppression totale une fois la confidentialité différentielle critique atteinte. Anticiper tout le cycle de vie d'un système confidentiel peut aussi permettre d'optimiser les réponses aux requêtes. En effet, à quantités d'information utile égales, un système confidentiel qui répond à des requêtes les unes après les autres sans anticiper les futures requêtes sera moins différentiellement confidentiel qu'un système confidentiel qui optimise l'ensemble des réponses aux questions.

Dans le premier cas, on parle de requêtes *online* (ou en temps réel). Dans le second cas, on parle de requêtes *offline*. Certaines applications pourraient alors permettre d'adresser directement des requêtes *offline* et pourraient alors supprimer la base de données une fois les réponses *offline* déterminées. Cependant, en pratique, le concepteur du système confidentiel est souvent contraint de se contenter de requêtes *online*, puisqu'au moment de la conception du système confidentiel, les requêtes ne sont en général pas connues. Ceci sera notamment le cas lorsque les requêtes sont celles des utilisateurs, et pas du concepteur du système.

Comme vous l'imaginez, il y aurait beaucoup plus à dire sur la confidentialité différentielle. Le concept n'a été inventé qu'en 2006, et la recherche dans ce domaine est encore très active.

En pratique, ça ne va pas !

Cependant, de l'idée théorique à son application pratique, les obstacles sont multiples et complexes, et il faudra sans doute encore du temps avant que les bases de données des hôpitaux et des institutions soient différentiellement confidentielles. Même si certains s'y mettent¹². En particulier, il faudra que les concepteurs et les utilisateurs de ces systèmes comprennent beaucoup mieux la confidentialité en général, et les limites des techniques de pseudonymisation classiquement utilisées.

Ceci étant dit, il n'est pas encore clair que la confidentialité différentielle soit le concept idéal pour garantir la confidentialité des données. La confidentialité différentielle pourrait ainsi être trop contraignante dans certains cas. En particulier, elle doit assurer la confidentialité quelles que soient les données dans la base de données, et quelles que soient les créances *a priori* des hackers qui chercheraient à casser la confidentialité de la base de données. C'est beaucoup lui demander ! Qui plus est, il faut imaginer que votre perte de confidentialité sur l'ensemble de votre vie sera la somme des pertes de confidentialité de tous les mécanismes auxquels vous aurez pris part.

¹²  *The Big Data Setup of the Human Brain Project* | ZettaBytes | A. Ailamaki (2017)

Une approche plus radicale pour garantir notre confidentialité est d’interdire les agrégations d’information et de faire de chacun le seul et unique possesseur de ces données. Il y a 25 ans, avant l’avènement du web, ceci aurait pu être imaginable, puisque les données de chacun étaient alors physiquement séparées des données des autres. À l’époque, chacun avait son ordinateur (ou ses disquettes), et nos données digitales étaient physiquement restreintes à rester dans nos logements (quoiqu’il restait le problème des données imprimées et conservées dans des institutions publiques...).

Cependant, de nos jours, nos données privées voyagent de serveurs à serveurs à travers l’Internet mondial. Aucun d’entre nous n’a même idée de la localisation géographique de nos données personnelles. Pire encore, beaucoup de ces données sont inéluctablement stockées dans les immenses centres de données que possèdent une poignée de géants du web, comme Google, Apple, Facebook et Amazon. Même la plupart des traitements de ces données, de votre profil Facebook à la réservation de vos billets d’avion en passant par la retouche de vos photos de vacances, se fait désormais dans ces centres de données. À tel point que Facebook est désormais probablement plus à même de prédire les prochains articles et vidéos que vous aimerez que ne le serait votre conjoint(e).

Dans cette guerre des données, les grandes entreprises de la Silicon Valley ont clairement plusieurs longueurs d’avance.

Le chiffrement homomorphe

Cependant, tout n’est pas encore perdu. En particulier, le chiffrement homomorphe pourrait bientôt révolutionner notre relation aux données et aux centres de calcul. Le principe de ce chiffrement est de déléguer le traitement de nos données privées aux centres de données, sans que ces centres de données soient capables de lire ou comprendre les données qu’ils sont en train de traiter.

Le chiffrement homomorphe a ainsi déjà été appliqué pour sécuriser et garantir la confidentialité de votes électroniques, avec des prototypes disponibles en ligne comme *Helios* et *Belenios*. Le principe est grossièrement le suivant. Chaque électeur possède une clé privée. À l’aide de cette clé privée, il peut coder et signer son vote, mais personne ne peut le décoder. Les votes encodés sont ensuite combinés via une opération publiquement vérifiable, et fournissent un résultat final encodé. Les clés privées des électeurs sont ensuite combinées pour former une sorte de super-clé, avec laquelle le résultat final encodé, et seulement ce résultat final encodé, peut être décodé. Ce faisant, on peut être mathématiquement assuré de la validité du résultat final, sans compromettre la confidentialité de chacun des votes des électeurs.

Si en principe, ces astuces de calcul et de cryptographie semblent avoir résolu le problème de la conception du vote électronique, le vote électronique ne garantit pas tout à fait toutes les bonnes propriétés du vote classique dans l’isoloir.

En particulier, dans l'isoloir, l'électeur est seul et ne peut être surveillé par personne. À l'inverse, si l'électeur vote sur son téléphone, il pourrait être alors sous la menace d'une personne malveillante qui le forceraient alors à voter d'une certaine manière. Pire encore, il demeurera un risque que l'appareil de l'électeur se fasse attaquer par un virus ou par un hacker, qui pourra alors faire croire à l'électeur qu'il a voté d'une manière, alors que le virus ou hacker le fera en fait voter d'une autre manière.

Neanmoins, il me semble qu'il ne s'agit pas de savoir si ce vote électronique chiffré est parfait ; il s'agit de déterminer s'il est meilleur que le vote pratiqué aujourd'hui — lequel a de nombreux défauts, par exemple celui d'être chronophage. Mais je ne compte pas disserter davantage sur cette question. D'autant qu'il possède une dimension morale qui sort du cadre de la philosophie du savoir.

Revenons-en au chiffrement homomorphe. Dans le cadre du vote électronique, ce chiffrement permet de combiner des votes encodés en un résultat final, encodé certes, mais qui reflète malgré tout parfaitement les votes des électeurs. C'est là toute la magie du chiffrement homomorphe : l'ordinateur vient d'effectuer une manipulation des données qui donne un résultat parfaitement juste, mais l'ordinateur en question est incapable de savoir quelle était la nature de ces données !

D'un point de vue mathématique, le chiffrement homomorphe utilisé par le vote électronique est relativement simple : ajouter les voix pour et les voix contre. Il ne s'agit que d'additions. Cependant, le Saint-Graal de la recherche moderne sur le chiffrement homomorphe serait de permettre des opérations beaucoup plus sophistiquées que l'addition. En fait, l'idéal serait de permettre à un ordinateur d'appliquer n'importe quel algorithme sur des données chiffrées, sans jamais avoir à les déchiffrer. Dès lors, de chez vous, sur votre téléphone ou sur votre ordinateur, vous pourriez demander à un centre de données à l'autre bout du monde, de manipuler vos données chiffrées, de calculer un résultat à partir de ces données, et de vous envoyer uniquement ce résultat. Vous déchiffreriez alors les données avec votre téléphone, sur lequel se trouveraient vos codes secrets. Vous pourriez alors lire vos emails, regarder vos photos de vacances ou écouter vos musiques, sans que le centre de données, ni qui que ce soit, sache quelles étaient les données que vous aviez téléchargées !

Ce chiffrement homomorphe existe en fait déjà. Malheureusement, il est encore trop inefficace. Pour effectuer les opérations exigées, les centres de données devraient effectuer toutes sortes de pirouettes complexes et variées. Avec les algorithmes de chiffrement homomorphe d'aujourd'hui, ces centres de données devraient dépenser énormément plus de temps, de place mémoire et d'énergie électrique, que si les mêmes opérations étaient effectuées avec des données déchiffrées. Cependant, la recherche avance vite...

Références en français

- ▶ *Le principe du chiffrement par clefs asymétriques* | Wandida | E.M. El Mhamdi (2014)
- ▶ *L'arithmétique utilisée par le chiffrement par clefs asymétriques* | Wandida | E.M. El Mhamdi (2014)
- ▶ *Les codes secrets* | Science Étonnante | D. Louapre (2015)
- ▶ *Le décryptage d'Enigma* | Science4All | R. Barbulescu et L.N. Hoang (2017)
- ▶ *Il donne du cannabis à son chat, ça tourne mal* | La statistique expliquée à mon chat (2017)
- ▶ *10 prouesses de la cryptographie* | Crypto | String Theory | L.N. Hoang (2018)

Références en anglais

- ☒ *Differential privacy* | Automata, languages and programming | C. Dwork (2006)
- ☒ *Differential privacy* | Encyclopedia of Cryptography and Security | Springer US | C. Dwork (2011)
- ➲ *The algorithmic foundations of differential privacy* | Foundations and Trends® in Theoretical Computer Science | C. Dwork and A. Roth (2014)

- ▶ *An Embarrassing Survey - Randomized Response* | Singingbanana | J. Grime (2010)
- ▶ *Mathematics of Codes and Code-Breaking* | Singingbanana | J. Grime (2012)
- Maths from the talk "Alan Turing and the Enigma Machine"* | Singingbanana | J. Grime (2013)
- ▶ *Diffie-Hellman Key Exchange* | Wandida | J. Goubault-Larrecq (2014)
- ▶ *The Diffie-Hellman Protocol* | ZettaBytes | S. Vaudenay (2016)
- ▶ *The Big Data Setup of the Human Brain Project* | ZettaBytes | A. Ailamaki (2017)

- ▶ *Differential Privacy* | Playlist | Wandida | L.N. Hoang (2017)
- ▶ *What is Privacy?* | Wandida | L.N. Hoang (2017)
- ▶ *The Formal Definition of Differential Privacy* | Wandida | L.N. Hoang (2017)
- ▶ *A Simple Differentially-Private Randomized Survey* | Wandida | L.N. Hoang (2017)
- ▶ *Interpretation of ϵ and δ 's of Differential Privacy* | Wandida | L.N. Hoang (2017)
- ▶ *Interpretation of ϵ and δ 's of Differential Privacy (Proof)* | Wandida | L.N. Hoang (2017)

Un pari est une taxe sur les bullshit.

Alex Tabarrok (1966-)

Ce que le jeu est, définit ce que les joueurs font. Notre problème aujourd’hui n'est pas juste le fait que les gens perdent confiance, c'est le fait que notre environnement agit contre l'évolution de la confiance.

Nicky Case



Les jeux sont faits

La magouilleuse

L'École Polytechnique est l'une des *Grandes Écoles* d'ingénieur et de science les plus prestigieuses de France. Fondée en 1794, elle fut ensuite militarisée en 1804 par Napoléon I^e, qui y voyait un pôle de recrutement intéressant pour la direction de ses armées. De nos jours, l'École Polytechnique est encore sous la tutelle du ministère des Armées. C'est pourquoi, à leur entrée à l'École Polytechnique, tous les étudiants français doivent suivre une formation militaire initiale de trois semaines. Après ces trois semaines, vient alors le moment des affectations.

Il y a 400 élèves. De façon grossière, 130 doivent être affectés à l'armée de terre, 60 à la marine, 60 à l'armée de l'air, 60 à la gendarmerie, tandis que les 90 élèves restants doivent se répartir entre d'autres formations militaires, la police, les pompiers ou les associations humanitaires civiles. Plutôt que de déterminer des affectations aléatoires, la direction de l'École, ayant de futurs ingénieurs en son sein, a eu la brillante idée de laisser les étudiants déterminer d'eux-mêmes comment affecter leurs camarades. C'est ainsi qu'un logiciel, appelé *magouilleuse*, a été développé. Chaque étudiant rentre alors ses aptitudes médicales et ses préférences d'affectation dans le logiciel. Le logiciel secoue tout ça, applique un algorithme obscur, et détermine les affectations des étudiants.

Étudier les propriétés de tels logiciels est un sujet passionnant qui aura occupé mes pensées pendant de nombreuses années. En fait, j'ai trouvé ce problème si

intrigant que j'en ai fait le sujet de ma thèse. Malheureusement, après des années de recherche, si elle m'a en effet permis de mieux comprendre le problème de la *magouilleuse*, ma thèse n'aura absolument pas clos le problème en question.

Une difficulté consiste simplement à déterminer l'objectif de la *magouilleuse*. La *magouilleuse* telle qu'implémentée à l'époque où j'en ai été victime est, si j'en crois ce qu'on m'en avait dit, une minimisation de pertes quadratiques. Autrement dit, la *magouilleuse* attribuait une pénalité de 1 point lorsqu'un premier choix était assigné à un élève, 4 si un second choix lui était assigné, 9 si c'était un troisième choix, et plus généralement, une pénalité de n^2 pour le n -ième choix d'un élève. La *magouilleuse* va alors minimiser la somme des pénalités. Il s'agit d'un choix finalement assez arbitraire et discutable — et j'ai notamment longuement discuté de cela dans ma thèse¹.

Mais laissons cette difficulté majeure et non-triviale de côté. Ce sur quoi je veux insister pour l'instant est un secret de polichinelle qui devint un sujet de conversation récurrent entre élèves polytechniciens. Pour bien s'en tirer, il ne fallait pas révéler ses vraies préférences. L'astuce était de mettre son affectation préférée en premier choix, et d'insérer juste après des affectations très prisées comme la marine et la gendarmerie. En effet, ces affectations étant très prisées, elles seront uniquement assignées à des premiers choix. Du coup, pour la *magouilleuse*, l'alternative au premier choix est un quatrième ou cinquième choix, dont la pénalité est très importante. Ce faisant, la *magouilleuse* favorisera ceux qui ont suivi cette stratégie au profit de ceux qui se sont contentés de rentrer leurs vraies préférences. La *magouilleuse* défavorise l'honnêteté.

Ce constat dérangeant ne se limite malheureusement pas à la *magouilleuse*. Les périodes pré-électorales conduisent bien souvent à d'éternels débats entre le cœur et la raison. Lors de l'élection présidentielle de 2002 en France, de nombreux candidats de gauche s'étaient présentés, et ont recueilli un très grand nombre de voix. Il y eut une dispersion des voix à gauche, qui fut fatale à Lionel Jospin, le grand candidat de gauche. Jospin fut éliminé du premier tour du scrutin uninominal à deux tours. Le deuxième tour serré qui s'annonçait entre Chirac et Jospin devint un deuxième tour sans appel entre Chirac et Le Pen. Chirac fut élu avec une légitimité bluffante — mais trompeuse. Et les millions de Français de gauche qui n'ont pas voté Jospin ont regretté ne pas avoir voté « utile ».

Les incitations engendrées par un mécanisme de décision comme la *magouilleuse* ou le scrutin uninominal à deux tours sont l'objet d'étude de la théorie de la conception de mécanismes en particulier. Le Saint-Graal de cette théorie est de déterminer des règles d'interaction entre des agents (par exemple les élèves de l'École Polytechnique ou les électeurs) qui incitent les agents à révéler leurs vraies préférences (ou plus généralement, à agir de manière éthique) et qui conduisent à des résultats relativement souhaitables (assigner équitablement les élèves aux affectations ou élire un candidat qui représente bien la volonté du

¹  *Measuring Unfairness Feeling in Allocation Problems* | Omega | L.N. Hoang, F. Soumis et G. Zaccour (2016)

peuple). Comme on le verra, la philosophie bayésienne a beaucoup apporté à cette théorie — et a été récompensée de nombreux prix Nobel.

Mais avant d'en arriver là, parlons de la théorie des jeux sur laquelle repose la théorie de la conception des mécanismes. Et pour cela, partons au Royaume-Uni.

Split or Steal

C'est la fin du jeu télévisé *Golden Balls*. Sarah et Steven se battent pour une cagnotte de 100 150 £ dans la dernière étape du jeu, appelée *Split or Steal*. Les deux candidats sont face à face. Chacun a deux boules. L'une de ses boules est *Split* (partage), l'autre est *Steal* (vol). Chaque candidat doit choisir une des deux boules. S'ils choisissent tous deux *Split*, alors ils partageront la cagnotte. Si l'un choisit *Split* mais l'autre choisit *Steal*, celui qui aura choisi *Steal* volera toute la cagnotte. Enfin, si les deux joueurs choisissent *Steal*, alors tous les deux rentreront chez eux les poches vides.

Avant que chacun ne choisisse sa boule, les deux candidats ont une trentaine de secondes pour discuter. Sarah supplie alors Steven de partager la cagnotte. Elle est au bord des larmes. Steven tente de rassurer Sarah, et lui promet qu'il partagera. La discussion s'achève. Chaque candidat choisit sa boule en secret. Le suspense est à son comble. Le présentateur demande alors aux candidats de révéler leurs choix. Stupeur générale. Si Steven a choisi *Split*, Sarah a joué *Steal*. Sarah a volé la cagnotte ! Steven est atterré. Abattu. Sarah est gênée, et ne sait plus où regarder. Mais c'est bien elle qui a gagné 100 150 £ !

Dans le jeu *Split or Steal* comme dans notre quotidien, l'incertitude est omniprésente. Nous dépendons fortement des décisions des uns et des autres, et notre pouvoir d'influence est très limité. Dans ce contexte, plutôt que de forcer les autres à agir de telle ou telle manière, il est souvent plus raisonnable d'anticiper la manière dont ils vont agir et s'y adapter — même si le militarisme a aussi ses effets. La difficulté conceptuelle qui survient alors, c'est que la manière dont les autres vont agir dépend de la manière dont on va agir, dont on vient de dire qu'elle dépend de la manière dont ils vont agir, laquelle dépend de la manière dont on va agir... et ainsi de suite à l'infini.

En 1951, dans une thèse de doctorat de seulement 28 pages et ne contenant que 2 citations, le futur prix Nobel d'économie John Nash propose la notion d'équilibre pour abréger ce raisonnement infini. Un équilibre de Nash est alors une situation où tout joueur a une réaction optimale vis-à-vis des réactions des autres. Ainsi, sachant ce que font les autres, chaque joueur a tout intérêt à persister dans sa stratégie. C'est d'ailleurs en cela que l'équilibre de Nash est un équilibre : une fois que les joueurs le jouent, on s'attend à ce qu'ils persistent dans cet état d'équilibre.

Étrangement, en première approximation, les stratégies de Sarah et Steven forment un équilibre de Nash. En effet, sachant que Steven choisit *Split*, Sarah a tout intérêt à jouer *Steal*, puisqu'elle double là ses gains. De son côté, si Steven sait que Sarah va jouer *Steal*, il n'a rien à gagner en passant de *Split* à *Steal*. En effet, dans les deux cas, Steven rentre chez lui les poches vides.

On a (presque) là un cas de dilemme du prisonnier. Ce dilemme fut imaginé par Merrill Flood et Melvin Dresher en 1950, puis formalisé par Albert Tucker. Dans ce dilemme, deux complices sont arrêtés par la police qui les interroge séparément. Si un complice se fait dénoncer par son partenaire, sa peine se voit augmentée. Cependant, la police promet une réduction de peine à chaque complice s'il dénonce son partenaire, que son partenaire le dénonce ou non. Ainsi, en dénonçant son partenaire, chaque complice diminue sa peine, peu importe ce que fait son partenaire. Par conséquent, dénoncer son partenaire est un équilibre de Nash. C'est même le seul équilibre de Nash.

Mais alors, les deux complices se voient dénoncés l'un par l'autre. Ils écoperont alors de lourdes peines, qu'ils auraient pu éviter en restant tous deux silencieux. La morale de cette histoire est que les incitatifs individuels ne sont pas toujours alignés avec l'intérêt du groupe. En première approximation, voler la cagnotte du *Split or Steal* est analogue à dénoncer son complice au dilemme du prisonnier. Dans les deux cas, il s'agit d'une stratégie individuellement optimale qui conduit à une sous-optimalité globale.

La persuasion bayésienne

Cependant, ces exemples sont certainement bien trop simplifiés pour vraiment représenter le processus intellectuel que suivent Steven, Sarah et les deux complices du dilemme du prisonnier. En effet, au-delà des gains financiers et de la dureté de la peine, il y a un énorme coût psychologique à être l'auteur d'une délation en public. En particulier, il n'est pas insensé d'imaginer que pour Steven, il est préférable de jouer *Split* plutôt que de choisir *Steal*. Gagner toute la cagnotte et devoir subir le regard désapprobateur de l'autre candidat, des spectateurs et de sa famille peut être un lourd fardeau à porter. Dès lors, il se pourrait que jouer *Split* est en fait une stratégie optimale pour Steven, peu importe ce que choisit Sarah.

Là où les choses deviennent intéressantes, c'est si Steven préfère voir Sarah perdre la cagnotte à partir du moment où il sait qu'elle joue *Steal*. Autrement dit, imaginons maintenant que Steven préfère, dans l'ordre, les issues suivantes : $(\text{Split}, \text{Split})$, $(\text{Steal}, \text{Steal})$, $(\text{Steal}, \text{Split})$, $(\text{Split}, \text{Steal})$. La première action de la paire est le choix de Steven, la seconde est celui de Sarah. En particulier, désormais, la stratégie optimale de Steven dépend de ce que joue Sarah. Si Sarah choisit *Split*, Steven préférera jouer *Split*. Mais si Sarah choisit *Steal*, Steven voudra alors jouer *Steal*.

C'est là que la discussion préalable entre Steven et Sarah — et la philosophie bayésienne — jouera un rôle très important. Cette discussion peut affecter les crédences de Steven en le choix à venir de Sarah. Visiblement, dans notre cas, Sarah a réussi à faire croire qu'elle jouera *Split*. Elle a bien fait. Steven a rapidement été convaincu que cela serait le cas, si bien qu'il a négligé l'alternative. D'où sa très grande déception.

La communication peut d'ailleurs elle aussi être formalisée par une approche bayésienne. Après tout, communiquer, c'est révéler une information, que l'autre pourra utiliser pour mettre à jour ses crédences. Ainsi, en 2011, les économistes Kamenica et Gentzkow se sont demandés comment un procureur pouvait incriminer au mieux un accusé devant un juge bayésien. De façon étonnante, ils ont montré qu'un bon procureur pouvait alors convaincre le juge de condamner plus d'accusés que le nombre d'accusés que le juge pense coupables !

Détaillons. En bon bayésien, ce juge a nécessairement un préjugé sur la culpabilité de l'accusé. Supposons que ce préjugé soit $\mathbb{P}[\text{🔴}] = 0,3$. On va supposer, de plus, que le juge bayésien ne condamnera l'accusé que si sa crédence en la culpabilité de l'accusé est supérieure ou égale à sa crédence en l'innocence de l'accusé. Pour convaincre le juge, le procureur va proposer de lancer une investigation un peu particulière. Lorsque l'accusé est coupable, elle va le démontrer. Mais lorsque l'accusé est innocent, l'investigation va parfois se tromper. 3 fois sur 7, l'investigation va incriminer un accusé pourtant innocent. Le procureur le sait. Le juge aussi.

Bien entendu, du coup, si l'investigation incrimine l'accusé, c'est peut-être uniquement parce que l'investigation s'est trompée. Néanmoins, ceci ne peut qu'augmenter les suspicions du juge en la culpabilité de l'accusé. Un calcul bayésien permet de déterminer précisément cette suspicion *a posteriori* :

$$\mathbb{P}[\text{🔴}|\text{🔴}] = \frac{\mathbb{P}[\text{🔴}|\text{🔴}]\mathbb{P}[\text{🔴}]}{\mathbb{P}[\text{🔴}|\text{🔴}]\mathbb{P}[\text{🔴}] + \mathbb{P}[\text{🔴}|\text{🟡}]\mathbb{P}[\text{🟡}]} = \frac{1 \cdot 0,3}{1 \cdot 0,3 + 3/7 \cdot 0,7} = 0,5.$$

Autrement dit, *a posteriori*, une fois qu'il apprend que l'investigation incrimine l'accusé, le juge bayésien aura autant de crédence en la culpabilité de l'accusé qu'en son innocence : il décidera donc de le condamner². Ainsi, tout accusé incriminé par l'investigation est condamné. Or, la probabilité d'être incriminé par l'investigation s'obtient par la loi des probabilités totales :

$$\mathbb{P}[\text{🔴}] = \mathbb{P}[\text{🔴}|\text{🔴}]\mathbb{P}[\text{🔴}] + \mathbb{P}[\text{🔴}|\text{🟡}]\mathbb{P}[\text{🟡}] = 1 \cdot 0,3 + 3/7 \cdot 0,7 = 0,6.$$

Autrement dit, *a priori*, le juge aura désormais une probabilité de 60 % de condamner un accusé, quand bien même sa crédence *a priori* en la culpabilité de l'accusé est de 30 %. Le juge condamnera nécessairement trop d'accusés !

²En remplaçant 3/7 par 3/7 – ε, on obtient essentiellement le même résultat.

Que l'on soit clair, toutefois, ceci n'est absolument pas une déficience du bayésianisme du juge. S'il n'avait pas utilisé la formule de Bayes et s'il s'était restreint à son *a priori*, le juge aurait acquitté tous les accusés. Il aurait donc connu le même taux d'erreurs. L'objectif du juge n'est pas de faire coïncider le nombre de condamnés avec le nombre de coupables. Son objectif est de minimiser son taux d'erreurs³.

Mais alors, serait-il possible de piéger un juge bayésien et d'augmenter son taux d'erreurs ? La réponse est non. En effet, l'inférence bayésienne possède une propriété remarquable qui la distingue de nombreuses autres formes d'induction : pourvu que le juge interprète correctement toute information additionnelle et applique la formule de Bayes, son taux d'erreurs espéré ne pourra pas avoir diminué⁴. Autrement dit, en espérance, le bayésien y gagne toujours à acquérir plus d'information⁵.

Les points de Schelling

Revenons-en au *Split or Steal*. Le jeu devient encore plus intéressant si l'on imagine maintenant que les préférences de Sarah sont en fait les mêmes que celles de Steven. Dans ce cas, tout deux ont envie de partager. Il suffirait en fait qu'ils se persuadent mutuellement qu'ils partageront pour que tout se passe très bien.

Cependant, la moindre suspicion peut rapidement dégénérer. Si, suite à une maladresse, Steven donne l'impression qu'il pourrait ne pas jouer *Split*, Sarah pourrait alors augmenter sa crédence en un choix *Steal* de Steven. Il pourrait alors survenir un point où cette crédence est si grande, que Sarah pourrait préférer jouer *Steal* que *Split*, de peur de subir l'humiliation de voir Steven raffler la mise sous ses yeux. Mais si jamais elle laissait transparaître ses hésitations, Steven pourrait alors anticiper la stratégie *Steal* de Sarah, et ainsi se conforter à jouer *Steal* lui aussi.

Il s'agit là d'un problème dit de *coordination*. On a là deux équilibres de Nash symétriques⁶. Dans l'un de ces équilibres, Sarah et Steven jouent tous les deux

³Notez aussi que l'on s'est placé ici dans le cas où condamner un innocent ou acquitter un coupable avait le même coût pour le juge. On peut aisément modifier le problème pour adresser des philosophies morales différentes de la condamnation juste sous incertitude.

⁴Plus généralement, considérons un bayésien avec une fonction d'utilité u conforme aux axiomes de von Neumann et Morgenstern, qui doit prendre une décision a , sans connaître une variable x . Sans information additionnelle y , le bayésien choisit $\sup_a \mathbb{E}_x[u(a, x)]$. Supposons qu'il apprend y . Il maximise alors $\sup_a \mathbb{E}_x[u(a, x)|y]$. Son gain espéré *a priori* est alors $\mathbb{E}_y [\sup_a \mathbb{E}_x[u(a, x)|y]]$. Pourvu que $\mathbb{P}[x|y]$ soit calculé via la formule de Bayes, le bayésien y gagne alors à apprendre y , i.e. $\mathbb{E}_y [\sup_a \mathbb{E}_x[u(a, x)|y]] \geq \sup_a \mathbb{E}_x[u(a, x)]$.

⁵Bien sûr, la notion d'espérance est fondamentale, puisque le juge est maintenant amené à condamner des innocents qu'il aurait acquittés sans l'investigation.

⁶En fait il y a même un troisième équilibre de Nash où Sarah et Steven jouent tous les deux au hasard *Split* ou *Steal*.

Split et repartent heureux. Dans l'autre, ils jouent tous les deux *Steal* et repartent tristes. Mais s'ils se ne se coordonnent pas, tout deux repartiront encore plus tristes, soit parce qu'ils auront vu l'autre candidat leur raffler la mise, soit parce que le public les aura fusillés d'un regard désapprobateur. Autrement dit, l'issue du jeu sera là déterminée par leurs crédences respectives en ce que l'autre fera — et des crédences erronées peuvent être catastrophiques pour tous les deux !

Ces problèmes de coordination transcendent bien sûr très largement le cadre du *Split or Steal*. Il y a bien sûr le cas hollywoodien de deux amoureux qui hésitent à révéler leurs flammes respectives, parce qu'ils doutent de l'intérêt que l'autre porte pour eux. De mauvaises crédences peuvent alors ruiner ce qui aurait été une belle histoire d'amour — et l'apprentissage rocambolesque des bonnes crédences fait souvent tout l'intérêt du film !

C'est pour résoudre ces problèmes de coordination que nous disposons de traditions, de conventions ou de protocoles. Ces éléments qui structurent nos relations sociales sont ce que l'on appelle les *points de Schelling*, du nom de Thomas Schelling, prix Nobel d'économie de 2005. En termes bayésiens, ces points de Schelling servent à établir des crédences *a priori* sur les comportements des uns et des autres dans une société.

Faire preuve de bon sens semble souvent requérir une bonne estimation de ces points de Schelling, qui prennent alors un rôle important dans le comportement des individus. Ainsi, on peut imaginer que dans une société où la confiance et l'honnêteté sont des points de Schelling fiables, Steven et Sarah finiraient toujours par jouer *Split*. À l'inverse, dans des sociétés où la norme n'est pas la confiance mutuelle et où la méfiance est conseillée, Steven et Sarah auraient peut-être plus tendance à jouer *Steal*.

L'équilibre mixte

Il y a une dernière variante du *Split or Steal* qui va nous permettre d'explorer d'autres subtilités des interactions entre individus. Supposons maintenant que Steven et Sarah sont vénaux, et promettent un pourboire de 10 000 £ à l'autre si l'un gagne toute la mise. Du coup, leur ordre de préférence est désormais : (*Steal, Split*) avec un gain de 90 k£, (*Split, Split*) avec un gain de 50 k£, (*Split, Steal*) avec un gain de 10 k£, et (*Steal, Steal*) avec aucun gain.

De façon étrange, si Sarah annonce qu'elle jouera *Steal*, Steven aura tout intérêt à jouer *Split*, ce qui confortera davantage Sarah dans son choix. À l'inverse, si Steven annonce jouer *Steal*, les rôles pourraient alors s'inverser. Ce jeu possède donc deux équilibres de Nash antisymétriques. L'asymétrie des deux équilibres pose alors une problématique intrigante : chacun des joueurs va alors vouloir convaincre l'autre de jouer l'équilibre de Nash qui l'avantage. En pratique, ceci

incitera chaque candidat à annoncer vouloir voler la cagnotte et à se montrer plus convaincant que l'autre dans cette affirmation.

De telles stratégies peuvent sembler farfelues. Pourtant, de façon surprenante mais terriblement efficace, quand Nick et Abraham furent opposés au *Split or Steal* avec une cagnotte de 13 k£, Nick lança tout à coup : « Abraham, je veux que vous me croyiez. À 100 %, je vais jouer *Steal*. » Déséparé, Abraham se sentit impuissant. Il se demanda : « d'où vient votre cerveau » ? Puis, « vous êtes un idiot » !

Mais Nick n'était pas idiot. Il voulait juste s'assurer qu'Abraham aurait tout intérêt à jouer *Split* — et parce qu'il n'était pas à une ruse près, Nick joua *Split* lui aussi ! Le plus stupéfiant, c'est que, dans une interview par Radio Lab⁷, Abraham avouera qu'il comptait initialement jouer *Steal*. La stratégie de Nick aura été parfaite.

Imaginons maintenant que Steven et Sarah, inspirés par Nick, choisissent tout deux d'imiter sa stratégie. Cependant, aucun d'eux ne s'avoue vaincu suite à la discussion préalable. On tombe alors sur un problème profondément bayésien. Chacun ne peut alors pas être certain de ce que fera l'autre. Il doit amorcer un raisonnement probabiliste, fondé sur des crédences *a priori*.

Supposons par exemple que Steven attribue une probabilité d'une chance sur deux que Sarah joue *Split* et une chance sur deux de jouer *Steal*. Si Steven joue *Steal*, alors il a une chance sur deux de gagner le gain de (*Steal, Split*) qui correspond à 90 k£, et une chance sur deux de gagner celui de (*Steal, Steal*) qui correspond à 0 k£. L'espérance de gain de Steven est alors 45 k£.

Par opposition, si Steven joue *Split*, alors il gagnera une fois sur deux le gain de (*Split, Split*), soit 50 k£, et l'autre fois le gain (*Split, Steal*), soit 10 k£. Son espérance de gain serait alors de 30 k£. Le calcul de Steven le pousse alors à jouer *Steal* plutôt que *Split*, car l'espérance de gain dans le premier cas est supérieur⁸.

Puisque ce livre traite des probabilités subjectives de la théorie bayésienne, il est bon d'insister sur le fait que les probabilités en jeu ici pour Steven et Sarah ne sont absolument pas fréquentistes. Malgré le langage fréquentiste que j'ai utilisé par souci pédagogique, Steven et Sarah ne jouent pas à *Split or Steal* tous les jours. De plus, les gains potentiels représentent des cas uniques et isolés dans leurs vies respectives. Et il n'est pas question de parler de gain moyen à ce jeu. Le gain espéré ici, ou espérance de gain, correspond aux gains qu'estiment leurs probabilités subjectives.

Revenons à Steven et Sarah. Supposons maintenant que tous deux soient très malins et savent que l'autre l'est tout autant. Lors des discussions préalables, Steven et Sarah ont tous deux vraiment donné l'impression qu'ils joueraient

⁷  *The Golden Rule* | Radio Lab (2017)

⁸ On suppose ici que les joueurs sont vénaux et averses au risque, c'est-à-dire que l'utilité $u(x)$ de gagner une quantité d'argent x est $u(x) = x$.

Steal. Ils tombent donc sur un mur. Pour sortir de l'impasse, Steven affirme tout à coup qu'il jouera *Steal* avec probabilité 4 sur 5, et jouera *Split* sinon. Il rajoute qu'il invite Sarah à faire de même.

Sarah se met alors à effectuer ses calculs d'espérance. En supposant que Steven fera ce qu'il a annoncé, en jouant *Steal*, elle gagnera 0 k€ quatre fois sur cinq, et 90 k€ une fois sur cinq, ce qui lui donne une espérance de gain de⁹ 18 k€. À l'inverse, en jouant *Split*, elle gagnera 10 k€ quatre fois sur cinq, et 50 k€ une fois sur cinq, pour une espérance de¹⁰ 18 k€ aussi. Elle en conclut donc que son choix n'affectera pas son espérance de gain. Qui plus est, elle se rend compte qu'en suivant le conseil de Steven, elle donnera les incitatifs à Steven de persister dans sa stratégie. Non seulement Sarah accepte. Elle confirme même qu'elle jouera bien *Steal* avec probabilité 4 sur 5. En termes techniques, on a là un équilibre de Nash dit *mixte*, c'est-à-dire où les joueurs ont des stratégies randomisées.

Si Steven et Sarah maximisaient leurs gains et refusaient l'asymétrie, la théorie de Nash affirme que c'est cet équilibre mixte qu'ils seraient amenés à jouer¹¹. Cependant, cet équilibre de Nash, à l'instar de la plupart des équilibres de Nash, est une situation sous-optimale, dans la mesure où Steven et Sarah s'en sortiraient tous deux mieux s'ils avaient tous deux joué *Split*.

C'est là qu'intervint le futur prix Nobel 2005 Robert Aumann. En 1974, Aumann proposa d'introduire un signal, un espèce de feu rouge, qui coordonnerait les joueurs dans leurs prises de décision.

Par exemple, il pourrait s'agir d'une pièce pile ou face. Si la pièce tombe sur pile, Steven jouerait *Steal* et Sarah *Split*. Si elle tombait sur face, ce serait l'inverse. Le génie de ce signal extérieur, c'est que, sachant ce qu'a donné la pièce, Steven et Sarah auront alors tout intérêt à écouter la pièce. Autrement dit, les stratégies optimales de Steven et Sarah, sachant le signal et sachant la manière dont les uns et les autres sont censés réagir au signal, sont de suivre les instructions suggérées par le signal. Il s'agit donc presque d'un équilibre de Nash. On parle d'équilibre corrélé. Tout équilibre de Nash peut d'ailleurs être interprété comme un équilibre corrélé sans signal.

Cependant, l'ajout du signal peut grandement améliorer les équilibres de Nash. Dans le cas de Steven et Sarah, l'espérance de gain de Steven et Sarah avant de voir le signal passe alors à 50 k€. Autrement dit, le signal leur permet de faire aussi bien que s'il n'y avait pas d'incitatifs individuels à dévier du¹² (*Split*, *Split*) !

⁹Ceci correspond au calcul $\frac{4}{5} \cdot 0 + \frac{1}{5} \cdot 90 = 18$.

¹⁰Ceci correspond au calcul $\frac{4}{5} \cdot 10 + \frac{1}{5} \cdot 50 = 18$.

¹¹Nash a ainsi prouvé que tout jeu symétrique possède un équilibre symétrique (possiblement mixte).

¹²Il n'est pas trop dur de voir que l'enveloppe convexe des équilibres de Nash est inclus dans l'ensemble des équilibres corrélés. Cependant, l'ensemble des équilibres corrélés est en général plus large que cette enveloppe convexe, ce qui permet parfois de déterminer des équilibres corrélés nettement plus optimaux que n'importe quelle combinaison d'équilibres de Nash !

Les jeux bayésiens

La combinaison de la théorie des jeux et des probabilités subjectives de la philosophie bayésienne constitue un arsenal puissant pour attaquer des problèmes complexes de décision en présence d'incertitude. C'est le cas par exemple pour l'étude du poker. Au début d'une main de poker, chaque joueur connaît ses cartes, mais ne connaît pas celles des autres joueurs. Cependant, au fur et à mesure que le jeu avance, certains joueurs vont renoncer aux gains (on dit qu'ils se « couchent »), tandis que d'autres vont augmenter la mise.

Notre *pure bayésienne* appliquera alors la formule de Bayes pour modifier ses crédences sur les mains des autres. Typiquement, si un adversaire est plus agressif que d'habitude et mise de grandes sommes, elle va davantage croire que l'adversaire a de bonnes cartes en main. Bien entendu, ceci ne prouve pas que ses cartes sont en effet bonnes. Mais pour prendre des décisions optimales, il est indispensable de mettre à jour nos crédences et d'ajuster nos stratégies à ces nouvelles crédences.

Sans grande originalité, la théorie des jeux dans de tels contextes d'incertitude est appelée théorie des jeux bayésiens. Cette théorie a été développée en 1967 dans une série de trois articles par John Harsanyi, qui gagnera ensuite le prix Nobel d'économie en 1994 en compagnie de Nash. En insérant notamment de l'incertitude vis-à-vis des cartes en main des autres joueurs, ou vis-à-vis de leurs préférences, Harsanyi permet de rendre la théorie des jeux plus réaliste et de l'adapter à de nombreux problèmes réels où l'information pertinente à nos prises de décisions est incomplète.

Harsanyi utilisa aussi le langage bayésien pour expliquer la pertinence des équilibres de Nash mixtes. Harsanyi explique ainsi que l'incertitude sur les stratégies des individus trouve son origine dans l'incertitude sur leurs préférences. C'est ce que suggère le théorème de purification de Harsanyi. De façon plus cruciale, Harsanyi mit enfin la philosophie bayésienne au cœur du raisonnement économiste. Roger Myerson, futur prix Nobel d'économie 2007, affirma en 2004 : « La cohésion et la portée de la théorie économique moderne de l'information viennent du formalisme de Harsanyi. »

Dans le contexte bayésien, l'équivalent de l'équilibre de Nash est désormais appelé équilibre de Bayes-Nash. Il s'agit là encore d'une situation où tout joueur suit une stratégie qui est une meilleure réponse aux stratégies des autres. Cependant, il y a une subtilité à comprendre derrière le concept de « stratégie » dans un contexte bayésien. Une stratégie est désormais une façon d'agir en fonction de son information privée, qu'il s'agisse des préférences individuelles ou de ses cartes en main. Par exemple, dans le cas du poker, une telle stratégie peut consister à doubler la mise si on a une paire d'as, et à renoncer à la main sinon — cette stratégie n'est probablement pas optimale.

La conception de mécanismes bayésiens*

On peut enfin en revenir à la *magouilleuse*. De prime abord, au lieu de faire une *magouilleuse*, on pourrait chercher à mettre en place un processus de négociations entre élèves qui finit par conclure à une décision globale d'affectations. Outre la complexité intimidante de cette alternative, la théorie de la conception de mécanisme nous a fourni un joli théorème qui, s'appuyant sur des principes bayésiens¹³, permet de montrer que l'on peut toujours se ramener au cas relativement simple où une autorité centrale, ici la magouilleuse, collecte des préférences des élèves pour prendre une décision globale. Le théorème en question est une conséquence du principe de révélation¹⁴.

Pour comprendre ce principe, considérons un mécanisme quelconque M où chaque individu se comporte de sorte à maximiser son utilité espérée à la lumière de ses crédences bayésiennes. Autrement dit, les individus jouent un équilibre de Bayes-Nash du mécanisme M . Une fois le mécanisme achevé, on obtient une décision globale x pour le groupe, par exemple une affectation dans les différentes armées. Le principe de révélation nous permet d'obtenir la même décision globale x via une sorte de *super-magouilleuse*. Cette *super-magouilleuse* collecte les préférences des individus et simule ensuite les individus jouant l'équilibre de Bayes-Nash dans le mécanisme M . En particulier, elle peut ainsi calculer les résultats et en déduire la décision globale x pour le groupe. C'est cette décision globale pour le groupe que la *super-magouilleuse* décide de sélectionner.

Le génie de ces simulations, c'est que, du point de vue des individus, tout ce qu'ils ont à faire, c'est révéler leurs préférences. Notre *super-magouilleuse* en déduit ensuite les conséquences. Mieux encore, et contrairement au cas des scrutins uninominaux ou de la *magouilleuse* utilisée à Polytechnique, chaque individu a alors tout intérêt à révéler ses préférences de manière honnête. En effet, si un individu révèle des préférences qui ne sont pas les siennes, notre *super-magouilleuse* simulera cet individu se comportant avec des préférences qui ne sont pas les siennes. Son comportement dans la simulation ne sera donc pas optimal pour lui, ce qui l'empêchera d'influencer la décision globale à son avantage.

C'est sur cette astuce que repose le principe du « privilège avocat-client ». En effet, ce qu'on aimeraient, c'est déterminer un verdict pour un accusé. Idéalement, ce verdict prendrait en compte des informations que seul l'accusé connaît. Cependant, celui-ci n'a, *a priori*, pas intérêt à révéler tout ce qu'il sait — notamment s'il sait qu'il est coupable. Du coup, les tribunaux mettent en place un système d'interaction où l'accusateur et l'accusé débattent, et la sentence finale est l'issue de ce système d'interaction. Le principe du « privilège avocat-client » consiste à ajouter des intermédiaires qui vont simuler le débat entre l'accusateur et l'accusé sans requérir l'intervention directe de l'accusateur et de l'accusé. Ces

¹³Il existe aussi des versions non-bayésiennes du principe de révélation.

¹⁴  Favoriser l'honnêteté | Démocratie 18 | Science4All | L.N. Hoang (2017)

personnes qui simulent l'accusateur et l'accusé sont le procureur et l'avocat de la défense. Cependant, pour que ces avocats soient des bonnes simulations de l'accusateur et de l'accusé, et pour qu'ils agissent de façon optimale pour leurs clients, ils se doivent de savoir tout ce que leurs clients savent. Et pour cela, il faut que les clients aient tout intérêt à révéler à leurs avocats tout ce qu'ils savent. C'est ce que cherche à garantir le « privilège avocat-client ».

En théorie des jeux, l'utilité première du principe de révélation est de justifier le fait que l'on puisse s'intéresser, (presque) sans perte de généralité, aux mécanismes centralisés qui consistent à collecter les informations secrètes des individus pour prendre une décision globale. Ce principe a notamment permis de découvrir un mécanisme très général pour s'assurer que la décision globale maximisera le « bien-être social », à savoir la somme des utilités des individus, tout en favorisant l'honnêteté des parties prenantes. Ce procédé remarquable est appelé mécanisme VCG du nom de William Vickrey, prix Nobel d'économie de 1996, Edward Clarke et Theodore Groves¹⁵.

L'enchère de Myerson

Dans le cas d'une enchère d'un seul bien, maximiser le bien-être social revient à attribuer le bien en vente à celui qui lui attribue le plus de valeurs. C'est ce que garantit le mécanisme VCG en faisant ensuite payer le bien au prix du second acheteur. Cependant, *a priori*, cette enchère, dite *au second prix*, ne semble pas maximiser les gains du vendeur. Pour le vendeur, ne vaudrait-il mieux pas tout simplement vendre le bien au prix du premier acheteur ?

La réponse surprenante de Roger Myerson, prix Nobel d'économie de 2007, fut non. Si le bien va toujours au premier acheteur, peu importe la manière dont le prix qu'il paie est déterminé et pourvu que les acheteurs se comportent tous comme des bayésiens qui maximisent leurs utilités espérées, le revenu espéré du vendeur sera toujours le même ! C'est le surprenant théorème d'équivalence des revenus de Myerson¹⁶.

Il y a toutefois une petit détail technique sur lequel repose ce théorème de Myerson, qui donne une idée de la complexité des interactions humaines en pratique. Pour que le théorème de Myerson soit valide, il faut aussi supposer que les acheteurs et le vendeur ont tous un *a priori* commun sur ce que chacun est prêt à payer pour acquérir le bien. Il s'agit là d'une hypothèse communément admise dans la théorie des jeux bayésiens, pour simplifier les calculs et éviter de se poser des questions métaphysiques. Cependant, cette hypothèse est fausse. « Tous les modèles sont faux. »

¹⁵  Socialement optimal : le mécanisme VCG | Démocratie 20 | Science4All | L.N. Hoang (2017)

¹⁶ Comme le mécanisme VCG, ce théorème suppose que les individus ont des utilités dites *quasi-linéaires*. Qui plus est, le théorème d'équivalence des revenus suppose que les valuations du bien par les acheteurs sont indépendantes et identiquement distribuées.

Dans le cas général, se pose en fait la question des *higher beliefs*, à savoir de ce qu'un premier acheteur croit concernant la crédence d'un autre acheteur sur ce que le premier acheteur est prêt à payer, voire sur ce qu'il croit que l'autre croit qu'il croit sur les préférences de l'autre, et ainsi de suite. Il est intéressant de se rendre compte que notre *pure bayésienne* prend constamment en compte de telles réflexions. Cependant, l'étude de ces *higher beliefs* est extrêmement ardue et dépasse de loin le cadre de ce livre. Revenons-en donc au cadre fixé par Harsanyi et Myerson, dans lequel le théorème d'équivalence de revenu est vérifié.

Est-il alors possible pour le vendeur d'acquérir une meilleure espérance de gain avec une autre enchère ? La réponse non moins surprenante de Myerson est oui. Dans un merveilleux article, en 1981, Myerson réussit à déterminer l'enchère optimale pour le vendeur¹⁷. Les détails de cette enchère sont un peu techniques, mais l'idée fondamentale est très simple. Pour maximiser ces gains, le vendeur doit utiliser son préjugé bayésien. En particulier, il doit refuser la vente de son bien si les prix annoncés sont nettement inférieurs à ses crédences sur ce que les acheteurs sont réellement prêts à payer. Bien entendu, ce faisant, le vendeur a une probabilité non nulle de ne tirer aucun revenu dans l'enchère. Mais ce qui importe, c'est le revenu espéré.

L'enchère de Myerson explique pourquoi les touristes étrangers ont tellement de mal à négocier les prix. C'est tout simplement parce que, dans de nombreux pays en voie de développement, les vendeurs estimate qu'un touriste étranger est prêt à dépenser de plus grandes sommes qu'un local. Pire encore, si un touriste étranger et un local se disputent un même bien, l'enchère de Myerson conseille au vendeur d'effectuer une négociation de prix à distance : il peut vouloir accepter la vente du bien à l'étranger uniquement si le prix qu'il annonce est au moins le double de celui de l'individu local. On pourrait croire que le vendeur et le local sont de mèche pour monter le prix du bien — et c'est peut-être le cas en pratique. Mais ceci n'a pas besoin d'être le cas pour justifier cette stratégie de vente du vendeur. Il ne fait là qu'appliquer l'enchère de Myerson qui discrimine les acheteurs dont on pense qu'ils sont prêts à payer plus. Ce n'est pas qu'une question de talent de négociation ! C'est avant tout une question de crédences bayésiennes !

Les conséquences sociétales du bayésianisme

Une conséquence majeure de l'enchère de Myerson est que le comportement optimal d'un bayésien conduit à une discrimination entre les individus par les préjugés. Cela ne veut absolument pas dire qu'il faut rejeter la théorie de Myerson pour des raisons éthiques. La théorie de Myerson est un théorème mathématique qui ne repose sur aucun fondement éthique. Il n'y a rien à rejeter.

¹⁷  *La négociation optimale* | Démocratie 19 | Science4All | L.N. Hoang (2017)

De la même façon, interdire le raisonnement bayésien parce que dans ce cas il conduit à des conséquences peu souhaitables moralement, ce serait comme interdire la réflexion sous prétexte qu'il aide les milliardaires à devenir toujours plus riches.

Là où des valeurs morales apparaissent dans la théorie de Myerson, c'est dans les préférences des vendeurs et des acheteurs. S'il y a une chose à combattre, ce seraient des préférences éthiquement discutables des vendeurs et des acheteurs. Cependant, ce que montre vraiment la théorie de Myerson, c'est que des préférences anodines et socialement admissibles des vendeurs peuvent conduire à des comportements moralement rejetés par nos sociétés. Bien souvent, le problème n'est pas que les responsables de discrimination ont voulu faire du mal ; le problème c'est qu'ils n'ont pas été suffisamment encouragés à faire le bien. *Le tort des pires torts n'est généralement pas la volonté de causer un tort ; c'est souvent davantage l'omission de la volonté de ne pas causer de tort.*

Bien entendu, le contexte de l'enquête de Myerson n'est *a priori* pas si sujet à controverse — encore que les touristes se plaignent souvent du traitement de faveur qu'ils subissent et que certaines ethnies sont stigmatisées pour leurs richesses. Cependant, je vous invite à décliner ses variantes plus sensibles. Faut-il utiliser ses préjugés au moment de recruter un nouvel employé, de juger un suspect ou de calculer les malus d'une assurance ?

Pour la *pure bayésienne*, il s'agit là d'une question qui sort du cadre de sa philosophie du savoir. On a là une question d'éthique, de morale, de valeur, de fonction objectif, de préférences. La *pure bayésienne* n'a pas d'avis à émettre sur cette question.

Ignorer les préjugés pour prendre des décisions est néanmoins une valeur morale qui gagne en popularité, notamment dans les milieux intellectuels. Comme on l'a vu, une grosse partie de la recherche en informatique consiste justement à garantir la confidentialité des données personnelles, pour empêcher d'autres d'affiner leurs crédences bayésiennes.

Cependant, il y a des cas où les inaptitudes de certains nous empêchent de les traiter comme tout le monde. C'est le cas notamment du problème de l'affectation des élèves polytechniciens aux différentes armées. Dans le cas de la *magouilleuse*, pour des raisons d'aptitudes médicales, il est impossible de ne pas discriminer les différents élèves. Dès lors, plutôt que d'ignorer les différences entre individus, il pourrait être plus souhaitable d'inclure dans notre philosophie morale une façon de gérer convenablement ces distinctions entre individus.

En fait, il est même en général utile de profiter des préjugés pour déterminer des politiques plus éthiques. Typiquement, de nombreux bus à travers le monde demandent aux passagers de laisser la priorité aux places assises aux personnes âgées, aux femmes enceintes et aux personnes handicapées. Autrement dit, ces bus invitent leurs passagers à fonder des préjugés à partir de l'apparence physique des uns et des autres pour favoriser les personnes à qui ces places assises ont plus de chance de profiter.

De façon plus générale, comme on le verra dans le dernier chapitre de ce livre, il existe une branche de la philosophie morale pour laquelle les préjugés sont moralement désirables. Cette philosophie morale est le conséquentialisme. Pour le conséquentialiste, seules les conséquences (probables) de nos actions importent. Le conséquentialiste bayésien devra alors utiliser tout l'arsenal bayésien pour optimiser les bienfaits de ses actions. Or, cet arsenal inclut ses préjugés. Ignorer ces préjugés serait alors un péché moral.

En particulier, les préjugés peuvent aider à aider plus rapidement et plus efficacement ceux qui en ont le plus besoin. Il semble alors immoral de les ignorer.

Références en français

- ▶ *La théorie des jeux* | Science Étonnante | D. Louapre (2017)
- ▶ *La vidéo pas drôle mais intéressante* | Squeezie | L. Hauchard (2017)

- ▶ *La démocratie sous l'angle de la théorie des jeux* | Science4All | L.N. Hoang (2017)
- ▶ *L'équilibre de Nash* | Démocratie 13 | Science4All | L.N. Hoang (2017)
- ▶ *Le poker résolu ! (ou non)* | Démocratie 15 | Science4All | L.N. Hoang (2017)
- ▶ *Favoriser l'honnêteté* | Démocratie 18 | Science4All | L.N. Hoang (2017)
- ▶ *La négociation optimale* | Démocratie 19 | Science4All | L.N. Hoang (2017)
- ▶ *Socialement optimal : le mécanisme VCG* | Démocratie 20 | Science4All | L.N. Hoang (2017)

- ⓘ *La démocratie à la moulinette des maths et de la science* | Podcast Science 84 | N. Tupégabé, D. Medernach et A. Vonlanthen (2012)
- ⓘ *Jeux* | Podcast Science 214 | R. Jamet (2015)

Références en anglais

- ☒ *Equilibrium points in n-person games* || Proceedings of the National Academy of Science | J. Nash (1950)
- ☒ *Non-cooperative games* | Annals of mathematics | J. Nash (1951)
- ☒ *Subjectivity and correlation in randomized strategies* | Journal of Mathematical Economics | R. Aumann (1974)
- ☒ *Games with randomly disturbed payoffs: A new rationale for mixed-strategy equilibrium points* | International Journal of Game Theory | J. Harsanyi (1973)
- ☒ *Optimal auction design* | Mathematics of Operations Research | R. Myerson (1981)

▣ *Comments on “Games with Incomplete Information Played by ‘Bayesian’ Players, I–III Harsanyi’s Games with Incomplete Information”* | Management Science | R. Myerson (2004)

▣ *Bayesian persuasion* | The American Economic Review | E. Kamenica and M. Gentzkow (2011)

▣ *Measuring Unfairness Feeling in Allocation Problems* | Omega | L.N. Hoang, F. Soumis et G. Zaccour (2016)

⌚ *The Golden Rule* | Radio Lab (2017)

🌐 *The Evolution of Trust* | ncase | Nicky Case (2017)

🌐 *Game Theory and the Nash Equilibrium* | Science4All | L.N. Hoang (2012)

🌐 *Bayesian Games: Math Models for Poker* | Science4All | L.N. Hoang (2012)

🌐 *Mechanism Design and the Revelation Principle* | Science4All | L.N. Hoang (2012)

🌐 *A Mathematical Guide to Selling* | Science4All | L.N. Hoang (2015)

Il y a de la grandeur dans cette vision de la vie avec cette puissance initialement insufflée dans quelques formes de vie ou une seule ; et dans le fait qu'alors que cette planète continuait son cycle de rotation selon la loi fixée par la gravité, à partir d'un début si simple, des formes infiniment belles et magnifiques ont évolué et évoluent encore.

Charles Darwin (1809-1882)

10

Darwin et Bayes font affaire

Le biais du survivant

Pendant la seconde guerre mondiale, l'armée de l'air anglaise embaucha le statisticien Abraham Wald pour investiguer le blindage optimal des avions de guerre. L'armée de l'air avait remarqué que les avions qui revenaient de la bataille étaient criblés d'impacts partout sauf à l'avant, où se trouvait le moteur. L'armée conclut que ce serait une bonne idée de réduire le blindage à l'avant pour le renforcer à l'arrière. *Faux*, s'exclama Wald ! Il rétorqua qu'au contraire, le fait que les avions étaient criblés de balles uniquement à l'arrière prouvait que c'était l'avant de l'avion qu'il fallait renforcer.

Cette remarque de Wald a de quoi surprendre. Il s'agit pourtant essentiellement du même argument que celui de Charles Darwin pour expliquer l'émergence des structures complexes du vivant. Dans les deux cas, le processus subtil qui échappe à la plupart d'entre nous est un processus d'élimination, ou, si l'on est concerné par les survivants, de sélection. Dans le cas de Wald, l'élimination est celle des avions dont l'avant a été touché. Ces avions ayant vu leurs moteurs être détruits, voire exploser, ils n'ont pas pu revenir au bercail. De façon similaire, Darwin affirme que les espèces animales dont les déficiences ont empêché leurs reproductions ont inéluctablement disparu. Par conséquent, celles qui subsistent aujourd'hui ont remarquablement peu de déficiences majeures.

Si elle est unanimement célébrée par la communauté scientifique, la théorie de l'évolution de Darwin connaît aujourd'hui encore de nombreuses critiques

pseudo-scientifiques. Ces derniers lui opposent l'argument du *design* intelligent. L'argument est le suivant. Imaginez-vous en plein désert. Si vous tombez sur un caillou difforme, vous ne serez pas surpris d'apprendre qu'il est le fruit de processus naturels. En revanche, si vous tombez sur une montre dont les rouages sont complexes, il paraît stupide de penser qu'elle a pu émerger de processus purement naturels. La sophistication de la montre ne semble pouvoir s'expliquer que par le travail minutieux d'un concepteur intelligent. De la même manière, la sophistication remarquable du corps humain, de la biomécanique de ses os et de ses muscles à l'organisation de son système immunitaire, en passant par l'ingéniosité de l'œil et la complexité incompréhensible du cerveau, ne peuvent être que le résultat d'un *design* intelligent — et ce concepteur intelligent ne peut être que Dieu.

Cet argument peut sembler très convaincant. Mais, outre l'amalgame discutable entre un concepteur intelligent et Dieu, ce serait sous-estimer le processus d'élimination dont on a parlé plus haut — que Darwin appelait sélection naturelle.

Les lézards colorés de Californie

Partons dans la Central Valley en Californie, où vivent trois variétés différentes de lézards mâles. De façon grossière, il y a les lézards orange, les lézards bleus et les lézards jaunes. Ces lézards mâles sont de la même espèce, et cherchent donc à se reproduire avec les mêmes femelles. Mais leurs attributs et leurs stratégies de reproduction sont très distincts. Les lézards orange sont des grosses brutes. Ils contrôlent un territoire et se reproduisent avec toutes les femelles de leur territoire. Les lézards bleus sont des monogames jaloux qui contrôlent tous les faits et gestes de leurs compagnes. Enfin, les lézards jaunes sont des Don Juan furtifs, qui se jettent sur toute femelle qu'ils rencontrent.

La théorie de l'évolution darwinienne suggère que les lézards les plus à même de se reproduire seront ceux qui subsisteront. Cependant, ce qui est amusant, c'est que la capacité des lézards mâles à se reproduire dépend des populations de lézards mâles présentes.

Par exemple, supposons que la plupart des lézards sont des brutes orange. Alors, chaque lézard orange va conquérir un large harem qu'il ne pourra pas bien surveiller. Dès lors, les furtifs jaunes pourront aisément profiter des lézards femelles non surveillées, de sorte que les femelles soient plus souvent inséminées par des furtifs jaunes que par les brutes orange. Petit à petit, on pourrait alors s'attendre à ce que les furtifs jaunes prennent le pas sur les brutes orange.

Imaginons maintenant que les furtifs jaunes soient prédominants. Les jaloux bleus pourront alors charmer les femelles et se les garder pour eux, de sorte que, petit à petit, toutes les femelles soient casées avec un jaloux bleu. Mais alors, les furtifs jaunes ne pourront pas profiter de femelles laissées vacantes, et ne

pourront pas se reproduire. Les jaloux bleus causeraient donc l'extinction des furtifs jaunes.

Postulons enfin que les lézards mâles soient presque tous des jaloux bleus. Dès lors, les brutes orange pourraient combattre les jaloux bleus, et ainsi agrandir leur harem, femelle après femelle. Les jaloux bleus deviendraient alors tous célibataires et ne pourraient plus se reproduire. Ils finiraient donc par disparaître au profit des brutes orange.

Faisons le point. En gros, orange perd contre jaune, jaune perd contre bleu et bleu perd contre orange. Voilà qui ressemble pas mal au jeu du chifoumi, où pierre bat ciseaux, ciseaux bat feuille et feuille bat pierre. Ce jeu possède un unique équilibre de Nash qui consiste alors à jouer aléatoirement les trois options. Devinez quoi. On observe en pratique que les trois variétés de lézards mâles co-existent dans la nature, comme s'ils avaient choisi de jouer l'équilibre de Nash du chifoumi ! Autrement dit, le concept d'équilibre de Nash, qui n'est censé n'être joué que par des joueurs intelligents, semble parfaitement s'appliquer à ce à quoi a conduit l'évolution darwinienne. Comme on va le voir, il ne s'agit pas là d'une coïncidence.

La dynamique de Lotka-Volterra*

En 1972, le biologiste John Maynard Smith inventa le concept de stratégies évolutionnairement stables. Smith définit ces stratégies comme étant une certaine composition de notre population qui soit robuste à l'invasion d'une (petite) population avec une composition différente (par exemple l'ajout de 100 mâles jaunes). En pratique, ceci correspond typiquement à des variations aléatoires de la population dues à des fluctuations statistiques. Ces fluctuations statistiques causeront-elles des modifications profondes de la population ? Ou l'évolution darwinienne ramènera-t-elle la composition de la population à ce qu'elle était avant ces fluctuations statistiques ?

Pour répondre à ces questions, on va plonger dans les détails mathématiques d'un modèle simplifié de l'évolution darwinienne. « Tous les modèles sont faux. » Mais celui dont on va parler a été utile à de nombreux biologistes.

Appelons $x(t)$ le nombre d'individus d'une variété à un instant t . À la génération suivante $t + 1$, on sait que la population va augmenter du nombre de naissances et diminuer du nombre de morts. Ces nombres de naissances et de morts vont être à peu près proportionnels à la population entière. Les coefficients de proportionnalité sont les taux de natalité (noté $\% \text{😊}$) et de mortalité (noté $\% \text{💀}$). La population devient alors $x(t + 1) = x(t) + (\% \text{😊})x(t) - (\% \text{💀})x(t) = x(t) + (\% \text{😊} - \% \text{💀})x(t)$. Autrement dit, la variation de la population est proportionnelle à la population $x(t)$. Appelons **fitness** = $\% \text{😊} - \% \text{💀}$ ce coefficient de proportionnalité. On obtient alors l'équation qui régit l'évolution de la population : $\dot{x} = x(t + 1) - x(t) = \text{fitness} \cdot x$.

L'équation ci-dessus est valable pour les individus d'une variété. Si on distingue maintenant les différentes variétés par un indice i , on obtient les équations de Lotka-Volterra $\dot{x}_i = \text{fitness}_i \cdot x_i$. En fait, ces équations sont un peu plus précises et nous indiquent comment les *fitness* des différentes variétés varient en fonction des populations des autres variétés. Comme on l'a vu, si la population dominante est celle des brutes orange, alors la *fitness* des furtifs jaunes augmentera¹.

Cependant, ce qui va nous intéresser n'est pas les nombres d'individus x_i de chaque variété i , mais les proportions z_i des différentes variétés i dans la population. Après quelques manipulations algébriques, que je vous laisse en exercice, on obtient l'équation qui régit les variations des proportions des variétés dans la population :

$$z_i(t+1) = \frac{(1 + \text{fitness}_i)z_i(t)}{(1 + \text{fitness}_i)z_i(t) + \sum_{j \neq i} (1 + \text{fitness}_j)z_j(t)}.$$

Vous le voyez venir ? L'équation qui régit l'évolution n'est autre qu'une formule de Bayes déguisée ! Incroyable ! Les probabilités subjectives correspondent aux proportions z_i . Au moment de passer de t à $t+1$, ces probabilités subissent une espèce d'inférence bayésienne où les termes d'expérience de pensée sont remplacés par les quantités $1 + \text{fitness}_i$. Enfin, au dénominateur, on retrouve la fonction de partition, qui permet de s'assurer que la somme des z_i est encore égale à 1 au temps $t+1$.

Voici la conclusion stupéfiante de cette analyse. Dans la mesure où la *fitness* à l'instant t d'une variété i peut s'identifier à la capacité d'une théorie i à expliquer les données survenues à l'instant t , l'évolution darwinienne est indiscernable d'un être rationnel !

Cette comparaison peut paraître farfelue. Elle est pourtant corroborée par un théorème remarquable (même s'il est mathématiquement trivial) que prouva le biologiste John Maynard Smith en 1973. Ce théorème affirme que les proportions de la population auxquelles l'évolution darwinienne conduit sont nécessairement des équilibres de Nash. Ce qui est étonnant, c'est que ces équilibres de Nash correspondent à des stratégies dans des jeux joués par agents intelligents et rationnels. Autrement dit, à l'instar de la montre trouvée en plein désert, les proportions qui décrivent des équilibres de Nash semblent ne pouvoir être que le résultat d'un dessein intelligent. C'est du moins ce que l'on pourrait croire naïvement. Mais il n'en est rien.

Ce qui semblait devoir être le fruit d'un dessein intelligent peut tout autant n'être que la conséquence inévitable de l'évolution darwinienne. Telle est la conclusion stupéfiante du théorème de Maynard Smith.

¹Dans leurs formes classiques, les équations de Lotka-Volterra supposent que les *fitness* sont des fonctions affines des populations.

Les algorithmes génétiques

L'évolution darwinienne est même bien plus qu'une pâle copie de l'intelligence humaine. En fait, elle est aisément capable de créer des structures qu'une intelligence humaine aurait du mal à penser d'elle-même. L'exemple classique à citer est celui du cerveau humain. L'évolution a su le concevoir. Mais le cerveau humain échappe encore complètement à la compréhension des neuroscientifiques, même une fois équipés de supercalculateurs.

Cette sophistication stupéfiante à laquelle donne lieu l'évolution darwinienne est telle que les informaticiens et mathématiciens appliqués se sont tournés vers des algorithmes dits *génétiques*, pour trouver des solutions à des problèmes qu'ils ne savent pas résoudre autrement. Ces algorithmes génétiques imitent la sélection naturelle, mais aussi les croisements et les mutations génétiques.

Supposons que l'on cherche par exemple à déterminer l'ordre dans lequel visiter les 100 plus grandes villes de France de sorte à minimiser le temps de trajet total. Ce problème est connu sous le nom de problème du voyageur de commerce. Chaque ordre de visite des villes est une solution du problème. L'objectif est de déterminer la solution optimale. La difficulté de ce problème réside dans le nombre monstrueux de solutions. Il y en a $100! \approx 10^{157}$. Même si on combinait tous les supercalculateurs sur Terre pour lister tous ces ordonnancements, il faudrait un temps beaucoup plus grand que l'âge de l'univers pour y parvenir.

Les algorithmes génétiques forment une approche remarquablement efficace pour ce genre de problèmes. Le principe de ces algorithmes est de maintenir vivante une certaine population diversifiée de solutions prometteuses mais sous-optimales. À chaque étape, l'algorithme sélectionne deux solutions de la population et les croisent, en y ajoutant des mutations (favorables), avant qu'une phase de sélection élimine les plus mauvaises solutions. Étrangement, cette approche darwinienne de l'optimisation s'en sort étonnement bien, au point d'être l'état de l'art dans de nombreux cas !

L'évolution darwinienne s'en sort là bien mieux que l'intelligence humaine. La sophistication de la Nature n'est donc pas un argument convaincant contre la théorie de l'évolution. On y reviendra d'ailleurs dans le prochain chapitre.

Se faire son propre avis ?

La distinction entre science et pseudo-science est le thème de prédilection d'un mouvement de pensée connu sous le nom de scepticisme, de pensée critique ou encore de zététique. Ce mouvement insiste sur les sophismes récurrents et les biais cognitifs des tenants des pseudo-sciences. Ces erreurs de raisonnement ont en effet le mauvais goût d'être aux fondements de nombreuses théories conspirationnistes, de nombreuses médecines alternatives et de nombreux phénomènes paranormaux.

Pour certains, la bonne réaction à ces questions est de se faire une idée par soi-même. Cependant, le danger d'une telle réaction est qu'elle mène inéluctablement à une remise en cause de tout ce qui nécessite un bagage intellectuel ou empirique important, pour se faire une opinion suffisamment informée pour être pertinente. Voire à des biais inévitables, des contre-sens et des erreurs. C'est le cas par exemple du problème de Linda, de la *p-value* ou du concept de confidentialité différentielle. C'est aussi le cas, de manière plus capitale, de l'efficacité des vaccins, des algorithmes utilisés par Google et Facebook et de l'origine anthropique du changement climatique. À moins de passer des années à étudier minutieusement ces questions, l'opinion que vous pourriez vous faire de vous-même ne sera pas bien informée ; elle ne sera donc pas pertinente.

Il est tentant de croire qu'en y passant plusieurs heures, on pourrait tout de même finir par pencher du bon côté de la balance sur ces questions. Ceci n'a rien d'évident. À l'instar du problème de Linda, il arrive que notre intuition nous conduise à un taux d'échec supérieur à celui d'un chimpanzé qui ferait des prédictions aléatoires. C'est ce que s'amusait à montrer le statisticien Hans Rosling. Sur de nombreuses questions comme le nombre d'années d'éducation des femmes, le nombre de décès par catastrophes naturels ou la pauvreté dans le monde, nous sommes pires qu'ignorants² ; nous penchons systématiquement du mauvais côté !

Pire, il nous est souvent extrêmement difficile d'estimer le bon degré de confiance à avoir en notre intuition. Ainsi, même après avoir pris le temps de réfléchir et de s'informer sur un sujet, il peut être remarquablement difficile de comprendre à quel point on l'a compris, et si notre avis sur le sujet peut vraiment être considéré comme un avis informé. Pire encore, la thèse de Derek Muller³ montre que le visionnage d'explications physiques parfaitement justes peut augmenter la confiance qu'ont les étudiants en leurs intuitions, alors même que ces intuitions sont contredites par les vidéos que les étudiants viennent de visionner !

Cet excès de confiance récurrent dont nous faisons tous beaucoup trop preuve est, comme vous l'aurez compris, le principal biais cognitif que je cherche à combattre dans ce livre. C'est ce que la formule de Bayes, les difficultés d'Erdős face au Monty Hall et l'incomplétude de Solomonoff devraient nous forcer à reconnaître : nous sommes constamment en excès de confiance. Comme le disait le grand logicien Bertrand Russell, « tout le problème avec le monde est que les imbéciles et les fanatiques sont toujours si sûrs d'eux-mêmes, alors que les personnes plus sages sont pleines de doutes ». Étienne Klein rajoute : « il faut se hâter de ne pas conclure ».

En fait, croire quoi que ce soit « par soi-même » est une tâche monstrueusement difficile et semée d'embûches insurmontables. Je la déconseille très fortement. S'il était si aisément de se forger des opinions justes, les études supérieures ne seraient pas aussi longues, et l'ensemble des connaissances ne serait pas aussi segmenté

²  *How not to be ignorant about the world* | TED | H. Rosling et O. Rosling (2014)

³  *Khan Academy and the Effectiveness of Science Videos* | Veritasium | D. Muller (2011)

en disciplines disjointes. Faute d'avoir les ressources financières, temporelles et cognitives pour se plonger dans l'étude détaillée de questions précises, force est de se reposer sur les avis des autres. Il ne s'agit pas d'un mauvais réflexe. Le *bayésien pragmatique* préfère ainsi profiter des décennies, voire des siècles de travaux faits par d'autres pour affiner sa compréhension du monde. Même notre *pure bayésienne* sait que les autres individus ont eu accès à de nombreuses données auxquelles elle n'a pas eu accès ; ils ont donc des choses à lui enseigner.

Un scientifique n'est pas crédible

L'argument d'autorité est donc un outil puissant pour comprendre le monde dans lequel on vit de manière efficace et pragmatique. Toutefois, ceci soulève alors la question suivante : quelles sont les autorités les plus fiables ? Un argument venant d'Einstein a-t-il plus de valeur, ou plus de crédence, qu'un argument de Shakespeare ? Peut-on attribuer une confiance aveugle aux scientifiques ?

Quand ils sont confrontés à ces questions, certains *zététiciens* (ces défenseurs la pensée critique) et certains scientifiques mettent en avant l'objectivité de la méthode scientifique. Selon ce raisonnement, les scientifiques sont parvenus à leurs conclusions via un raisonnement parfaitement rigoureux, objectif et contrôlé par des pairs. Leurs conclusions ont donc une valeur supérieure à celle des pseudo-scientifiques, lesquels ne suivraient pas cette même démarche.

Cependant, les meilleurs zététiciens mettent en garde face à ce raisonnement grossier et caricatural. Pour commencer, certaines pseudo-sciences suivent plus ou moins les grandes lignes de la méthode scientifique. Pire, le bayésianisme rejette l'objectivité de cette méthode scientifique — et même sa validité ! Mais surtout, les scientifiques ne suivent quasiment jamais la méthode scientifique.

Prenez un article au hasard dans la littérature scientifique. Il y a de bonnes chances que les auteurs de l'article n'aient pas émis d'hypothèse, puis identifié un protocole restreint, puis effectué l'expérience conforme au protocole, puis conclu avec la *p-value*, puis complété leur article. Les sciences, modernes ou passées, sont bien plus une suite d'essais et d'erreurs, de modélisations et de simulations, d'ajustements de paramètres et de questionnements en pleine expérience. Ce n'est souvent qu'une fois les résultats obtenus que l'écriture de l'article commence. Dès lors, l'angle choisi par les auteurs consiste souvent à ignorer la quasi-totalité des fausses pistes initiées dans le laboratoire pour mieux synthétiser l'ensemble des trouvailles et dégager une conclusion intrigante — ce que la plupart des lecteurs trouveront fort appréciable.

Pire encore, les scientifiques sont inéluctablement victimes des mêmes biais cognitifs, et même de sophismes, qui affectent les pseudo-sciences. En effet, comme on l'a vu dans les deux premiers chapitres, même les meilleurs scientifiques sont impuissants face à des problèmes d'une simplicité pourtant déconcertante, à l'instar d'Erdös face au problème de Monty Hall. Il y a eu des époques où les

meilleurs scientifiques pensaient que la Terre était le centre de l'univers⁴, que la géométrie était forcément euclidienne⁵ ou que les réseaux de neurones artificiels étaient une *dead-end* de la recherche en intelligence artificielle — ce fut ma réaction quand j'en découvris la formalisation mathématique en 2011 !

Même le grand Einstein, dont les percées paraissent miraculeuses aux yeux de beaucoup de physiciens, se trompa à de nombreuses reprises, en défendant une théorie de la relativité générale erronée⁶ en 1913 ou en introduisant une constante cosmologique⁷ dans ses équations pour forcer l'univers à être stable et éternel — ce qu'il appellera « la plus grosse bêtise de sa vie ». Aussi intelligents soient-ils, les plus grands scientifiques jouissent néanmoins, et jouiront toujours, de capacités cognitives limitées.

Mais il y a pire encore. Le système académique induit des incitatifs qui ne sont pas compatibles avec une lutte permanente contre les biais cognitifs. En effet, le prestige d'un scientifique, ou sa simple faculté à conserver son poste, reposent entre autres sur l'originalité de ses idées et sur le nombre de ses publications. Dans ce contexte, un scientifique a tout intérêt à défendre outrageusement ses idées, souvent au-delà de ce que la formule de Bayes autoriserait. Il aura même intérêt à ne jamais dénigrer les théories qu'il a développées dans le passé et qui ont fait sa gloire, même si ces théories finissent par être réfutées. Enfin, il n'a pas intérêt à prendre le temps de vérifier la validité des théories concurrentes, puisque les journaux ne publient pas de telles consolidations de théories déjà existantes.

Enfin, il y a les cas extrêmes, mais avérés, de scientifiques dont les sources de financement exigeaient des conclusions prédéterminées, à l'instar d'un Ronald Fisher qui avait vendu son âme à l'industrie du tabac. Or, l'existence de ces sources de financement malsaines ne peut jamais être complètement exclue.

Ces nombreux arguments semblent mettre à mal la crédibilité des scientifiques. D'ailleurs, quand je vois certains raccourcis utilisés par les scientifiques de renom lors de conférences grand public, ma crédence en leurs discours en prend un mauvais coup. Moi-même, lorsque je fais des vidéos sur Science4All ou ZettaBytes dont le but premier est de promouvoir les mathématiques et l'informatique, je préfère violemment esquiver les difficultés techniques pour délivrer un message clair, convaincant et divertissant. Voilà qui, à plusieurs reprises, m'a conduit à mentir à mon auditoire — y compris dans ce livre. D'autres scientifiques que j'admire profondément m'ont d'ailleurs précédé dans ce mensonge intentionnel. Mais ce n'est absolument pas étonnant. Celui qui présente le théorème de Gödel sans avoir introduit la logique du premier ordre commet nécessairement de petits mensonges. L'effort de promotion des sciences auprès d'un grand public nous contraint à préférer la fluidité du discours à la rigueur.

⁴ *La Terre est-elle le centre du monde ?* | Relativité 14 | Science4All | L.N. Hoang (2016)

⁵ *La géométrie hyperbolique* | Relativité 12 | Science4All | L.N. Hoang (2016)

⁶ *Et Einstein découvre la gravité...* | Infini 20 | Science4All | L.N. Hoang (2016)

⁷ *La fin du monde et la plus grosse bêtise d'Einstein* | Relativité 21 | Science4All | L.N. Hoang (2016)

L'argument d'autorité

Ceci étant dit, les opinions de certains experts sur certaines questions pointues ont une toute autre valeur à mes yeux. Ce fut typiquement le cas de mon professeur de mathématiques de première année de classes préparatoires. Comme beaucoup d'autres élèves, je fus subjugué par la pertinence de ses remarques. S'il y avait un conflit entre nos croyances concernant un problème mathématique, non seulement me mis-je immédiatement à me questionner violemment, voire à rejeter mes croyances, mais qui plus est, je me mis rapidement à croire en ses croyances, et à chercher à comprendre leur origine.

De la même façon, si un scientifique de renom, qui m'a maintes fois stupéfait par l'intelligence de ses positions, émet un avis surprenant au sujet d'une question précise de son domaine d'expertise, alors, peu importe ce que je croyais avant d'entendre cet avis, je vais rapidement augmenter mes crédences en l'avis que le scientifique a exprimé.

Ce fut le cas lorsqu'un ami logicien m'affirma que, contrairement à ce qu'indique un raisonnement grossier et contrairement à ce qu'affirmait Wikipedia, il existe des modèles mathématiques dans lesquels tous les nombres réels sont définissables. Ayant constaté maintes fois son expertise en logique mathématique, et malgré ma grande crédence en les pages mathématiques de Wikipedia, je remis grandement en cause ce que je pensais et j'en vins même rapidement à croire mon ami. Même si je ne comprenais pas pourquoi il croyait ce qu'il croyait.

Aussi étrange que cela puisse paraître, ma réaction était rationnelle ! En effet, la formule de Bayes nous force à accepter l'argument d'autorité dans ce cas présent. Notons \clubsuit le fait qu'une autorité défende une thèse, et \checkmark et \times le statut de la thèse. La formule de Bayes nous invite à calculer la crédence *a posteriori* comme suit :

$$\mathbb{P}[\checkmark|\clubsuit] = \frac{\mathbb{P}[\clubsuit|\checkmark]}{\mathbb{P}[\clubsuit|\checkmark]\mathbb{P}[\checkmark] + \mathbb{P}[\clubsuit|\times]\mathbb{P}[\times]}\mathbb{P}[\checkmark].$$

Supposons que votre *a priori* soit négatif. Autrement dit, $\mathbb{P}[\checkmark] \approx 0$ et $\mathbb{P}[\times] \approx 1$. On peut raisonnablement penser que si la thèse est vraie, alors l'autorité va la défendre, d'où⁸ $\mathbb{P}[\clubsuit|\checkmark] \approx 1$. On obtient alors l'approximation suivante :

$$\mathbb{P}[\checkmark|\clubsuit] \approx \frac{1}{1 \cdot 0 + \mathbb{P}[\clubsuit|\times] \cdot 1} \mathbb{P}[\checkmark] = \frac{\mathbb{P}[\checkmark]}{\mathbb{P}[\clubsuit|\times]}.$$

Ainsi, en première approximation, votre crédence en la thèse sera multipliée par $1/\mathbb{P}[\clubsuit|\times]$. Autrement dit, la formule de Bayes vous invite à accepter l'argument d'autorité si et seulement s'il est hautement improbable que l'autorité prenne la position qu'il a prise sachant que la thèse est fausse.

⁸Le raisonnement tient tant que $\mathbb{P}[\clubsuit|\checkmark]$ n'est pas très petit.

Voilà qui explique pourquoi l'autorité climatosceptique est rejetée par la *pure bayésienne*. Étant donné les énormes intérêts économiques des entreprises du pétrole, il n'y a absolument rien d'étonnant à ce qu'ils trouvent des individus prêts à défendre leurs positions. À cela s'ajoute un biais de sélection énorme. Si une émission veut donner la parole à un climatosceptique, alors la probabilité que la personne invitée défende le climatoscepticisme sera nécessairement égale à 1, même si sa thèse est fausse.

Néanmoins, cet argument est tout aussi valide pour le camp opposé. La probabilité qu'un militant écologique défende la thèse du réchauffement climatique dans le cas où ce réchauffement est faux est aussi proche de 1. Ce qui est vrai du militant peut même être dit du scientifique choisi pour apparaître dans les médias, sachant tous les biais cognitifs dont le scientifique est victime et dont on a parlé plus haut. En bref, pour la question du réchauffement climatique, comme pour tout sujet controversé, polarisant et associé à d'énormes intérêts économiques ou politiques, l'argument d'autorité n'a quasiment aucune validité.

De façon générale, la formule de Bayes montre que, *si vous savez ce qu'un individu va dire, alors vous n'apprendrez rien en l'écoutant parler*. Ou plus précisément, votre crédence en une thèse ne peut pas rationnellement augmenter grâce à une prise de position d'une autorité dont vous connaissez déjà la position. En effet, supposons que vous soyiez quasiment certain que l'individu va défendre la thèse. Autrement dit, supposons $\mathbb{P}[\checkmark] \approx 1$. La formule de Bayes fournit alors l'approximation suivante :

$$\mathbb{P}[\checkmark|\checkmark] = \frac{\mathbb{P}[\checkmark|\checkmark]}{\mathbb{P}[\checkmark]} \mathbb{P}[\checkmark] \approx \mathbb{P}[\checkmark|\checkmark] \mathbb{P}[\checkmark].$$

Or, toute probabilité est au plus égale à 1. On a donc $\mathbb{P}[\checkmark|\checkmark] \leq 1$, d'où l'inégalité (approximative) $\mathbb{P}[\checkmark|\checkmark] \lesssim \mathbb{P}[\checkmark]$. L'argument d'autorité ne tient alors pas⁹. Le corollaire de ce théorème bayésien, c'est que *si votre interlocuteur (bayésien) sait ce que vous voulez dire, vous ne changerez pas ses croyances avec vos discours*.

A contrario, les paroles de mon ami logicien m'ont stupéfait. En particulier, la probabilité $\mathbb{P}[\checkmark|\text{X}]$ qu'il affirme ce qu'il affirme en supposant que ce qu'il affirme n'a aucun fondement est quasi-nulle. En fait, cette probabilité est même encore plus faible que mes crédences *a priori* en une erreur conjointe de mon raisonnement et de Wikipedia, quand bien même ces crédences étaient infimes. C'est pour cela que suite à notre discussion, et même si je ne comprenais pas le raisonnement de mon ami, j'étais déjà persuadé qu'il avait raison¹⁰.

⁹On peut tout de même raisonnablement supposer que l'autorité a au moins autant de chances de défendre sa thèse lorsque celle-ci est vraie que quand celle-ci est fausse, auquel cas on aurait aussi l'inégalité $\mathbb{P}[\checkmark|\checkmark] \geq \mathbb{P}[\checkmark|\text{X}]$. Dans ce cas, on a vraiment $\mathbb{P}[\checkmark|\checkmark] \approx \mathbb{P}[\checkmark]$. Autrement dit, vous n'avez rien appris en apprenant \checkmark .

¹⁰ *Numbers and Constructibility* | Science4All | L.N. Hoang (2013)

 *4 paradoxes de la logique mathématique* | Infini 17 | Science4All | L.N. Hoang (2017)

Il m'est donc arrivé de croire quelque chose sans l'avoir compris. Pire, si je l'ai cru, c'est uniquement via un argument d'autorité. Certains affirmeraient que c'est irrationnel. Mais, même si je ne le savais pas à ce moment-là, il s'agissait en fait du seul *a posteriori* rationnel — en tout cas si l'on en croit la formule de Bayes¹¹.

Le consensus scientifique

Revenons-en au changement climatique. On a vu qu'aucun scientifique ne pouvait faire autorité. D'ailleurs, plutôt que de pointer vers un expert en climatologie, les zététiciens vont souvent davantage mettre en avant l'avis de la communauté climatologue. Or cet avis est univoque. La quasi-totalité de cette communauté croit au changement climatique et à son origine anthropique — un chiffre souvent supérieur à 98 % de cette communauté est avancé. Mais si chaque scientifique n'est pas crédible, pourquoi la communauté le serait-elle davantage ?

La *pure bayésienne* a une réponse remarquable à cette question : la communauté scientifique applique mieux la formule de Bayes que chacun de ses membres. Imaginons que la communauté scientifique est un territoire et que les théories T sont des espèces animales vivant sur ce territoire. À chaque instant t , les théories les plus crédibles vont davantage se reproduire. Elles seront acceptées par davantage de scientifiques. Appelons $p_T(t)$ la fraction de scientifiques qui adoptent la théorie T à l'instant t . Les équations de Lotka-Volterra s'appliquent alors à l'évolution des idées¹² :

$$p_T(t+1) = \frac{\text{fitness}(t, T)p_T(t)}{\text{fitness}(t, T)p_T(t) + \sum_{A \neq T} \text{fitness}(t, A)p_A(t)}.$$

Vous voyez là où je veux en venir ? *L'évolution darwinienne des théories dans la communauté scientifique est une inférence bayésienne déguisée !*

Autrement dit, tout se passe donc comme si la communauté scientifique applique la formule de Bayes pour faire émerger les théories les plus crédibles. C'est pour cette raison que la communauté mérite une crédence qui transcende de loin les opinions de chacun de ses individus. Pour peu que les *fitness* des théories soient corrélées avec les termes d'expérience de pensée, la communauté scientifique, mieux que n'importe lequel de ses membres, applique la formule de Bayes¹³.

¹¹  *How I use "meta-updating"* | J. Galef (2015)

¹² Par commodité et sans perte de généralité, j'ai remplacé les $1 + \text{fitness}$ par fitness .

¹³ On peut même ajouter une similarité entre l'exploration des théories par la communauté scientifique et les algorithmes MCMC dont on parlera au chapitre 17.

Le putaclic

Par analogie, on pourrait croire que les opinions les plus répandues dans une population sont aussi les plus crédibles. Cet argument, souvent évoqué par ailleurs pour défendre le principe de la démocratie, est cependant fallacieux. La raison est simple : les théories qui se propagent le mieux dans une population, celles qui ont les meilleurs *fitness*, ne sont pas nécessairement les théories les plus crédibles ; ce seront plutôt les théories les plus *virales*¹⁴.

Dans une vidéo¹⁵ sortie peu après les résultats de l'élection présidentielle américaine de 2016, l'excellent Derek Muller de la chaîne Veritasium a confessé son optimisme initial et naïf. Comme bien d'autres, il pensait qu'Internet permettrait de partager plus rapidement des faits. Il espérait que ceci conduirait à une convergence globale vers des valeurs et des croyances (scientifiques) communes. Cependant, comme Derek Muller l'explique lui-même, cela n'a pas été le cas. L'explication à la divergence, voire la bipolarisation idéologique, constatée par Derek Muller semble résider dans une vidéo¹⁶ d'un autre excellent YouTuber éducatif, CGP Grey. CGP Grey y suggère que la faculté d'une théorie à proliférer sur Internet réside davantage dans sa faculté à susciter des réactions émotionnelles que dans celle à expliquer des données observées. Autrement dit, à l'instar des émotions¹⁷, les *fitness* des théories sur Internet et dans le grand public semblent davantage liées à leur effet « *putaclic* » qu'aux termes d'expériences de pensée de la formule de Bayes.

Pire encore, en se fondant sur une publication de Berger et Milkman¹⁸, CGP Grey suggère que les théories qui se propagent le mieux sont celles qui causent de la colère, et que ces théories sont d'autant plus aptes à se propager qu'elles s'opposent à d'autres théories antinomiques qui causent tout autant de colère. Deux théories en parfaite opposition sont alors telles deux espèces animales en symbiose. Elles se nourrissent mutuellement l'une de l'autre et gagnent ensemble la totalité du territoire animal.

Autrement dit, les *fitness* des théories sur le web favorisent la bipolarisation idéologique et la haine entre les tenants des deux théories qui s'opposent, essentiellement indépendamment de leurs fondements logiques et empiriques. C'est sans doute là que se trouve l'origine de la montée de l'extrémisme idéologique depuis le début du XXI^e siècle. En particulier, de nombreuses données montrent que cela fait très longtemps que les États-Unis n'avaient plus été aussi divisés¹⁹.

¹⁴  *Chère conviction, mute-toi en infection VIRALE !!!* | Démocratie 7 | Science4All | L.N. Hoang (2017)

¹⁵  *Post-Truth: Why Facts Don't Matter Anymore* | Veritasium | D. Muller (2016)

¹⁶  *This Video Will Make You Angry* | CGP Grey (2015)

¹⁷  *Experimental evidence of massive-scale emotional contagion through social networks* | PNAS | A. Kramer, J. Guillory and J. Hancock (2014)

¹⁸  *What Makes Online Content Viral?* | Journal of Marketing Research | J. Berger and K. Milkman (2012)

¹⁹  *Partisanship and Political Animosity in 2016* | Pew Research (2016)

De façon plus générale, la prolifération de l'information brève, choquante, saisissante, énervante, séduisante, critique, intrigante, arriviste, partisane, attristante, inspirante, militante, extrémiste, diffamatoire, accusatrice et infondée semble comparable à une dangereuse tumeur cancéreuse. À l'échelle de nos sociétés, le *putaclic* semble favoriser les positions politiques tranchées et fondées sur le court terme, les sentiments d'insécurité et des espoirs injustifiés, plutôt que des visions informées, réfléchies et concernées par le long terme. Mais surtout, il rend le vote démocratique mal informé, biaisé et irrationnel.

La puissance prédictive des marchés

Dans un livre audacieux²⁰ où il ose affirmer tout haut ce que personne n'ose admettre en secret, l'économiste Bryan Caplan dresse un portrait peu flatteur de l'électeur médian. En s'appuyant sur des données empiriques des votes des Américains et sur des réponses à des sondages, Caplan conclut que l'électeur médian est pire qu'ignorant. L'électeur médian est même, selon Caplan, irrationnel. Caplan justifie même cette conclusion avec un modèle économique. Au cœur de ce modèle est la remarque selon laquelle tout vote a une probabilité quasi-nulle d'avoir un effet quelconque. Dès lors, le plaisir d'exprimer ses croyances irrationnelles au moment du vote surpassé de loin les effets hautement improbables d'un vote raisonné, d'autant qu'un vote raisonné requiert d'énormes efforts cognitifs. Autrement dit, l'électeur médian est rationnellement irrationnel²¹.

Bryan Caplan propose une alternative à l'opinion démocratique : le capitalisme et la loi des marchés. Caplan va même jusqu'à affirmer que c'est grâce aux marchés et aux lobbies que nos démocraties ne sombrent pas dans un chaos mercantile, sur-régulé et trop protectionniste. Ainsi, les grands défenseurs des droits des immigrés aux États-Unis ne seraient pas les citoyens américains ; ce seraient les Google, Facebook et autres industriels dont la puissance économique dépend très fortement de l'immigration hautement qualifiée qui occupe la grande majorité de leurs bureaux, et dont les revenus dépendent très fortement de leur image à l'international.

Caplan va même jusqu'à se faire l'avocat du très controversé *Policy Analysis Market* (PAM) mis en place par l'Agence américaine pour les projets de recherche avancée de défense (DARPA). Dans ce marché en ligne, les internautes pouvaient, par exemple, parier sur le nombre de victimes américaines en Irak, ou sur une attaque terroriste contre Israël dans l'année à venir. Comme vous pouvez l'imaginer, cette initiative gouvernementale fut vivement critiquée et rapidement renommée le « marché de la Terreur ». Le projet fut immédiatement arrêté.

 *Is America More Divided Than Ever? The Good Stuff* (2016)

²⁰  *The Myth of the Rational Voter: Why Democracies Choose Bad Policies* | Princeton University Press | B. Caplan (2011)

²¹  *Rationnellement irrationnel* Démocratie 11 | Science4All | L.N. Hoang (2017)

Pourtant, les analyses préliminaires étaient remarquablement prometteuses. De façon stupéfiante, un tel marché en ligne de paris semble drôlement efficace pour effectuer des prédictions, à l'instar des paris sur les courses de chevaux, les élections ou sur les invasions militaires. Ce que les parieurs pensent est remarquablement proche de ce qui finit par advenir. Après tout, contrairement au vote démocratique, parce qu'il y a là de l'argent en jeu, les parieurs prennent le soin de longuement s'informer et réfléchir avant de s'exprimer. Mieux encore, s'ils n'ont pas suffisamment de crédences en leurs prédictions, contrairement aux votants, les parieurs ne vont pas s'exprimer du tout. Ils éviteront ainsi de polluer les données avec des convictions mal informées, biaisées et irrationnelles.

Cependant, ramener le pouvoir prédictif des marchés à l'expertise des parieurs serait erroné. En 1988, 4 employés du *Wall Street Journal* se sont amusés à parier en bourse aléatoirement. Chaque mois, ils lancèrent des fléchettes pour choisir une action à acheter. Mois après mois, les performances des joueurs de fléchettes furent ensuite comparées avec celles de 4 investisseurs professionnels. Cent mois plus tard, les résultats furent compilés : les joueurs des fléchettes avaient battu les investisseurs 39 fois sur 100. Autrement dit, les investisseurs ont gagné de manière peu significative. Pire, plusieurs économistes affirment que si les investisseurs ont gagné, c'est uniquement parce que leurs choix d'action avaient été publiés dans le *Wall Street Journal*, créant ainsi un effet d'annonce ! Pire encore, malgré ce biais, les investisseurs n'ont battu la moyenne du marché (appelé Dow Jones) que 51 fois sur²² 100.

Le prix Nobel Daniel Kahneman a étudié les investisseurs et les marchés de près. Son constat est encore plus sévère. La corrélation entre les classements successifs des *traders* mois après mois était quasi-nulle, comme si les succès des *traders* étaient des variables aléatoires indépendantes et identiquement distribuées. Pire encore, Kahneman montra que les *traders* les plus performants étaient les moins actifs sur les marchés, comme si, pour maximiser ses gains, il suffisait de faire une confiance aveugle au marché plutôt que de chercher à le battre²³.

Toutes ces expériences semblent montrer, encore et encore, que le marché est plus compétent que n'importe lequel de ses *traders* — et beaucoup plus que n'importe lequel d'entre nous. Comment est-ce possible ?

La réponse que je propose est la même que celle que j'ai proposée pour justifier la pertinence du consensus scientifique : le marché applique la formule de Bayes mieux que n'importe quel participant du marché ! Pour comprendre cela, commençons par considérer que le cerveau de chaque *trader* est une théorie prédictive. En bons bayésiens, on va alors vouloir démultiplier la crédence en les cerveaux dont les paris passés sont justes, et réduire celle en les cerveaux dont les paris passés sont erronés. Devinez quoi ! C'est précisément ce que font les marchés !

²²  Un singe ferait-il mieux que votre conseiller financier ? Science Étonnante | D. Louapre (2013)

²³  Thinking Fast and Slow | SpringerFarrar, Straus and Giroux | D. Kahneman (2013)

Pour comprendre cela, considérons un *trader* T . Appelons $\text{fortune}(T, t)$ la fortune de T à l'instant t . De par la nature multiplicative des marchés, sa fortune à l'instant suivant est alors $\text{fortune}(T, t + 1) = \text{perf}(T, t) \cdot \text{fortune}(T, t)$. C'est exactement la même équation que les équations de Lotka-Volterra, où la variété i d'une population est remplacée par le *trader* T , et où la *fitness* de la variété est remplacée par la performance du *trader*.

En particulier, en s'attardant sur la part de marché du *trader* T , on obtient alors l'équation évolutionniste suivante :

$$\text{part}(T, t + 1) = \frac{\text{perf}(T, t)\text{part}(T, t)}{\text{perf}(T, t)\text{part}(T, t) + \sum_{A \neq T} \text{perf}(A, t)\text{part}(A, t)},$$

où A désigne les *traders* autres que T . Encore une fois, on obtient une espèce d'inférence bayésienne ! Encore une fois, la formule de Bayes est mieux appliquée par l'ensemble de la bourse que par chacun de ses membres. À l'instar du consensus scientifique, le consensus des marchés semble donc plus fiable que l'avis de chacun des experts des marchés !

Il y a toutefois trois bémols à mettre à cette analyse. Le premier est que la prédiction du marché, contrairement à une prédiction bayésienne, n'est pas la moyenne pondérée des prédictions des *traders*. En fait, le mécanisme qui transforme les prédictions des *traders* en une prédiction des marchés est davantage une médiane pondérée qu'une moyenne pondérée, dans la mesure où le prix d'équilibre divisera les *traders* en deux parties : ceux qui pensent que le prix est sous-estimé seront exactement ceux qui auront investi (et auront donc conduit à l'augmentation du prix). Reste que, de façon cruciale, plus un *trader* a de parts de marché, plus il influencera le prix d'équilibre.

Les deux autres bémols sont plus problématiques. D'un côté, il faut prendre en compte le flux continu de nouveaux *traders* dont la fortune ne vient pas de paris sur les marchés. L'exemple récent le plus spectaculaire est celui des nouveaux parieurs sur les crypto-monnaies, notamment le bitcoin, dont les investissements n'ont pas été gagnés suite à des bons paris passés. Ce flux de nouveaux *traders* agit alors comme un bonus accordé à des théories qui n'ont pas fait leurs preuves. En particulier, il gomme ainsi les échecs cuisants du passé. Il agit comme un effacement de la mémoire à long terme. Et favorise donc le court terme.

De l'autre, il y a le départ anticipé de certains *traders*, qui vient par exemple du fait que les *traders* prévoient rarement de continuer longtemps à faire ce qu'ils font — il est même devenu courant pour les jeunes cadres de ne pas rester plus de 5 ans dans leurs entreprises ! Typiquement, pour avoir une promotion, il faut avoir brillé dans les années à venir. Autant alors miser gros et prendre des risques sur le court terme.

Ces deux effets, et certainement de nombreux autres auxquels je n'ai pas pensé, nuisent beaucoup à la capacité prédictive des marchés. Ils sont typiquement à l'origine des bulles financières.

Les bulles financières

Partons aux Pays-Bas. Au début du XVII^e siècle, la tulipe devint un objet très prisé. Les Néerlandais s'arrachèrent les bulbes de tulipe. Il y eut alors rapidement une explosion de la demande sans croissance de l'offre, ce qui a conduit à une rapide augmentation du prix des bulbes de tulipe. En fait, la croissance des prix semblait inexorable, de sorte que les bons investisseurs achetaient de grandes quantités de bulbes de tulipes à un très grand prix, et pouvaient ensuite les vendre à un prix encore plus élevé. Ces investisseurs s'enrichirent et devinrent des grands acteurs des marchés financiers. En 1635, il semble que la folie de la tulipe fut telle qu'une seule tulipe pouvait désormais valoir autant qu'un manoir. C'était la tulipomanie.

Mais tout à coup, en 1637, le prix de la tulipe cessa d'augmenter. Les investisseurs de bulbes de tulipe commencèrent à craindre la chute de la valeur de leurs biens. Ils se mirent alors à écouter leurs stocks, quitte à vendre à perte. Ils baissèrent donc leur prix. Mais plus le prix chutait, plus les investisseurs voulaient vendre vite, et plus ils baissèrent leurs prix de vente. Pire, plus les prix chutaient, plus les acheteurs s'attendaient à ce que les prix continuent à chuter, et plus il était dur de trouver acheteurs, et plus les vendeurs durent baisser leur prix de vente. C'était l'implosion d'une bulle spéculative²⁴.

Ce phénomène est loin d'être unique dans l'histoire de l'humanité. Dernièrement, en 2008, la crise dite des *subprimes* a violemment frappé le marché des crédits aux États-Unis, avec des conséquences mondiales. Cette crise a été amorcée par l'implosion de la bulle spéculative du marché de l'immobilier américain. Avant la crise, les Américains n'hésitaient pas à s'endetter pour acheter des maisons, dans l'espoir que l'augmentation de la valeur de la maison rembourse largement les prêts. Et plus les Américains et leurs banques croyaient en l'augmentation des prix de l'immobilier, plus ils voulaient acheter, et plus les prix de l'immobilier augmentaient. Et plus les acheteurs voulaient s'endetter pour acheter des maisons.

Cependant, dès que les prix cessèrent d'augmenter, les stratégies de remboursement ultérieur furent remises en cause. De plus en plus de familles se mirent en défaut de paiement, et durent vendre leur maison. Mais plus il y eut de vendeurs de maisons, plus les prix des maisons se mirent à chuter, et plus il y eut de défauts de paiement. Ce cercle vicieux amplifia le phénomène.

Pire encore, les prêts bancaires étaient devenus de complexes produits dérivés revendus à des investisseurs de Wall Street. Or ces derniers n'ont rien vu venir. C'est ainsi que les nombreux défauts de paiements des particuliers américains se transformèrent en gouffres financiers pour les grandes firmes d'investissement, dont l'effondrement entraîna celui de nombreuses autres entreprises à leur tour. Ce fut alors un désastre mondial²⁵.

²⁴  *What causes economic bubbles?* | Ted-Ed | P. Singh (2015)

²⁵  *Les Subprimes 1ère Partie : La boulette !* Heu?reka | G. Mitteau (2017)

On en arrive là à une limite fondamentale de la capacité des marchés à prédire. L'entrée en jeu fréquente de nouveaux *traders* et le départ régulier d'autres *traders* conduisent à la divergence entre la dynamique des marchés et la formule de Bayes. En particulier, ces caractéristiques des marchés favorisent le court terme. Voilà qui est incompatible avec le temps plus lent des politiques et des bulles spéculatives. En termes statistiques, le temps accéléré des marchés les voue à sur-interpréter leur passé immédiat. C'est pourquoi les prédictions à long terme des marchés ne méritent pas les mêmes crédences que le consensus scientifique.

Références en français

- ⌚ *Un singe ferait-il mieux que votre conseiller financier ?* | Science Étonnante | D. Louapre (2013)
- ▶ *Les arguments fallacieux* | Hygiène Mentale | C. Michel (2016)
- ▶ *Les OGMs sont-ils nocifs ? (non)* | Dirty Biology | L. Grasset (2016)
- ▶ *Biotope et Équilibre Proies - Prédateurs* | Goana (2017)
- ▶ *Les Subprimes 1ère Partie : La boulette !* Heu?reka | G. Mitteau (2017)
- ▶ *Les Subprimes 2ème Partie : Une crise imprévisible* | Heu?reka | G. Mitteau (2017)
- ▶ *Les Subprimes 3ème Partie : Ceux qui ont prédit la crise* | Heu?reka | G. Mitteau (2017)
- ▶ *4 paradoxes de la logique mathématique* | Infini 17 | Science4All | L.N. Hoang (2017)
- ▶ *Petit communautarisme deviendra grand* | Démocratie 6 | Science4All | L.N. Hoang (2017)
- ▶ *Chère conviction, mute-toi en infection VIRALE !!!* Démocratie 7 | Science4All | L.N. Hoang (2017)
- ▶ *Rationnellement irrationnel* Démocratie 11 | Science4All | L.N. Hoang (2017)
- ▶ *Le paradoxe de la morale* | Démocratie 25 | Science4All | S. Debove et L.N. Hoang (2017)

Références en anglais

- 📚 *Designing Effective Multimedia for Physics Education* | University of Sydney | PhD Thesis | D. Muller (2008)
- 📚 *The Myth of the Rational Voter: Why Democracies Choose Bad Policies* | Princeton University Press | B. Caplan (2011)
- 📚 *Thinking Fast and Slow* | SpringerFarrar, Straus and Giroux | D. Kahneman (2013)

- ☒ *The theory of games and the evolution of animal conflicts* | Journal of Theoretical Biology | J.M. Smith (1974)
 - ☒ *What Makes Online Content Viral?* | Journal of Marketing Research | J. Berger and K. Milkman (2012)
 - ☒ *Experimental evidence of massive-scale emotional contagion through social networks* | PNAS | A. Kramer, J. Guillory and J. Hancock (2014)
 - ☒ *Partisanship and Political Animosity in 2016* | U.S. Politics & Policy | Pew Research Center (2016)
-
- ▶ *How not to be ignorant about the world* | TED | H. Rosling et O. Rosling (2014)
 - ▶ *How I use "meta-updating"* | J. Galef (2015)
 - ▶ *Rock Paper LIZARDS* | Numberphile | H. Fry (2015)
 - ▶ *Khan Academy and the Effectiveness of Science Videos* | Veritasium | D. Muller (2011)
 - ▶ *This Video Will Make You Angry* | CGP Grey (2015)
 - ▶ *Post-Truth: Why Facts Don't Matter Anymore* | Veritasium | D. Muller (2016)
 - ▶ *What causes economic bubbles?* | Ted-Ed | P. Singh (2015)
 - ▶ *Is America More Divided Than Ever?* | The Good Stuff (2016)
 - ▶ *That Time Tulips Crashed the Economy (Maybe)* | The Good Stuff (2018)
-
- 🌐 *Partisanship and Political Animosity in 2016* | Pew Research (2016)
 - 🌐 *Evolutionary Game Theory* | Science4All | L.N. Hoang (2012)
 - 🌐 *Numbers and Constructibility* | Science4All | L.N. Hoang (2013)

Une analyse de l'histoire des technologies montre que les changements technologiques sont exponentiels, contrairement au sens commun qui correspond à « l'intuition linéaire ». Nous n'allons donc pas faire l'expérience de 100 années de progrès au cours du XXI^e siècle — ce sera plutôt de l'ordre de 20 000 années de progrès (au rythme d'aujourd'hui).

Ray Kurzweil (1948-)

11

Exponentiellement contre-intuitif

Les nombres archi-méga-super géants

« Une croissance linéaire est 1, 2, 3. Une croissance exponentielle est 1, 2, 4. Ça n'a pas l'air différent [...]. Mais à la 30-ième étape, la croissance linéaire, et c'est ça notre intuition, en est à 30. La progression exponentielle en est à un milliard », explique le futuriste Ray Kurzweil. « Notre intuition est linéaire, mais la réalité des technologies de l'information est exponentielle, et ça fait une différence profonde. »

Pour Kurzweil, notre mécompréhension de la croissance exponentielle nous conduit, peut-être à légèrement surestimer l'influence des nouvelles technologies sur le court terme, mais indubitablement à profondément la sous-estimer sur le long terme — à l'échelle de 5 ans ou plus. Mais avant d'en arriver là, il est bon de d'abord goûter à l'immensité des très grands nombres.

Partons d'un nombre à l'apparence raisonnable. Un million est un nombre auquel notre quotidien nous confronte si régulièrement qu'il est tentant de penser qu'on en comprend l'immensité. En novembre 2016, Dr Nozman fut ainsi le premier YouTuber scientifique francophone à franchir la barre du million d'abonnés. Les salaires des joueurs de football se comptent en millions. La population d'un pays est souvent de l'ordre du million.

Cependant, il ne faut pas confondre notre familiarité avec ce nombre avec la compréhension que l'on en a. Un million, c'est beaucoup. Un homme aurait bien du mal à compter jusqu'à un million en une année, même au rythme in-

imaginable de presque un nombre par seconde. Un million est un nombre que l'on peut penser, mais auquel il est impossible d'accéder en partant de zéro.

Et pourtant, un million est un grain de sable devant l'immensité des nombres. Un milliard est déjà énormément plus gigantesque ! Les plus riches gagnent des milliards par an. Ceci représente cent mille fois mon salaire ! Pour se rendre compte de l'immensité de ces nombres, on peut souligner qu'un euro pour moi correspond, proportionnellement, à cent mille euros pour ces milliardaires. Ainsi, de la même manière qu'il m'arrive de ne pas prendre la peine d'aller ramasser une pièce d'un euro qui traîne dans la rue, le milliardaire ne prendrait pas la peine d'acquérir un bien de cent mille euros qui lui est offert s'il lui faut lever le petit doigt pour ce faire !

On peut voir les choses autrement. Certaines études montrent qu'à partir de 70 000 euros par an, l'argent ne fait pas le bonheur. Ou plutôt, quelqu'un qui gagne plus de 70 000 euros par an n'est statistiquement pas plus heureux que quelqu'un qui gagne 70 000 euros¹. Imaginons maintenant qu'un milliardaire apprend cela et décide de dépenser exactement 70 000 euros par an. Un milliard d'euros lui permettrait alors de vivre 14 000 ans sans travailler ! Ou vu encore autrement, s'il gagne un milliard d'euros par an, il peut alors se permettre de garantir l'apport financier suffisant à un bonheur maximal à 14 000 personnes — lui-même inclus.

Mais un milliard reste un nombre ridiculement faible à l'échelle de la physique. Notre galaxie contient ainsi des centaines de milliards d'étoiles, nos cerveaux sont composés des millions de milliards de connexions neuronales, notre Terre possède des milliards de milliards de grains de sable et une goutte d'eau est un agglomérat de millions de milliards de milliards de molécules. Pour des nombres aussi grands, on préfère utiliser des notations comme 10^{25} , qui est le nombre décrit par un 1 suivi de 25 zéros. Ces nombres sont littéralement astronomiques.

Cependant, ces nombres ne sont qu'astronomiques. Et si la physique moderne permet de rencontrer des nombres encore plus grands, les puissances de 10 suffisent à atteindre les limites de la physique. Par exemple, selon certaines théories modernes de la physique, le temps serait discret et s'écoulerait à pas de temps de Planck d'environ 10^{-44} secondes. Du coup, il ne se serait déroulé que 10^{60} pas de temps élémentaires depuis le Big Bang. Par ailleurs, on ne peut compter environ que 10^{80} atomes dans l'univers observable².

Le plafond de verre des calculs

Les limites de la physique se traduisent alors nécessairement en limites de nos puissances de calcul. Ainsi, de nos jours, les informaticiens considèrent souvent

¹  *L'argent fait-il le bonheur ?* Stupid Economics | V. Levetti et A. Gantier (2017)

²  *Les nombres archi-méga-super géants* | Infini 1 | Science4All | L.N. Hoang (2016)

qu'un algorithme qui nécessite plus de 10^{70} étapes de calculs ne terminera jamais dans l'histoire de l'humanité. Après tout, ce nombre d'étapes de calculs est supérieur au nombre de pas de temps élémentaires depuis le Big Bang.

Ce chiffre peut être déduit d'autres hypothèses physiques. Le principe de Landauer, qui se déduit des équations de Boltzmann, postule l'existence d'une limite physique à l'énergie minimale nécessaire pour effectuer un bit d'opération irréversible. Cette énergie est de 10^{-21} joules par bit d'information à température ambiante. Or, l'énergie du système solaire est finie. On peut même l'estimer à 10^{47} joules. Par conséquent, le nombre d'opérations irréversibles qui peuvent être effectuées grâce à toute l'énergie du système solaire n'est que 10^{68} .

C'est sur cette hypothèse de plafond de verre des calculs que s'appuient les cryptologues pour sécuriser nos communications. Ils supposent que les protocoles utilisés en pratique nécessitent plus de 10^{70} étapes de calculs pour être cassés. S'ils voient juste, ceci signifierait que les messages encodés aujourd'hui avec leurs technologies ne seront non seulement pas craqués demain ; ils ne seront même jamais craqués par l'humanité, garantissant ainsi parfaitement la confidentialité des données chiffrées.

Bien entendu, il pourrait encore être possible de hacker ces données en sondant les bonnes personnes ou leurs connaissances — une technique appelée *social engineering* qui est souvent la principale cause de l'insécurité informatique. Pire encore, les capacités des algorithmes sont encore mal comprises et l'on ne peut pas exclure aujourd'hui l'existence d'algorithmes efficaces pour court-circuiter les 10^{70} opérations que les cryptologues pensent nécessaires — il s'agit même là du problème P versus NP, le plus prestigieux problème ouvert de l'informatique théorique. En fait, depuis 1997 et la découverte de Peter Shor, on sait même que si des ordinateurs quantiques avec des espaces mémoires quantiques suffisamment grands peuvent être construits, alors il sera possible d'utiliser un algorithme quantique pour hacker bon nombre de protocoles cryptographiques utilisés aujourd'hui, y compris le protocole RSA.

Détaillons le protocole RSA. RSA est une cryptographie dite *asymétrique*. Ceci signifie que l'utilisateur de RSA possède deux clés qui fonctionnent en symbiose. La première est une clé publique e , la seconde est une clé privée d . De façon cruciale, étant donné la clé publique e , l'utilisateur peut choisir secrètement deux nombres premiers p et q pour calculer efficacement une clé privée compatible³ d . Cependant, pour que d'autres puissent lui envoyer des messages chiffrés, l'utilisateur doit aussi rendre public le produit $N = pq$ de ses deux nombres premiers. Du coup, si un hacker arrive à décomposer le nombre N en le produit de deux nombres entiers p et q , il pourra suivre les pas de l'utilisateur, et ainsi calculer efficacement la clé privée d à partir de la clé publique e . Il aura cassé le protocole RSA.

³Plus précisément, il doit déterminer d tel que $ed \equiv 1 \pmod{(p-1)(q-1)}$, ce qui peut être fait rapidement grâce à l'algorithme d'Euclide et au calcul $\text{pgcd}(e, (p-1)(q-1))$. Je passe certains détails sous silence.

Toute la sécurité de RSA repose alors sur l'hypothèse selon laquelle un utilisateur ne pourra pas décomposer l'entier N en un produit de nombres premiers. Il s'agit du problème de factorisation. Aujourd'hui, on ne sait pas s'il existe un algorithme (classique) rapide de factorisation. Et pour la sécurité de RSA, faute de preuve de non-existence, on ne peut que prier pour que personne ne trouve un tel algorithme. Pire, l'algorithme quantique de Shor résout rapidement le problème de la factorisation. Or, il semblerait qu'il ne soit qu'une question de temps avant que des ordinateurs quantiques dignes de ce nom voient le jour...

La difficulté de la factorisation peut toutefois sembler étonnante. On pourrait croire qu'il suffit d'essayer de diviser N par tous les nombres a qui lui sont inférieurs. Il est même possible de prouver qu'il suffit de tester la division par tous les nombres entre 2 et \sqrt{N} . Cependant, le nombre N utilisé en pratique est un nombre gigantesque, qui va typiquement être de l'ordre de 10^{300} . Mais le nombre de divisions à tester est alors \sqrt{N} , qui fait environ 10^{150} . Or, on l'a vu, on peut considérer que 10^{70} opérations ne peuvent pas être effectuées dans notre monde physique.

L'explosion exponentielle

Le plafond de verre de tout calcul semble être une limite inatteignable. Elle est en effet physiquement inatteignable pour des croissances linéaires. Cependant, de manière extrêmement contre-intuitive, ces ordres de grandeur sont en fait rapidement atteints par l'étonnante croissance exponentielle.

Selon la légende, le roi Belkib d'Inde adora le jeu des échecs que le sage Sissa lui avait présenté. Il proposa donc au sage de choisir lui-même sa récompense. Sissa répondit humblement qu'il serait heureux avec 1 grain de riz pour la première case le premier jour, 2 pour la suivante au deuxième jour, 4 pour celle d'après, 8 ensuite, et ainsi de suite. Le roi, surpris par la modestie de la requête, accepta. Quelle erreur ! Après 64 jours, le roi avait une dette de plusieurs milliards de milliards de grains de riz, ce qui représente environ mille fois la production mondiale annuelle de riz d'aujourd'hui ! Autant dire que le roi Belkib eut une dette éternelle envers le sage Sissa⁴.

La croissance exponentielle de la dette du roi Belkib a de quoi surprendre. De la même manière, il suffit de plier une feuille de papier 42 fois pour que son épaisseur fasse la hauteur de la distance Terre-Lune, et 103 fois pour qu'elle fasse la largeur de notre univers observable ! Or, l'univers observable est très, très grand. Il fait presqu'un million de milliards de milliards de kilomètres de large ! Et pourtant, en seulement 103 étapes, la croissance exponentielle a dépassé les limites de l'astrophysique !

On retrouve cette croissance folle dans l'arbre généalogique. En effet, puisque chacun d'entre nous est l'enfant de deux parents (biologiques), le nombre de nos

⁴  *La Légende de Sessa* | Scientificfiz (2017)

ancêtres croît exponentiellement quand on remonte l'histoire. Ainsi, à l'aide de simulations informatiques qui tiennent compte des déplacements géographiques des civilisations du passé, Rohde, Olson et Chang estiment⁵ que tous les humains vivants aujourd'hui descendent d'un même ancêtre ayant vécu il y a entre 2 000 et 5 000 ans⁶. Oui, nous sommes tous consanguins et cousins germains à quelques centaines de degré ! Mieux encore, presque tout humain ayant vécu avant ce plus récent ancêtre commun est soit un ancêtre de tous les humains aujourd'hui vivants, soit l'ancêtre d'aucun d'entre nous⁷ !

De même, en sciences, on dit souvent de nos directeurs de thèse qu'ils sont nos pères et nos mères académiques, et les parents académiques de nos parents académiques sont nos grands-parents académiques. Mieux encore, on peut remonter la généalogie académique et déterrer nos ancêtres académiques. Si Mickaël Launay est un descendant académique de G.H. Hardy, Isaac Newton et autre Galilée, je suis de mon côté un descendant de Georg Dantzig, Carl Friedrich Gauss et Leonhard Euler. J'ai même découvert descendre de Jerzy Neyman (on ne choisit pas ses ancêtres !) et de Pierre-Simon Laplace (avec une fierté irrationnelle !). Cependant, la présence de grands noms dans nos généalogies n'a rien de surprenant. D'une part, le nombre de mathématiciens il y a plusieurs siècles était très res-treint. D'autre part, parce que certains ont plusieurs parents académiques, le nombre d'ancêtres croît exponentiellement.

Si la croissance exponentielle est stupéfiante, sa décroissance l'est tout autant. En divisant un morceau de sucre en deux une soixantaine de fois, on finit par devoir diviser les molécules de sucre. De la même manière, une dilution homéopathique divise par environ 100 le nombre de molécules d'une substance active. Et bien, malgré le nombre astronomique de molécules initiales, il suffit de 12 dilutions pour garantir statistiquement la disparition de toute molécule de la substance active ! Et tout ça vient de l'incroyable décroissance exponentielle des concentrations en molécules, dilution après dilution.

Pour beaucoup, l'émergence de la complexité du vivant à partir de (presque) rien semble irréaliste. Mais notre inaptitude à concevoir la complexification des espèces animales par sélection naturelle est sans doute liée à notre intuition linéaire du monde. Or le vivant, lui, se démultiplie, comme on l'a vu au moment de parler des équations de Lotka-Volterra. Lors d'une mitose, une cellule se divise en deux cellules, puis chaque cellule fille se divise en deux autres, et ainsi de suite. À chaque étape, le nombre de cellules est multiplié par 2. C'est ainsi qu'en l'espace de quelques jours ou de quelques mois, une unique cellule œuf peut s'être transformée en un organisme vivant complexe composé de milliers de milliards de cellules.

⁵  *Modelling the recent common ancestry of all living humans* | Nature | D. Rohde, S. Olson, et J. Chang (2004)

⁶ Il ne faut pas confondre ce plus récent ancêtre commun avec d'autres concepts comme l'Ève mitochondriale, la mère des mères. En effet, le nombre de mères des mères ne croît pas exponentiellement — l'Ève mitochondriale est donc nécessairement beaucoup plus ancienne.

⁷  *Vous êtes de sang royal* | Dirty Biology | L. Grasset (2018)

Certes, la croissance exponentielle de la complexification du vivant par variations et sélection naturelle est beaucoup plus « lente », notamment comparée à la croissance des technologies. Cependant, celle-ci s'est prolongée sur des échelles de temps qui nous sont impossibles à concevoir, comme le milliard d'années. Ainsi, si elle produit peu de changements perceptibles à l'échelle du siècle, il faut bien se rendre compte que des dizaines de millions de siècles se sont écoulés depuis l'apparition des premières cellules vivantes. L'immensité de ces quantités dépasse déjà notre entendement. Une croissance exponentielle sur de telles durées le dépasse encore plus !

Affirmer que l'évolution darwinienne n'a pas eu le temps de générer la complexité du vivant sans modèle mathématique, c'est faire reposer nos conclusions sur notre intuition des grands nombres et de la croissance exponentielle. Cette intuition étant profondément erronée, elle ne mérite pas nos crédences. Et les conclusions tirées de cette intuition non plus.

Un autre exemple permet d'illustre cela. En ce moment, la population humaine augmente chaque année de 11 %. Cette croissance est non-négligeable, mais elle ne semble pas déraisonnable. Et pourtant. Un rapide calcul⁸ montre qu'avec une telle croissance, d'ici 8 604 ans, la population humaine sera telle que le nombre de particules qui composent les individus humains excédera le nombre total de particules dans l'univers ! Une croissance exponentielle peut être imperceptible sur le très court terme, et néanmoins envahir tout l'univers à pas si long terme.

À l'inverse, toutefois, il suffit que le nombre d'enfants par femme devienne inférieur à 2 (ce qui est déjà le cas dans bon nombre de pays développés) pour que le nombre total d'humains dans l'histoire de l'univers soit étonnamment restreint ! Ainsi, à 1,9 enfant par femme, la population totale de toute l'humanité à travers tous les âges, passés et futurs, sera uniquement de l'ordre de la centaine de milliards. La surpopulation sera alors un problème limité, même à supposer que les biologistes arrivent à nous rendre immortels⁹ !

Pour sentir la folie de la croissance exponentielle, je recommande vivement le jeu en ligne *Universal Paperclips*¹⁰. Attention toutefois. Ce jeu est terriblement addictif. Il illustre la fable de Nick Bostrom au sujet d'une intelligence artificielle qui maximisera le nombre de trombones. Pour atteindre son objectif, cette intelligence artificielle investirait massivement dans la recherche scientifique, ce qui lui permettrait d'accélérer le rythme de ses avancées technologiques. Plus elle serait technologiquement avancée, plus elle progresserait vite. Sa croissance serait exponentielle. Et cette croissance est effrénée. Après seulement deux jours de jeu, en partant d'une simple production de trombones individuels, *Universal Paperclips* nous amène inéluctablement à la conquête et à l'invasion de tout l'univers.

⁸  *How many particles in the univers?* | Numberphile | T. Padilla (2017)

⁹  *Immortalité = surpopulation... ou pas ?* Alexandre Technoprog | A. Maurer (2017)

¹⁰ decisionproblem.com/paperclips/

La magie des chiffres indo-arabes

Revenons-en à la cryptographie. On a vu que les cryptologues utilisaient quotidiennement des nombres $N \approx 10^{300}$ si grands que \sqrt{N} étapes de calcul étaient un processus qui transcende les limites physiques. Comment est-ce possible de parler d'un nombre si grand qu'il échappe à la physique ?

Pendant longtemps, les grands empires de l'humanité, comme les empires égyptiens et romains, avaient une manière peu efficace de représenter les nombres. Ainsi, le nombre 1 888 s'écrit, en chiffres romains, MDCCCLXXXVIII. Pire encore, les Romains étaient contraints d'inventer des nouveaux symboles pour les nombres toujours plus grands, si bien qu'ils étaient incapables d'écrire des nombres comme un million — à moins d'aligner mille M à la suite !

À l'inverse, les Babyloniens, suivis des Chinois et des Japonais, mais aussi et surtout des Indiens et des Arabes, ont eu la brillante idée de développer le système de notation positionnelle. La particularité majeure de ce système est le fait que la position d'un symbole détermine sa valeur numérique. Ainsi, comme on le sait tous, 12 est un nombre différent de 21, même si les symboles utilisés sont les mêmes.

Je ne vais pas décrire comment fonctionne le système numérique indo-arabe. Vous devriez l'avoir appris dès votre plus jeune âge. Cependant, ce dont vous ne vous étiez peut-être pas rendu compte, c'est de la concision remarquable que ce système numérique permet. Bien qu'il utilise un nombre restreint de symboles (appelés chiffres), le nombre de chiffres requis pour représenter un nombre est beaucoup plus faible que le nombre lui-même. Ainsi, 10^{100} est un nombre qui transcende les limites de la physique. Cependant, notre système de numération est si efficace qu'il permet de le représenter avec seulement 101 chiffres — un « 1 » suivi de cent « 0 ».

Notre système de numération est même optimal en un certain sens. En effet, il n'est pas possible de représenter plus succinctement tous les nombres entre 0 et $10^{100} - 1$. Après tout, toutes les combinaisons des 100 chiffres ont été utilisées pour représenter tous ces nombres. En fait, la taille des nombres est exponentielle en la longueur de leurs représentations. Autrement dit, tout nombre entier x est à peu près égal à $10^{\#\text{chiffres de } x}$. Une façon équivalente de dire cela consiste à dire que les nombres ont des représentations de taille logarithmique. Le logarithme d'un nombre est approximativement le nombre de chiffres nécessaires pour représenter ce nombre. Ainsi $\log_{10}(10^{100}) = 100$.

À l'inverse de la croissance exponentielle, la croissance logarithmique est incroyablement lente. Par exemple, le temps qu'il faut pour chercher un élément dans un tableau trié est logarithmique. Ceci veut dire que, même si le tableau en question fait la taille de l'univers, alors il sera possible de trouver l'élément en question en quelques centaines d'itérations — on ne serait limité que par la finitude de la vitesse de la lumière !

Cette découverte est aussi au cœur de l'une des notions les plus fondamentales de l'informatique, la notion d'adresse. Imaginez que vous vouliez trouver une information sur le web. De façon étonnante, alors qu'il y a des exaoctets, voire des zettaoctets de données sur le web, armé de l'adresse dite URL de l'information que vous cherchez, il vous sera possible de trouver quasiment instantanément l'information en question !

Qui plus est, l'URL de l'information est incroyablement brève. Sa taille est logarithmique en la taille de tout le web ! En particulier, l'URL entière peut aisément être gardée en mémoire — contrairement à l'information à laquelle cette URL fait référence¹¹. On reviendra d'ailleurs sur cette notion fondamentale d'adresse au moment où on abordera la gestion de la mémoire par le *bayésien pragmatique*.

De manière plus précise, le logarithme, comme la fonction exponentielle, dépend d'un paramètre appelé la base. La suite 1, 2, 4 est une croissance exponentielle de base 2, puisqu'à chaque itération, le facteur par lequel on multiplie les éléments de la suite est 2. Ou dit autrement, le $(n + 1)$ -ième élément de la suite s'écrit 2^n . À l'inverse, le logarithme d'un nombre x en base 2 compte le nombre de fois qu'il faut multiplier par 2 pour atteindre le nombre x . Ainsi, $\log_2(2^n) = n$, puisque 2^n s'obtient en multipliant n fois par le nombre 2. Quand x n'est pas une puissance de 2, le logarithme en base 2 va trouver une façon naturelle et astucieuse de déterminer un nombre réel y tel que $2^y = x$.

La loi de Benford

Ouvrez la page Wikipedia qui liste les pays du monde ordonnés selon leurs populations. Notez les premiers chiffres de ces nombres d'habitants. Vous devriez remarquer avec stupéfaction quelque chose de troublant : ces premiers chiffres des nombres d'habitants sont plus souvent des « 1 » que des « 9 » ! Cette remarque bluffante n'est d'ailleurs pas spécifique aux nombres d'habitants d'un pays. Considérez les longueurs des fleuves, les nombres d'abonnés des chaînes YouTube ou les salaires des multi-millionnaires. Dans tous ces cas, le premier chiffre le plus fréquent sera le « 1 », et il sera environ 6 à 7 fois plus fréquent que le « 9 » ! Telle est la surprenante loi de Benford¹².

L'origine de la loi de Benford réside dans une propriété récurrente de nombreux systèmes. À l'instar de l'équation de Lotka-Volterra dont on a parlé dans le chapitre précédent, bien des systèmes dynamiques conduisent à des réactions en chaîne qui caractérisent la croissance exponentielle. Par exemple, la croissance des nombres d'abonnés YouTube est typiquement exponentielle. Ceci signifie qu'elle double par exemple tous les six mois. Mais alors, pendant six mois, le nombre d'abonnés sera entre 1 000 et 2 000. Puis, les six mois suivants, il sera

¹¹  *DNA Encoding* | ZettaBytes | C. Dessimoz (2018)

¹²  *La loi de Benford* | Passe-Science | T. Cabaret (2015)

entre 2 000 et 4 000. Puis, les six mois suivants, entre 4 000 et 8 000, et six mois entre 8 000 et 16 000. On voit là apparaître l'origine de la loi de Benford. Le nombre d'abonnés sera alors entre 1 000 et 2 000 pendant 6 mois, alors qu'il sera moins d'un mois entre 9 000 et 10 000. Voilà qui suggère que le nombre d'abonnés commence six fois plus souvent par un « 1 » que par un « 9 ».

Pour mieux comprendre ces croissances exponentielles, il est utile de changer l'échelle que l'on étudie. Plutôt que d'étudier le nombre d'abonnés, on pourrait vouloir étudier le logarithme en base 2 du nombre d'abonnés. Rappelons que $1024 = 2^{10}$ et $2048 = 2^{11}$. Du coup, en six mois, le logarithme en base 2 du nombre d'abonnés passe environ de 10 à 11. Six mois plus tard, il passe à 12, puis à 13, et ainsi de suite. Vous voyez ce qu'il se passe ? Tous les six mois, le logarithme en base 2 du nombre d'abonnés augmente tout simplement d'une unité.

Mais alors, sur l'échelle logarithmique que l'on vient de construire, le logarithme du nombre d'abonnés passe autant de temps entre 10 et 11 qu'entre 11 et 12 et qu'entre 12 et 13. Supposons désormais que l'on observe les nombres d'abonnés de différentes chaînes prises à des moments distincts de leur croissance. On peut alors s'attendre à ce que les logarithmes de ces nombres d'abonnés sont tout aussi souvent entre 10 et 11 qu'entre 13 et 14. En termes techniques, la distribution des logarithmes des nombres d'abonnés est à peu près uniforme¹³. Telle est la condition technique à la validité de la loi de Benford : si l'échelle naturelle d'une quantité est l'échelle logarithmique (et si elle s'étend sur plusieurs ordres de grandeur), alors les premiers chiffres de cette quantité seront six fois plus souvent des « 1 » que des « 9 »¹⁴.

L'échelle logarithmique

La physique et la chimie adorent ces échelles logarithmiques. Il est quasiment impossible de dessiner le système solaire à l'échelle, parce que les tailles des planètes sont ridiculement faibles comparées aux distances qui les séparent, qui sont elles-mêmes quasiment nulles comparées à la taille de notre galaxie, qui est elle-même inexisteante vis-à-vis des distances intergalactiques, qui, elles, ne sont rien devant l'immensité de notre univers observable ! À l'inverse, les échelles microscopiques, nanoscopiques, voire subatomiques s'étalent sur des ordres de grandeur tels qu'il est impossible de représenter à la fois les constituants des protons des noyaux des atomes et les molécules qui les combinent. Pour penser toutes ces échelles à la fois, l'échelle logarithmique est incontournable.

¹³En fait, l'étude de David Louapre montre que les logarithmes des nombres d'abonnés d'une sélection de chaînes YouTube scientifiques sont à peu près distribués selon une loi normale.

De quoi le succès d'une chaîne Youtube de vulgarisation dépend-il ? Science Étonnante | D. Louapre (2017)

¹⁴La formule rigoureuse s'écrit $\mathbb{P}[\text{1er chiffre} = d] = \log_{10}(d+1) - \log_{10}(d)$. Les probabilités que ce chiffre soit 1,2,3,4,5,6,7,8,9 sont environ 30 %, 18 %, 13 %, 10 %, 8 %, 7 %, 6 %, 5 %, 5 %.

De la même manière, les volumes sonores, les magnitudes sismiques et l'acidité des solutions sont habituellement mesurés sur des échelles logarithmiques. On parle alors de décibels, d'échelle de Richter et de pH. Ces unités usuelles sont en fait égales (parfois à un signe près) aux logarithmiques des amplitudes des variations de pression, de l'énergie des ondes sismiques et de la concentration en ions H^+ .

Ces échelles logarithmiques révèlent la nature multiplicative des objets d'étude. Cette interprétation additive des différences de mesure nous est en fait très familière pour traiter de nombreux cas. Prenons un exemple. Typiquement, on a envie de dire que la différence entre une chaîne à 200 000 abonnés et une chaîne à 1 million d'abonnés est moindre que celle entre la chaîne à 200 000 abonnés et une chaîne à 50 abonnés. Cependant, cette phrase n'a aucun sens sur une échelle additive. En effet, la différence dans le premier cas est alors de 800 000 abonnés. Par opposition, dans le second cas, la différence n'est « que » de 199 950 abonnés.

Mais à l'inverse, notre intuition est parfaitement en phase avec l'échelle multiplicative. En effet, pour passer de 200 000 à 1 million, il suffit de multiplier le nombre d'abonnés par 5. Cependant, passer de 50 à 200 000 requiert de multiplier le nombre d'abonnés par 4 000.

Ce qu'affirme là l'échelle multiplicative est d'ailleurs équivalent à ce que révèle l'échelle logarithmique. Ainsi, la différence entre 200 000 et 1 million sur l'échelle logarithmique est égale à $\log_2(1\text{million}) - \log_2(200\,000) \approx 2,3$, alors que la différence entre 50 et 200 000 sur cette échelle logarithmique est environ 12. C'est pour cette raison que ces échelles logarithmiques sont si souvent utilisées pour représenter et comparer les objets dont les variations sont multiplicatives plutôt qu'additives.

Curieusement, les jeunes enfants et les autochtones semblent avoir une intuition additive des nombres plutôt qu'additives. Quand on leur demande de placer les nombres de 1 à 10 sur une échelle, ces personnes espacent davantage les premiers chiffres, et rapprochent les derniers. C'est ce à quoi ressemble l'échelle logarithmique — même si l'échelle de ces personnes n'est en fait pas tout à fait l'échelle logarithmique.

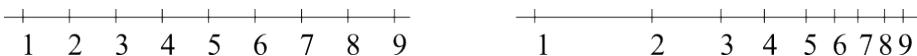


Figure 11.1. Échelle additive *versus* échelle logarithmique.

À l'inverse, toute personne formatée par l'éducation mathématique espacera les nombres de manière régulière, se plaçant ainsi dans une échelle additive. Notre intuition des nombres est fortement influencée par ce que l'on en a appris. L'école nous a appris à penser en mode additif. Elle nous a appris à laisser tomber notre intuition multiplicative. Pourtant, l'échelle additive n'a rien de plus naturel que l'échelle multiplicative (ou logarithmique).

Le logarithme

Comme Mickaël Launay l'explique très bien sur sa chaîne¹⁵, les échelles additives et multiplicatives semblent appartenir à deux mondes distincts. Le premier est celui des additions, des soustractions, des moyennes arithmétiques et des intégrales. C'est le monde linéaire. Le second est celui des multiplications, des divisions, des moyennes géométriques et des équivalences asymptotiques. C'est le monde des nombres d'abonnés YouTube, celui des intensités des secousses sismiques et celui de l'évolution darwinienne. Pour Ray Kurzweil, c'est surtout aussi le monde du progrès technologique. Et pour nous, c'est avant tout le monde de la formule de Bayes !

D'un point de vue mathématique, les deux mondes ne sont cependant pas sans lien. Il y a même deux intermédiaires, deux sortes de traducteurs, qui font le lien entre ces deux mondes. Ces intermédiaires, on en a déjà parlé. Il s'agit des fonctions logarithmes et exponentielles. La fonction exponentielle traduit les objets du monde additif en des objets du monde multiplicatif. En particulier, l'addition entre nombres deviendra alors une multiplication entre leurs exponentielles. À l'inverse, la fonction logarithme traduit les objets du monde multiplicatif en objets du monde additif. Cela veut dire que le logarithme permet de ramener des objets et des opérations qui nous sont étrangers, au monde qui nous est le plus familier¹⁶.

Parce qu'additionner est plus simple que multiplier, ces traducteurs ont joué un rôle central dans le calcul numérique avant l'avènement des calculatrices. Pour multiplier a et b , il y a quelques décennies, les élèves et les scientifiques commençaient par utiliser des tables (ou des règles) logarithmiques pour déterminer les logarithmes de a et de b . Puis, ils additionnaient les logarithmes. Enfin, ils utilisaient l'exponentielle pour traduire le résultat de l'addition vers le monde multiplicatif, et obtenir ainsi le résultat escompté¹⁷. Cette approche peut sembler alambiquée. Pourtant, c'était ce qu'il y avait de mieux pour effectuer des multiplications sophistiquées rapidement et avec peu d'erreurs.

De même, parce qu'additionner est plus simple que multiplier, Alan Turing utilisa le logarithme pour effectuer ses calculs bayésiens en temps de guerre. L'unité additive obtenue en traduisant les probabilités bayésiennes sur l'échelle additive était (grossièrement) ce que Turing appela les *banburismus*¹⁸ et, comme on le verra plus loin, ils sont intimement liés aux fameux *bits* de Shannon — et donc à des notions comme l'entropie et la divergence KL.

¹⁵  *Addition contre multiplication* | MicMaths | M. Launay (2014)

¹⁶ Formellement, on a $2^{m+n} = 2^m \cdot 2^n$ et $\log(xy) = \log(x) + \log(y)$.

¹⁷ Autrement dit, on calcule $ab = 10^{\log_{10}(a)+\log_{10}(b)}$.

¹⁸ Turing s'est en fait intéressé uniquement aux probabilités relatives, aussi appelées cotes ou *odds* en anglais. Ainsi les *banburismus* de Turing étaient les unités de quantités de la forme $\log_{10}(\mathbb{P}[T_1|D]/\mathbb{P}[T_2|D]) = \log_{10}\mathbb{P}[T_1|D] - \log_{10}\mathbb{P}[T_2|D]$ (ce qui revient à une différence de *bits* de Shannon, à un facteur multiplicatif près, puisque les *bits* utilisent la base 2). Voilà qui permet à Turing de court-circuiter le calcul de la fonction de partition.

Pour l'heure, je vous propose de simplement écrire la formule de Bayes sur l'échelle logarithmique :

$$\log \mathbb{P}[T|D] = \log \mathbb{P}[D|T] + \log \mathbb{P}[T] - \log \mathbb{P}[D].$$

C'est souvent cette version de la formule de Bayes que les chercheurs en intelligence artificielle préfèrent étudier ; et elle a aussi des applications en physique statistique et en sciences cognitives !

Bayes rafle un prix Gödel

Ce n'est d'ailleurs que très récemment que les informaticiens ont compris comment exploiter la folie des croissances exponentielles. En 2012, Arora, Hazan et Kale publièrent un article remarquable¹⁹ qui synthétisa et unifia de nombreuses idées disparates mais similaires, en un algorithme d'une simplicité et d'une efficacité déconcertantes. Cet algorithme est l'algorithme par *multiplicative weights update*. Le mot important là est bien sûr le mot « *multiplicative* ». L'astuce de cet algorithme est d'utiliser des échelles multiplicatives pour orienter les choix, par opposition à l'échelle additive utilisée pour mesurer sa performance. De façon stupéfiante, cette astuce pourtant simpliste permit aux trois chercheurs de résoudre efficacement de nombreux problèmes sur lesquels avaient butté les générations précédentes.

L'élégance du *multiplicative weights update* est telle qu'Arora, Hazan et Kale ont suggéré que leur algorithme était l'une des idées les plus importantes de l'informatique. Ils proposent ainsi au début de l'article d'inclure l'enseignement de leur algorithme dès les premiers cours d'informatique, avec d'autres méthodes bien connues comme *divide and conquer*.

L'efficacité de l'algorithme par *multiplicative weights update* est un testament à la différence majeure qui existe entre le monde additif et le monde multiplicatif. Le fait qu'il n'ait été découvert et compris que très récemment est un témoignage du fait que cette distinction est très contre-intuitive. Il vient corroborer le postulat de Kurzweil. Notre intuition de la croissance exponentielle est très erronée.

L'un des succès les plus remarquables du *multiplicative weights update* est la technique du *boosting* en *machine learning*. Mais avant d'expliquer le *boosting* et le *multiplicative weights update*, faisons un détour par le siècle des Lumières en France et une réflexion remarquable du marquis de Condorcet.

À une époque où plusieurs philosophes français défendaient la décentralisation du pouvoir, Condorcet en vint à se demander s'il était préférable que les déci-

¹⁹  *The Multiplicative Weights Update Method: a Meta-Algorithm and Applications* | Theory of Computing | S. Arora, E. Hazan and S. Kale (2012)

sions judiciaires soient prises par un juge compétent ou par un jury composé de plusieurs citoyens moins compétents. Condorcet proposa un modèle simpliste où chaque citoyen avait une probabilité p strictement supérieure à $1/2$ de prendre une bonne décision. Dans ce contexte, la probabilité que le jury prenne une mauvaise décision décroît alors exponentiellement avec la taille du jury²⁰. Condorcet en déduisit qu'un jury suffisamment grand est bien plus fiable qu'un juge compétent.

Cependant, le cas de Condorcet est trop simpliste pour être réellement pertinent. En particulier, l'hypothèse majeure sous-jacente est l'indépendance des convictions des citoyens. Or, notamment si les membres du jury interagissent entre eux, il faut s'attendre à ce que les plus loquaces influencent les autres — d'autant que des expériences de psychologie comportementale comme celles de Solomon Asch montrent que l'effet de groupe peut rapidement conduire les individus à croire en des faits assez clairement faux.

Pire encore, il est difficile d'anticiper la manière dont les convictions individuelles vont se corréler — et les études empiriques²¹ des délibérations entre jurés sont très inquiétantes, puisqu'il semble que la délibération pousse les jurés vers des conclusions plus extrêmes que l'avis du juré le plus extrême.

Est-il alors possible de combiner différentes opinions, plutôt justes mais très peu fiables et possiblement corrélées, pour en déduire une opinion remarquablement fiable ? Telle est la question que posèrent les informaticiens Michael Kearns et Leslie Valiant en 1988.

En 1997, Robert Schapire et Yoav Freund répondirent à cette question par l'affirmative. Leur solution fut nommée *Adaboost*, leur valut le prestigieux prix Gödel en 2003 et conduisit aux premiers algorithmes de détection de visage par Viola et Jones. Le succès époustouflant d'*Adaboost* pourrait laisser croire qu'il s'agit là d'un algorithme très sophistiqué. Pourtant, *Adaboost* ne fait qu'exploiter la disproportion entre la croissance exponentielle et la croissance linéaire, entre l'échelle logarithmique et l'échelle usuelle, entre le monde multiplicatif et le monde additif.

Mieux encore, à l'instar du *multiplicative weights update* qui le généralise, *Adaboost* n'est rien d'autre qu'une équation de Lotka-Volterra déguisée. Autrement dit, *Adaboost* n'est qu'une approximation de la formule de Bayes !

Détaillons un petit peu. Supposons que l'on dispose d'experts aux opinions diverses et variées. Pour commencer, faute de données pour distinguer les experts, on va commencer par les considérer tous indiscernables les uns des autres. Du coup, l'opinion que l'on va se faire va être une simple moyenne des opinions des experts. Cependant, au fur et à mesure que les opinions des experts sont confrontés aux données, le poids associé à l'opinion d'un expert donné va être

²⁰En utilisant l'inégalité de Chernoff et en appelant n la taille du jury, on peut montrer que la probabilité d'erreur est majorée par $\exp\left(-n^2 \frac{(p-\frac{1}{2})^2}{2p(1-p)}\right)$.

²¹ *The law of group polarization* | Journal of Political Philosophy | C. Sunstein (2002)

multiplié par la cohérence de son opinion avec les données. L’opinion que l’on se fera désormais sera toujours une moyenne des opinions des experts, mais cette moyenne sera désormais pondérée par les poids que l’on associe aux opinions des différents experts. Plus précisément, la crédence d’un expert sera le poids de ses opinions relativement aux poids des opinions des autres experts.

Bayes part en vacances

Prenons un cas simpliste pour mieux comprendre *Adaboost*. Supposons que, tous les ans, vous partez en vacances dans un pays lointain. Chaque année, pour choisir une destination, vous demandez à chacun de vos n amis de conseiller une destination. Pour commencer, ne sachant pas quel ami écouter, vous allez écrire chaque proposition sur un bout de papier, et tirer au hasard un papier parmi ces propositions. Telle sera votre destination.

Cette année, vous avez tiré le Nigeria. Il est maintenant temps de confronter la prédiction de votre ami qui a suggéré le Nigeria avec les données expérimentales. Et pour cela il vous faut malheureusement subir un heureux voyage à l’autre bout du monde ! Vous rentrez bronzé et heureux. Vos vacances étaient absolument merveilleuses — et le résultat positif au test d’Ebola ne vous inquiète pas plus que cela.

Même si vous êtes sur un nuage, il ne faut pas oublier de mettre à jour vos crédences en vos amis. Pour cela, vous allez multiplier l’opinion de l’ami qui a suggéré le Nigeria par un nombre qui représente votre appréciation du voyage, sur une échelle de 0 à 1. Vous avez adoré le Nigeria ? Vous pouvez multiplier le poids de l’opinion de votre ami par²² 0,9.

Cependant, pour affiner vos crédences en vos autres amis, il va falloir également imaginer les vacances que vous auriez eues si vous aviez écouté leurs suggestions. Idéalement, il faudrait tester toutes les vacances qu’ils suggèrent. Mais vous manquez de congés. Heureusement pour vous, vos amis ont eux-mêmes testés les vacances qu’ils ont suggérées. Et ils adorent parler de leurs vacances. Vous pouvez donc aisément vous faire une idée de l’appréciation que vous auriez eue de leurs voyages. Il vous faut alors multiplier le poids des opinions de vos amis par votre estimation de votre appréciation pour leurs voyages²³.

²²Pour faire de jolis calculs, on va supposer que l’on multiplie le poids de l’opinion de l’ami par $(1 + \eta m)$ où $m \in [0, 1]$ est votre appréciation du voyage.

²³Appelons w_i le poids de l’opinion de l’ami i . On multiplie alors w_i par $(1 + \eta m_i)$, où m_i est votre estimation de votre appréciation pour son voyage. La crédence en l’ami i est alors donnée par la formule quasi-bayésienne qui suit :

$$\text{newCredence}(i) = \frac{(1 + \eta m_i) \text{oldCredence}(i)}{(1 + \eta m_i) \cdot \text{oldCredence}(i) + \sum_{j \neq i} (1 + \eta m_j) \cdot \text{oldCredence}(j)}.$$

L'été suivant approche, et vous devez choisir votre prochaine destination de vacances. Pour ce faire, vous réunissez vos amis, et recueillez leurs nouvelles suggestions. Encore une fois, vous allez tirer au hasard parmi leurs suggestions. Sauf que cette fois, vous ferez en sorte que la probabilité de tirer la suggestion de votre ami i sera proportionnelle au poids de l'opinion de l'ami i . Autrement dit, cette probabilité sera la crédence de l'ami i . Ainsi, si l'ami i avait vu juste les années passées, alors sa proposition aura plus de chances d'être tirée.

Cette approche naïve peut sembler bien alambiquée et peu convaincante de prime abord. Pourtant, *Adaboost* et le *multiplicative weights update* garantissent mathématiquement que vous ferez ainsi presque aussi bien que si vous n'aviez écouté que l'avis de l'ami le plus fiable ! Ce faisant, *Adaboost* garantit que vos décisions successives seront, en un sens qui peut être rendu rigoureux, presque optimales²⁴.

Un point important à souligner est l'influence quasi-nulle des crédences initiales en les différents amis. On les avait supposées égales pour tous les amis. Cependant, parce que ces crédences ont évolué de manière exponentielle, après quelques itérations de l'algorithme, les crédences initiales ont complètement disparu.

Plus généralement, l'évolution exponentielle des crédences bayésiennes implique qu'après une poignée d'inférences bayésiennes, le rôle de l'*a priori* se dissipe. C'est pour cette raison qu'en présence de suffisamment de données, les crédences bayésiennes sont en fait beaucoup moins arbitraires que ce que notre intuition linéaire laisse à penser. *La subjectivité n'a rien d'arbitraire.*

La singularité

Kurzweil va plus loin et reproche au monde technologique et académique de ne pas saisir la croissance exponentielle des technologies, à l'instar de la loi de Moore qui affirme (grossièrement) que la puissance des ordinateurs double tous les deux ans. De la même manière, les économistes Brynjolfsson et McAfee suggèrent que l'on attache trop d'importance à divers événements du passé. Si l'on se concentre sur des métriques économiques, que ce soit le nombre d'habitants ou la production agricole, un phénomène majeur saute aux yeux et transcende tous les faits ponctuels comme la chute de l'empire romain, l'invention de l'imprimerie et la conquête de l'Amérique. Ce phénomène majeur est l'inéluctable croissance exponentielle des métriques économiques.

²⁴Plus précisément, on peut montrer qu'on a alors

$$\mathbb{E} \left[\sum_t m(t) \right] \geq (1 - \eta) \left\{ \max_{i \in [n]} \sum_t m_i(t) \right\} - \frac{\ln n}{\eta},$$

où $m_i(t)$ est votre appréciation (estimée) du voyage suggéré par l'ami i l'année t , et $m(t)$ est votre appréciation du voyage que vous avez effectivement effectué l'année t .

Pour Brynjolfsson, McAfee, Kurzweil et bien d'autres, cette croissance exponentielle, portée par les nouvelles technologies, annonce un futur très distinct du passé. L'intelligence artificielle, les imprimantes 3D, la nanotechnologie et les progrès de la génétique annoncent un futur dans lequel le travail humain ne sera plus nécessaire pour produire des objets de consommation en quantité et en qualité. La faim dans le monde et les maladies pourraient alors être éradiquées. Nos modes de vie seraient bouleversés.

Dans une excellente vidéo intitulée *Humans Need Not Apply*, le YouTuber CGP Grey suggère même que dans un futur proche, le travail humain ne sera même plus souhaité. Pour CGP Grey, les technologies pourraient bientôt être telles que (presque) tout travail effectué par un humain pourra être mieux effectué et à un coût bien moindre par une machine²⁵. Brynjolfsson et McAfee partagent ce constat, et annoncent l'extinction à venir des métiers. J'ai moi-même succombé à de tels arguments, ayant prédit en 2014 un taux de chômage supérieur à 80 % dès 2034 — n'hésitez pas à me rappeler que j'aurai eu tort à ce moment-là !

Mais McAfee n'est pas défaitiste. Bien au contraire. Dans une conférence TED²⁶, McAfee parle de « la meilleure nouvelle économique de notre époque ». « Non pas que la compétition soit là », rajoute-t-il. L'abondance sera garantie. Nous pourrons repenser nos sociétés, sociétés où il n'y aura pas à requérir le travail des uns et des autres — à condition de redistribuer adéquatement les biens produits par les machines. Ce n'est jamais arrivé dans l'histoire de l'humanité !

Nick Bostrom voit plus loin encore. Bostrom suggère que la croissance des technologies pourrait bien être non pas exponentielle, mais super-exponentielle. Pour comprendre cela, revenons-en à ce qui caractérise la croissance exponentielle : à chaque itération, la technologie, ou une certaine mesure plus quantitative de celle-ci, est multipliée par une constante. Or, pour Bostrom, plus les technologies se développent, plus la vitesse à laquelle elle se développe pourrait être grande. Dit autrement, à chaque itération, la technologie est multipliée par un nombre qui augmente à chaque itération.

Ce phénomène peut se modéliser à l'aide d'une équation différentielle. Je vais me permettre un peu de jargon mathématique dans ce paragraphe. N'hésitez pas à le sauter si vous n'êtes pas à l'aise avec ce jargon. La croissance exponentielle correspond à l'équation différentielle $\dot{x} = x$. La croissance technologique qu'envisage Bostrom serait davantage de la forme $\dot{x} = x^2$. Mais une telle équation a une solution de la forme $x(t) = 1/(1-t)$. Autrement dit, cette croissance est si rapide que x atteint une quantité infinie au bout d'un temps $t = 1$ fini.

La conclusion de cette petite réflexion est qu'il pourrait y avoir une singularité dans l'évolution technologique, un point où les technologies atteindraient tout à coup les limites de la physique. Cette singularité technologique est souvent interprétée comme étant le moment où une superintelligence, c'est-à-dire une intelligence artificielle supérieure à tout humain, se mettrait à améliorer sa propre

²⁵  *Humains versus machines* | IA 1 | Science4All | L.N. Hoang (2017)

²⁶  *What Will the Future of Jobs Look Like?* TED | A. McAfee (2013)

intelligence. Étant plus intelligente que ses concepteurs, cette superintelligence trouverait des solutions technologiques hors de notre portée, et accélérerait son auto-amélioration à un rythme effréné. En l'espace de très peu de temps, elle changerait du tout au tout le monde dans lequel on vit. Et son comportement serait fondamentalement imprévisible — car obtenue via une intelligence qui nous est très supérieure²⁷.

Bostrom ne s'aventure pas à deviner la date précise de cette singularité hypothétique. Cependant, il n'exclut pas son émergence d'ici 50 ans. Ray Kurweil, lui, va plus loin. Il prédit que cette singularité technologique aura lieu en 2045. Cette prédiction en fera sauter plus d'un. Pour la plupart d'entre nous, elle paraît même honteusement ridicule. Mais, pour Kurzweil, c'est parce que la plupart d'entre est incapable de s'extirper de l'intuition linéaire.

J'ai beau avoir mes réserves quant à de telles prédictions, j'émetts toutefois encore plus de réserves concernant mes réserves, ayant eu maintes fois l'occasion de constater les limites de mon intuition de la croissance exponentielle.

Références en français

- ➲ *Le grand roman des maths : de la préhistoire à nos jours* | Flammarion | M. Launay (2016)
- ➲ *De quoi le succès d'une chaîne Youtube de vulgarisation dépend-il ?* Science Étonnante | D. Louapre (2017)

- ▶ *Addition contre multiplication* | MicMaths | M. Launay (2014)
- ▶ *Merveilleux logarithmes* | MicMaths | M. Launay (2014)
- ▶ *La loi de Benford* | Passe-Science | T. Cabaret (2015)
- ▶ *La Légende de Sessa* | Scientificfiz (2017)
- ▶ *Vous êtes de sang royal* | Dirty Biology | L. Grasset (2018)

- ▶ *Le top 5 des études de psychologie sociale qui vous feront questionner les choses* | Outside The Box (2015)
- ▶ *L'argent fait-il le bonheur ?* Stupid Economics | V. Levetti et A. Gantier (2017)
- ▶ *Êtes-vous un hooligan politique ?* Démocratie 10 | Science4All | L.N. Hoang (2017)
- ▶ *Humains versus machines* | IA 1 | Science4All | L.N. Hoang (2017)

²⁷ ➔ *What is Singularity, Exactly?* Up and Atom | J. Tan-Holmes (2018)

Références en anglais

- ➲ *The singularity is near: When humans transcend biology* | Penguin | R. Kurzweil (2005)
- ➲ *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies* | W. W. Norton & Company | E. Brynjolfsson and A. McAfee (2005)
- ➲ *Superintelligence: Paths, Dangers, Strategies* | Oxford University Press | N. Bostrom (2014)

- ➲ *Modelling the recent common ancestry of all living humans* | Nature | D. Rohde, S. Olson, et J. Chang (2004)
- ➲ *The law of group polarization* | Journal of Political Philosophy | C. Sunstein (2002)
- ➲ *The Multiplicative Weights Update Method: a Meta-Algorithm and Applications* | Theory of Computing | S. Arora, E. Hazan and S. Kale (2012)

- ➲ *Universal Paperclips* | Decision Problem | F. Lantz (2007)

- ▶ *What Will the Future of Jobs Look Like?* TED | A. McAfee (2013)
- ▶ *Humans Need Not Apply* | CGP Grey (2014)
- ▶ *The Accelerating Future* | R. Kurzweil (2016)
- ▶ *How many particles in the universe?* Numberphile | T. Padilla (2017)
- ▶ *What is Singularity, Exactly?* Up and Atom | J. Tan-Holmes (2018)
- ▶ *DNA Encoding* | ZettaBytes | C. Dessimoz (2018)

- ▶ *The Multiplicative Weights Update Algorithm* | Wandida | L.N. Hoang (2016)
- ▶ *Motivations and Applications of the Multiplicative Weights* | Wandida | L.N. Hoang (2016)
- ▶ *Theoretical Guarantee for the Multiplicative Weights Update* | Wandida | L.N. Hoang (2016)

Il est inutile de faire avec plus ce qui peut être fait avec moins.

Guillaume d'Ockham (1285-1347)

La simplicité est la sophistication ultime.

Léonard de Vinci (1452-1519)

12

Tranchons avec le rasoir d'Ockham

Jeudi dernier...

En 2002, dans l'État américain de l'Ohio, Tonda Lynn Ansley fut jugée pour le meurtre de la propriétaire de son logement. Ansley se défendit en affirmant qu'elle croyait vivre dans la *Matrix*, en référence à la trilogie cinématographique du même nom. Dans cette série de films hollywoodiens, la *Matrix* est une simulation informatique dans laquelle vit l'écrasante majorité des humains. Les humains interagissent entre eux dans cet univers virtuel depuis si longtemps que (presque) aucun d'entre eux ne distingue la simulation du monde réel. Ils confondent leur univers simulé avec la réalité.

Cependant, *Matrix* n'est qu'un film et se croire dans ce film est souvent interprété comme un signe de déraison. D'ailleurs, Ansley fut jugée victime de troubles mentaux, ce qui lui valut d'être acquittée. Pour beaucoup, la *Matrix* n'est que fiction. Il faut être mentalement dérangé pour y croire.

Pourtant, des scientifiques de renom comme Stephen Hawking n'hésitent pas à considérer sérieusement l'hypothèse de la *Matrix*. Nick Bostrom propose même un argument très convaincant en faveur de cette hypothèse : si la technologie le permet, il est raisonnable de penser que les humains préféreraient descendre des pistes de ski virtuelles où le froid n'est pas glacial et dont les avalanches sont sans danger pour leur intégrité physique. Petit à petit, l'univers virtuel pourrait être l'univers préféré des humains. Ainsi, la *Matrix* pourrait être le futur de toute civilisation suffisamment avancée. Or, les civilisations les plus

peuplées sont celles qui sont développées. Du coup, on pourrait s'attendre à ce que la plupart des êtres intelligents de l'univers vivent dans des *Matrix*. Mais alors, si on prend n'importe quel être intelligent de l'univers au hasard — nous-même par exemple — la probabilité que cet être soit dans une *Matrix* est très proche de 1. Dès lors, l'hypothèse de la *Matrix* est non seulement à considérer, elle est même très crédible. En particulier, il n'y a rien de complètement insensé à y accorder une crédence non-négligeable¹ !

Il est même possible d'aller plus loin encore dans les théories métaphysiques obscures. Une théorie extrême est le *Last-Thursdayism*. Selon cette théorie, tout l'univers a été créé jeudi dernier. Tout l'univers, toute notre Terre, toutes nos civilisations, tous nos monuments, tous nos livres, mais aussi tous nos souvenirs. Si vous croyez avoir passé des vacances au Nigeria l'été dernier, c'est uniquement parce que jeudi dernier, lorsque tout a été créé, votre cerveau contenait des souvenirs de vacances passées au Nigeria. Mieux encore, le *Last-Thursdayism* est irréfutable et parfaitement cohérent avec les lois de la physique. Peu importe l'observation future, il sera possible d'en déterminer des causes datant de jeudi dernier².

Pour Karl Popper, cependant, le *Last-Thursdayism*, comme l'hypothèse de la *Matrix*, n'ont aucune valeur car il s'agit précisément de théories irréfutables. Une telle réponse peut paraître séduisante. Cependant, on a vu dans le chapitre 4 que la réfutabilité de Popper n'a ni pendant empirique ni fondement théorique. Je ne reviendrai pas dessus.

Le concept traditionnel pertinent pour dénigrer la théorie du *Last-Thursdayism* ou l'hypothèse de la *Matrix* n'est pas la philosophie de Popper. Le concept traditionnel pertinent, c'est le *rasoir d'Ockham*, du nom du philosophe Guillaume d'Ockham, que l'on appelle aussi le principe de parcimonie, d'économie ou de simplicité. En 1319, Ockham écrivit : « *Pluralitas non est ponenda sine necessitate.* » En français, « la multitude ne doit pas être avancée sans nécessité. » Ou dit autrement, les théories plus simples sont préférables.

Cependant, il n'est pas évident de voir en quoi le *Last-Thursdayism* serait moins simple que l'alternative selon laquelle l'univers observable aurait émergé il y a 13 milliards d'années, générant ensuite toute la complexité des galaxies, des étoiles, des planètes, du vivant et de nos cerveaux humains. Malgré sa simplicité apparente, le principe de simplicité d'Ockham n'est pas simple ! En particulier, ce qui peut paraître simple ne l'est pas forcément — et ce qui paraît compliqué n'est pas nécessairement compliqué non plus !

En fait, une compréhension rigoureuse de la simplicité des théories semble nécessiter une théorie de la complexité comme la théorie de la complexité algorithmique. Typiquement, les travaux de Solomonoff semblent être un socle incontournable pour une bonne formalisation du rasoir d'Ockham.

¹ *Sommes-nous des simulations ? L'argument de la simulation de Nick Bostrom | Argument frappant | Monsieur Phi | T. Giraud (2016)*

² *Is anything real? VSauce | M. Stevens (2013)*

Dans le football, rien n'est écrit d'avance

Mais pour l'heure, insistons sur l'importance cruciale du rasoir d'Ockham, notamment dans l'optique de déterminer une théorie prédictive. Comme l'ont découvert à leurs dépens les chercheurs en statistiques et en *machine learning*, sans rasoir d'Ockham, le piège récurrent dans lequel on tombe à pieds joints est celui de l'*overfitting*, que l'on pourrait traduire en « sur-interprétation³ ». Et pour se rendre compte des conséquences néfastes de l'*overfitting* et du rôle salvateur que le rasoir d'Ockham joue (ou pourrait jouer), faisons un détour vers le monde où l'*overfitting* est roi : le monde du sport.

Les prolongations ont déjà commencé, et l'image de la frappe de Gignac sur le montant droit du gardien portugais hante encore les esprits des joueurs français et de leurs supporters. Cette finale France-Portugal de l'Euro 2016 leur semble pourtant promise, eux qui ont remporté les deux grandes compétitions internationales de football qu'ils ont organisées — il y en a bien une avant la guerre qui leur avait fait défaut, mais c'était à une toute autre époque. Qui plus est, la France avait déjà gagné l'Euro en 1984, puis en 2000, comme si une règle sous-jacente semblait promettre l'Euro à la France tous les 16 ans. Enfin, l'Histoire de l'équipe de France montre qu'elle ne gagne que lorsqu'elle est portée par un joueur d'exception. En 1984, ce fut Platini. Pour la coupe du monde 1998 et à l'Euro 2000, ce fut Zidane. Cette année, c'est Griezmann qui transcende la compétition.

Cependant, au bout de la prolongation, c'est le Portugal qui marquera le seul et unique but de cette finale. Le Portugal devient champion d'Europe, défiant alors tous les pronostics et toutes les règles statistiques que l'on semblait avoir établies. Les statistiques nous auraient donc menti !

Ou peut-être pas. Les journaux titrent alors que cet Euro 2016 est la compétition de toutes les revanches. En quarts de finale, l'Allemagne a battu l'Italie pour la première fois de l'histoire des tournois internationaux de football. En demi-finale, la France a battu l'Allemagne pour la première fois depuis une petite finale en 1958. Et en finale, le Portugal a battu la France pour la première fois, alors qu'il restait sur une série de 10 défaites consécutives toutes compétitions confondues. Les bêtes noires ont été vaincues.

De son côté, Griezmann semble vivre une année pleine et remarquable, et ses performances intrinsèques le rendent favori pour l'attribution du ballon d'or, l'équivalent du prix Nobel pour le football. Cependant, Griezmann a perdu la finale de l'Euro 2016 contre le Portugal de Cristiano Ronaldo, après avoir gagné contre l'Allemagne de Manuel Neuer. Et quelques mois plus tôt, son équipe de club, l'Atletico Madrid, a perdu la finale de la ligue des champions contre le Real de Madrid de Cristiano Ronaldo, après avoir battu le Bayern de Munich

³La traduction française « surapprentissage » me semble être très mal choisie, car elle risque fortement de conduire à des contre-sens.

de Manuel Neuer. Quelques mois plus tard, c'est Cristiano Ronaldo qui sera récompensé du ballon d'or — Griezmann sera classé troisième.

Toutes les analyses que je viens de mentionner sont des analyses typiques des journaux sportifs. Les statistiques y sont utilisées pour révéler des régularités intrigantes, saisissantes, voire troublantes. Cependant, pour l'expert en *machine learning*, elles n'ont sans doute aucune valeur, car elles correspondent très probablement à de l'*overfitting*. En effet, si on jette un œil à l'histoire du football et si on torture les statistiques des matchs passés, on trouvera toujours tout plein de régularités statistiques remarquables. À chaque nouveau résultat, certaines de ces régularités sont détruites — comme le fait que la France gagne l'Euro tous les 16 ans — mais les régularités statistiques potentielles sont suffisamment nombreuses pour que toutes ne périssent pas. Bien au contraire, plus les données s'accumulent, plus il y a de façons de torturer les données pour y trouver des régularités statistiques apparentes.

C'est là que réside l'*overfitting*. Lorsque le nombre d'explications *ad hoc* augmente plus vite que le nombre de données, alors, peu importe les données, on trouvera une explication à ces données. C'est typiquement ce qu'il se passe quand des commentateurs sportifs prennent le temps de croiser toutes les informations sur tout plein de joueurs au cours de tout plein de matchs. C'est ainsi que tous les quatre matins, on découvre un joueur qui a établi un nouveau record.

Le fléau de la sur-interprétation

Ce phénomène d'*overfitting* est tourné en dérision par l'excellent Tyler Vigen sur son site *Spurious Correlation*. Vigen s'est ainsi amusé à croiser de nombreuses données temporelles du web et à y chercher systématiquement des corrélations étonnamment significatives. Sauf que ces corrélations sont si improbables *a priori* qu'il est impossible de les prendre au sérieux.

Ainsi, on y découvre que les années où Nicolas Cage apparaît dans le plus de films sont aussi les années où il y a le plus de morts par noyade dans les piscines, que les années de grande consommation de margarine sont accompagnées d'un grand taux de divorce dans l'État du Maine, et que, lorsque la miss États-Unis est âgée, les meurtres par vapeur et objets chauds sont nombreux. Heureusement, même une fois ces statistiques bien connues, les politiciens ne chercheront ni à arrêter la carrière cinématographique de Nicolas Cage, ni à interdire la margarine, ni à faire pression sur le jury de miss États-Unis...

Les cas que présente Tyler Vigen sont fascinants précisément parce que l'on a tendance à rejeter tout lien de causalité, malgré des corrélations très marquées. Ce sont des cas pédagogiquement excellents. Ils sont l'occasion de rappeler que corrélation n'est pas causalité, surtout lorsque le risque d'*overfitting* est élevé — et dans notre cas, il l'est, car les jeux de données dont on peut tester les

corrélations sont bien plus nombreux que le nombre de données dans chacun des jeux de données. Ici, les corrélations sont les équivalents des explications *ad hoc*, et ils sont en bien plus grand nombre que les tailles des échantillons pour chaque type de données.

Cependant, rejeter tout lien de causalité malgré une corrélation marquée n'est pas un réflexe que la plupart d'entre nous avons, et le piège de l'*overfitting* ne se restreint pas au monde du sport. L'actualité est souvent pleine de sur-interprétations prises très au sérieux, et dont les conséquences peuvent être majeures.

Dans un but pédagogique, le site FiveThirtyEight⁴ vous propose ainsi une interface web où vous pouvez aisément jouer avec des données politiques américaines. Après quelques bidouilles, vous pourrez construire une statistique qui suggérera que votre parti préféré a un effet positif sur l'économie des États-Unis. Et le plus fort, c'est qu'en y passant quelques secondes, vous réussirez à en déterminer une qui passe le seuil de la *p-value* exigée par la « méthode scientifique » ! Autrement dit, votre statistique sera suffisamment significative pour être publiée par une revue scientifique — elle le sera donc clairement pour être publiée dans le *New York Times* !

La raison pour laquelle l'approche de FiveThirtyEight peut amener à n'importe quelle conclusion prédéfinie, est que le site propose une grande multitude de façons de mesurer l'effet d'un parti politique sur l'économie. Il y a des métriques différentes (chômage, inflation, PIB, marché financier), des représentations différentes du parti au sein des autorités dirigeantes (président, gouverneurs, sénateurs, représentants) avec différentes façons de comparer les importances relatives de ces dirigeants, et même la possibilité de prendre en compte ou non les récessions économiques. Comme, qui plus est, il est possible de sélectionner des combinaisons de ces différents paramètres, par exemple à la fois le chômage et le PIB, le nombre d'explications possibles proposées par le site web de l'effet d'un camp politique sur l'économie atteint 2 048.

Or, souvenez-vous, même en l'absence d'un réel effet significatif, la méthode de la *p-value* signale un cas sur 20 comme étant significatif ! Par conséquent, dans notre cas, on s'attend à ce qu'une centaine de statistiques soient scientifiquement publiables ! Plus étrange encore, en jouant avec les données du site web, on se rend compte qu'il est tout aussi facile d'obtenir des statistiques significatives en faveur des démocrates qu'en faveur des républicains. Autrement dit, en jouant suffisamment longtemps avec les données de FiveThirtyEight, vous pourrez aisément publier un article au titre *putaclic*⁵ « 50 statistiques qui prouvent que x est mauvais pour l'économie », quelle que soit la valeur de $x \in \{\text{Democrats}, \text{Republicans}\}$!

⁴  Hack your way to scientific glory | FiveThirtyEight (2015)

⁵ Ces étapes de calcul sont très approximatifs et ne sont pas vraiment valides. Mais ils donnent bel et bien une idée de ce que l'on peut tirer d'une analyse telle que le propose FiveThirtyEight.

Or, l'interface web de FiveThirtyEight est en fait extrêmement limité. Un journaliste pressé par son comité éditorial et suffisamment habile avec l'informatique — ou ayant une connaissance qui, elle, est habile avec l'informatique — pourra aisément générer des millions, voire des milliards d'explications possibles de l'effet d'un camp politique sur l'économie, et ainsi publier mille statistiques significatives tous les jours pendant les cent prochaines années. Telle est l'ampleur du danger de l'*overfitting*. En explorant toujours plus d'explications plausibles, on est voué à trouver des statistiques significatives pour défendre n'importe quelle position — souvent sans même se rendre compte que la découverte de ces statistiques n'a en fait rien de miraculeux. *Même s'il est improbable pour chacune des statistiques d'être significative, il est encore plus improbable qu'aucune de ces statistiques ne le soit.*

Cette remarque simpliste est la cause des nombreux articles qui se contredisent mutuellement sur les sujets de société, de la politique au racisme, en passant par le terrorisme, l'alimentation et la religion. Après tout, plus un sujet suscite la curiosité de la population, plus les journalistes vont passer de temps à faire des recherches sur le sujet, plus ils trouveront des statistiques sensationnelles susceptibles de faire le *buzz*, et plus la population s'intéressera à ce sujet. Il s'agit d'un cercle vicieux, qui a le malheur de créer des convictions tranchées. Ces convictions reposent alors presque exclusivement sur de l'*overfitting*. Sauf que cet *overfitting* est invisible pour la plupart d'entre nous qui ne lisons que les statistiques significatives habilement sélectionnées et cueillies par des journalistes dont les patrons imposent de faire le *buzz*. Et quand on combine cela au *putaclic* dont on a parlé dans un chapitre précédent, on semble inéluctablement se diriger tout droit vers une profusion incontrôlable d'informations trompeuses.

En particulier, l'*overfitting* est un piège dans lequel l'écrasante majorité des militants convaincus saute à pieds joints à tout moment. Si l'on cherche à défendre une position, il suffit d'explorer suffisamment d'explications possibles pour en trouver une qui paraîsse justifier cette position. Pourvu qu'on la cherche suffisamment longtemps, on trouvera toujours une explication *ad hoc*⁶.

Malheureusement, d'après le psychologue Jonathan Haidt, les expériences en sciences sociales ne cessent de montrer, encore et encore, que nous autres humains prenons d'abord position, et justifions ensuite notre position par des arguments (que l'on croit) rationnels. La raison ne nous sert que d'instrument pour rechercher et « éternuer » des explications *ad hoc* à nos convictions préétablies. Or, les arguments *ad hoc* étant omniprésents, il nous suffit de nous doter d'une raison compétente pour être sûr de ce que l'on veut croire⁷.

Tel est le travers dans lequel nous tombons constamment. Tel est le travers des superstitions et des croyances surnaturelles. Tel est aussi le travers dans lequel tombe le *Last-Thursdayism*.

⁶  *La sur-interprétation (overfitting)* | IA 11 | Science4All | C. Michel et L.N. Hoang (2018)

⁷  *Êtes-vous un hooligan politique ?* Démocratie 10 | Science4All | L.N. Hoang (2017)

Pour toute nouvelle observation, il existe une nouvelle explication pour rendre cette observation compatible avec le *Last-Thursdayism*. En fait, il est même probable que, pour expliquer le monde qui nous entoure, un adepte du *Last-Thursdayism* finira par développer un modèle de l'univers semblable à celui que les scientifiques ont développé. Mais alors, l'hypothèse du *Last-Thursdayism* deviendra superflue. Elle ne permettra pas d'expliquer plus de choses que ce que les autres éléments de la théorie permettent déjà d'expliquer. Cette hypothèse étant désormais superflue, elle se verra tranchée par le rasoir d'Ockham.

La complexe quête de simplicité

Vous l'aurez compris. Le rasoir d'Ockham est l'outil de prédilection dans la lutte contre l'*overfitting*. Plutôt que de faire des va-et-vient entre théories concurrentes à chaque nouvelle donnée découverte, le rasoir d'Ockham suggère de négliger les théories trop sophistiquées, quitte à mal expliquer toutes les données. Après tout, les données sont généralement le fruit de causes si multiples qu'il est illusoire d'espérer toutes les expliquer à la perfection.

Quand un dé tombe sur un six, la position de chaque particule d'air peut potentiellement avoir eu un impact sur le résultat. Or, il est illusoire de suivre le mouvement de chaque particule d'air — notamment parce que le nombre de ces particules dépasse de très très loin l'espace mémoire combiné de tous les ordinateurs créés jusque-là. Or, un lancer de dé est beaucoup, beaucoup, beaucoup plus simple que les questions de société qui nous intriguent tant. S'il nous est impossible d'expliquer entièrement la chute d'un dé, il est dès lors complètement illusoire d'espérer avoir le fin mot des explications de la politique, du terrorisme et de la nutrition. Il nous faut accepter et embrasser l'incomplétude de nos modèles. « Tous les modèles sont faux. » Et c'est une bonne chose !

L'un des premiers à se rendre compte de cette nécessité de ne pas tout expliquer parfaitement est sans doute Galilée, le père des sciences modernes. L'un des plus grands coups de génie de Galilée fut de défier la physique d'Aristote, en affirmant que, non, les objets plus lourds ne tombent pas intrinsèquement plus vite. Ce postulat de Galilée, connu sous le nom de la loi de la chute des corps, est une absurdité expérimentale. Prenez une plume et un caillou, et laissez-les tomber. Vous verrez que Galilée avait tort.

Cependant, le génie de Galilée fut de se rendre compte que la chute intrinsèque des objets n'était qu'une partie de ce qui les faisait se mouvoir. Les objets, et surtout la plume, subissent les effets de l'air. Cet air freine le mouvement de certains objets davantage que celui d'objets plus lourds. Il permet même aux oiseaux de voler. Galilée postula alors qu'en l'absence d'air, les effets de l'air disparaîtraient, et l'on observerait alors les chutes intrinsèques des objets, lesquelles seraient alors indépendantes des masses des objets. Dans le vide, postula Galilée, tous les objets tombent à la même cadence.

On raconte souvent que Galilée serait monté en haut de la tour de Pise pour tester sa loi de la chute des corps. Mais il est très probable que cette histoire fut inventée de toute pièce par l'étudiant de Galilée. Après tout, si Galilée avait fait l'expérience, il aurait observé que l'objet plus lourd, moins freiné par les frottements de l'air, tombe plus vite. Ce n'est pas l'expérience empirique qui donna raison à Galilée. C'est en fait une expérience de pensée, que je ne détaillerai pas ici⁸, qui montre que l'hypothèse selon laquelle la masse des objets est la seule variable à affecter les chutes des objets est auto-contradictoire — à moins de supposer que la masse des objets n'a en fait aucun effet intrinsèque sur leur chute.

De la même manière, l'autre idée de génie de Galilée fut le principe de relativité. Ce principe affirme qu'un homme assis dans une cale sans fenêtre d'un bateau serait incapable de savoir si le bateau est en mouvement. « Le mouvement est comme rien », disait Galilée. Là encore, il n'est pas clair que l'expérience aurait donné raison à Galilée — on peut imaginer que le bateau s'agiterait plus lorsqu'il est en mouvement que s'il était amarré à quai. Cependant, les différences entre la théorie et la pratique étaient suffisamment faibles et suffisamment aléatoires pour que Galilée eut pleinement confiance en sa relativité du mouvement. Peu de temps après, cette crédence en son principe de relativité conduira d'ailleurs Galilée à mettre le Soleil au centre de l'univers⁹.

Dans les deux cas, Galilée eut le génie de préférer la simplicité et l'élégance de ses principes à leur adéquation avec la pratique. Voilà une brillante application du rasoir d'Ockham pour éviter le piège de l'*overfitting* dans lequel d'autres avant lui sont tombés. Un demi-siècle plus tard, ce sera au tour d'Isaac Newton de postuler le principe fondamental de la dynamique que l'on résume désormais en quatre symboles : $\vec{F} = m\vec{a}$. Puis, deux siècles plus tard, James Clerk Maxwell mit en avant la simplicité et l'élégance de ses équations pour suggérer qu'elles puissent expliquer à la fois l'électricité, le magnétisme et la lumière. Toutes ces brillantes théories reposent sur un même principe : remplacer une multitude d'explications disparates et *ad hoc* par des principes simples et universels — quitte à ce que tous les phénomènes ne soient pas parfaitement expliqués par la théorie.

Tout n'est pas simple

Cependant, il serait erroné de croire que les meilleures théories sont toujours simples. Les modèles de météorologie sont d'ailleurs bien connus pour être affreusement sophistiqués, tandis que les neurosciences modernes suggèrent fortement que la compréhension du cerveau humain nécessitera inéluctablement des modèles horriblement complexes — peut-être nécessairement aussi complexes que le cerveau lui-même !

⁸  *La loi de la chute des corps* | Relativité 13 | Science4All | L.N. Hoang (2016)

⁹  *La Terre est-elle au centre du monde ?* | Relativité 14 | Science4All | L.N. Hoang (2016)

Ainsi, en 2016, l'intelligence artificielle AlphaGo qui réussit à battre Lee Sedol au jeu de Go était si compliquée que tout un ordinateur était nécessaire pour le représenter. Il en était de même pour Cepheus et Libratus, ces intelligences artificielles qui ont triomphé des joueurs de poker humains¹⁰.

En fait, quand on a parlé du démon de Solomonoff, on a déjà vu quelle était la complexité nécessaire pour étudier un phénomène : il s'agit de la complexité de Solomonoff (de la loi probabiliste) des données. S'il ne connaissait pas la version formelle de ce concept, Alan Turing l'avait toutefois déjà bien compris avant n'importe qui. Dans son article historique de 1950, Turing se posa la question de la complexité minimale d'un ordinateur capable de parler comme un humain. En s'appuyant sur les premiers pas de la neuroscience, Turing estima à des gigaoctets la taille du plus simple algorithme capable de modéliser aussi bien qu'un humain la manière dont les humains communiquent. Autrement dit, pour Turing, la complexité de Solomonoff du langage parlé est probablement de l'ordre du gigaoctet. On en reparlera longuement au chapitre 14.

De même, la complexité de Solomonoff de nombreux phénomènes biologiques, sociologiques et économiques pourrait surpasser de loin cette quantité — rendant alors impossible la compréhension de ces phénomènes par nos cerveaux dont l'espace mémoire semble limité à quelques pétaoctets. Tout modèle simple serait alors voué à l'échec face à la biologie, à la sociologie et à l'économie.

Cependant, les gros modèles nous exposent à l'*overfitting*. Ce qui va alors nous permettre de développer de la complexité sans *overfitting* est le *buzzword* des *data sciences* : le *Big Data*. Plus on a de données, plus on peut complexifier nos modèles. Il existe même une formulation rigoureuse de ce principe : le théorème fondamental de l'apprentissage statistique¹¹. De façon grossière, ce théorème détermine le nombre d'échantillons nécessaires pour ajuster les paramètres d'un modèle. Ou à l'inverse, étant donné des échantillons, ce théorème nous suggère la complexité adéquate des modèles à considérer.

La mesure quantitative de la complexité utilisée dans le théorème fondamental de l'apprentissage statistique est la notion de dimension VC, du nom des informaticiens Vladimir Vapnik et Alexey Chervonenkis. La définition rigoureuse de ce concept est un peu trop compliquée pour nous¹². De façon grossière, la dimension VC compte le nombre d'explications *ad hoc* pour expliquer les données. La règle grossière que l'on peut déduire du théorème est la suivante : le nombre d'échantillons doit être 100 fois plus grand que la dimension VC de l'ensemble des explications considérées¹³.

¹⁰ *Le poker résolu ! (ou non)* | Démocratie 15 | Science4All | L.N. Hoang (2017)

¹¹ *Le théorème fondamental de l'apprentissage statistique* | IA 15 | Science4All | L.N. Hoang (2018)

¹² Une hypothèse est là une fonction $X \rightarrow Y$. La dimension VC d'un ensemble d'hypothèses $\mathcal{H} \subset X^X$ est la taille maximale $|X_{max}|$ d'une partie $X_{max} \subset X$ telle que toutes les fonctions $X_{max} \rightarrow Y$ peuvent être obtenues par restriction des hypothèses de \mathcal{H} à X_{max} . Voir :

Les explications ad hoc (dimension VC) | IA 14 | Science4All | L.N. Hoang

¹³ La formulation formelle repose sur la notion de PAC-Learning. De façon grossière, le

La validation croisée

Jusque-là, j'ai surtout insisté sur l'*overfitting*, parce que c'est sans doute le piège dans lequel nous tombons le plus souvent. Cependant, il y a également le pendant opposé qui est l'*underfitting*, ou sous-interprétation. L'*underfitting* consiste à ne pas donner suffisamment d'importance à l'écart entre la théorie et la pratique. C'est typiquement ce dont on est victime lorsqu'on ignore nonchalamment des données qui contredisent nos croyances — même si dans notre cas et contrairement à des algorithmes d'apprentissage, c'est souvent davantage la faute à un biais cognitif.

Trouver un juste milieu entre *underfitting* et *overfitting* est un problème classique — et souvent considéré non-résolu — des *data sciences*. Il est parfois illustré par le dilemme biais-variance¹⁴. Imaginons que l'on cherche à prédire la propriété y d'une donnée x . Pour ce faire, on peut collecter plein d'exemples de paires (x_i, y_i) . Appelons S l'ensemble de ces paires. On dit que S est un *training set*. Puis on applique une certaine approche pour prédire y à partir de la donnée x et du *training set* S . Appelons $f(x, S)$ notre prédition.

Supposons maintenant que S est un *training set* aléatoire. L'erreur quadratique moyenne qu'on est amené à faire est alors

$$\mathbb{E}_S \left[(f(x, S) - y)^2 \right] = (\mathbb{E}_S [f(x, S)] - y)^2 + \text{Var}_S (f(x, S)).$$

Cette équation est souvent réécrite en $\text{erreur}^2 = \text{biais}^2 + \text{variance}$. Autrement dit, l'erreur peut se décomposer en deux morceaux. D'une part, il y a l'erreur due au fait qu'en moyenne, notre algorithme prédit mal. C'est le biais. D'autre part, il y a l'erreur due aux fluctuations de la prédition d'un *training set* à l'autre. C'est la variance.

L'*underfitting* correspond alors à utiliser un algorithme d'apprentissage trop rigide. Cette rigidité l'empêche de s'adapter aux données, et cause alors le biais dans la prédition. Pour résoudre l'*underfitting*, la solution la plus simple est souvent d'augmenter la complexité de notre algorithme d'apprentissage. Typiquement, on peut augmenter le nombre de ses paramètres. Cependant, on risque alors l'*overfitting*. L'*overfitting* va trop coller aux données. Il va trop subir l'influence des aléas de l'échantillonnage du *training set*. Pour éviter de telles fluctuations, il est alors souhaitable de réduire le nombre de paramètres. Le problème, c'est que déterminer le *fitting adéquat a priori* est délicat, puisque celui-ci semble être une propriété intrinsèque des données.

théorème fondamental de l'apprentissage statistique dit que l'on peut déterminer qu'une explication est « ϵ -optimale » d'un ensemble d'hypothèses \mathcal{H} avec grande probabilité $1 - \delta$ si la taille de l'échantillon d'apprentissage est au moins $\Omega\left(\frac{VCdim(\mathcal{H}) + \log(1/\delta)}{\epsilon^2}\right)$.

¹⁴  *Gros Tony et Dr. John (dilemme biais-variance)* | IA 12 | Science4All | G. Mitteau et L.N. Hoang (2018)

En pratique, les *data scientists* utilisent la *validation croisée*. La recherche du meilleur modèle est alors découpée en deux phases. Dans un premier temps, on considère les modèles plus simples qu'un certain niveau de complexité K — typiquement les modèles avec au plus K paramètres. Parmi ces modèles, on sélectionne celui qui explique le mieux le *training set*. Ensuite, on calcule les performances du modèle sélectionné sur un autre jeu de données appelé *test set*.

La *validation croisée* consiste alors à optimiser le degré de complexité K . Commençons par une valeur de K très faible. On est pour l'instant dans le régime de l'*underfitting*. Notre ensemble de modèles est trop rigide pour expliquer les données. Quand K augmente, les performances au *test set* s'améliorent. Ce n'est pas étonnant, puisqu'on s'autorise alors typiquement à plus de flexibilité dans nos modèles. Cependant, vient un point où ces performances cessent d'augmenter. On entre là dans le régime de l'*overfitting*. Alors que les performances du meilleur modèle s'améliorent sur le *training set*, les performances sur le *test set* se dégradent désormais. Trouver la valeur de K où a lieu cette transition est l'une des meilleures façons de lutter contre les dangers de l'*overfitting*.

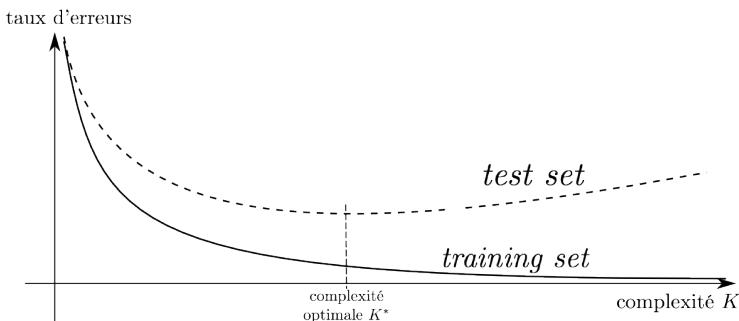


Figure 12.1. La courbe pleine représente le taux d'erreur au *training set*. Plus la complexité des modèles est grande, plus faible est ce taux d'erreur. La courbe en pointillée représente le taux d'erreur au *test set*. Elle représente la généralisabilité des paramètres calculés avec le *training set*. On voit qu'il y a un compromis à trouver. Trop de complexité nuit à la généralisabilité.

La quantité K dans la *validation croisée* est ce que les *data scientists* appellent un hyperparamètre, par opposition aux paramètres du modèle qui sont optimisés dans la première phase de la *validation croisée*¹⁵.

La *validation croisée* a toutefois ses limites. En particulier, elle suppose que le *test set* n'est utilisé que pour tester les hyperparamètres du modèle. Or, il arrive souvent que le *test set* soit utilisé pour tester un très grand nombre de modèles d'apprentissage différents, notamment lors des compétitions de *machine learning* comme ImageNet, CIFAR ou autres MNIST. Dès lors, le *test set* devient une sorte de *training set*. On risque alors l'*overfitting* sur le *test set*.

¹⁵ [La validation croisée | IA 13 | Science4All | La statistique expliquée à mon chat et L.N. Hoang \(2018\)](#)

La régularisation de Tibschirani

En 1996, le statisticien Robert Tibschirani eut l'idée d'introduire un autre hyperparamètre pour ajuster le *fitting* des *régressions linéaires*. Les régressions linéaires sont sans doute les techniques les plus utilisées en sciences. Dès la fin du XVIII^e siècle, Boscovich, Laplace, Legendre et Gauss avaient défini et utilisé de telles régressions pour gommer les erreurs de mesures des objets astronomiques et effectuer des prédictions malgré ces erreurs¹⁶.

En particulier, la régression linéaire permet d'expliquer une variable d'intérêt par p causes potentielles. Supposons que l'on dispose d'un échantillon de n données. Lorsque n est beaucoup plus grand que p , on peut appliquer tranquillement la régression linéaire. Cependant, dans de nombreux problèmes comme en génétique, à l'inverse, le nombre p de causes potentielles est plus grand que la taille n de l'échantillon. Dès lors, la régression linéaire multidimensionnelle est une très mauvaise idée, puisqu'elle conduira inéluctablement à un sévère *overfitting*.

Tibschirani proposa de mesurer la complexité de la régression linéaire et de la pénaliser. Typiquement, pour être retenue, une régression linéaire qui fait beaucoup intervenir beaucoup de causes devra expliquer bien mieux les données que d'autres régressions linéaires faisant peu intervenir peu de causes. La formalisation de ce principe a donné naissance à la régression dite LASSO¹⁷. L'astuce de la régression LASSO a depuis été généralisée et réutilisée dans de nombreux problèmes de *machine learning*. On parle alors de *régularisation*.

C'est sans doute une forme de régularisation qui permet à notre cortex cérébral et ses très nombreux neurones d'éviter en partie le piège de l'*overfitting*. Après tout, nous vivons environ 10^9 secondes, mais notre cerveau contient environ 10^{14} connexions neuronales. Le risque d'*overfitting* est énorme. Cependant, la régularisation permet d'ajuster le *fitting* des modèles vis-à-vis des échantillons. En particulier, les techniques de régularisation ont maintes fois démontré leur utilité en pratique, et sont devenues un ustensile incontournable de l'analyse de données, qu'il s'agisse de régressions linéaires, de classifications linéaires ou de réseaux de neurones.

Cependant, la régularisation a quelque chose de mystérieux. Pourquoi la régularisation serait-elle un guide pertinent vers les meilleures explications ? Le théorème fondamental de l'apprentissage statistique fournit une réponse bien incomplète à cette question. Un meilleur début de réponse nous vient de l'optimisation robuste.

¹⁶  Régressions et classifications linéaires | IA 9 | Science4All | L.N. Hoang (2018)

¹⁷ La régression linéaire consiste à décomposer l'explication d'une variable y par une combinaison linéaire de causes x_1, \dots, x_p et d'une erreur ϵ . Ainsi, on a $y = \beta_1 x_1 + \dots + \beta_p x_p + \epsilon$. L'approche classique consiste à déterminer les coefficients β_1, \dots, β_p qui minimisent la somme des carrés des erreurs de prédictions ϵ^2 pour les instances de l'échantillon. Tibschirani eut la brillante idée de minimiser une combinaison de cette somme de carrés et de la norme du vecteur β_1, \dots, β_p , typiquement la norme 1, c'est-à-dire la somme des valeurs absolues des β_i .

L'optimisation robuste

La motivation de l'optimisation robuste vient du constat suivant : toute donnée est minée d'imprécisions, voire d'erreurs. En *machine learning*, on parle de *bruits* dans les données. Ainsi, toute solution obtenue suite à une optimisation est sujette à n'être optimale que vis-à-vis de données erronées. Elle pourrait alors être totalement inadéquate vis-à-vis des vraies données.

Pour déterminer une solution performante malgré le bruit des données, l'optimisation robuste commence par identifier un ensemble d'incertitudes¹⁸. Cet ensemble est construit de sorte qu'avec une très grande probabilité, les données réelles se trouvent à l'intérieur de cet ensemble d'incertitudes. L'optimisation robuste consiste ensuite à choisir une solution qui convient pour toutes les données de l'ensemble d'incertitudes. Mieux encore, elle choisira la solution qui convient le mieux, même pour la pire donnée de l'ensemble d'incertitudes. Elle optimise le pire cas.

De façon étonnante, insister sur l'imprécision des données mesurées permet d'expliquer l'*utilité* du dysfonctionnement récurrent des neurones. Loin d'être une faille de fabrication, le manque de fiabilité des neurones pourrait en fait être un atout. Quand un neurone est victime d'un *bug*, il perturbe alors le signal, comme si l'on modifiait le jeu de données légèrement pour intégrer l'incertitude initiale concernant ces données. En affinant son modèle du monde encore et encore, notre cerveau explore ainsi tout un ensemble d'incertitudes et s'y ajuste — plutôt que de s'ajuster aux données brutes initiales.

Cette astuce est d'ailleurs utilisée aujourd'hui par de nombreux praticiens du *deep learning*. Ces derniers utilisent des réseaux de neurones artificiels pour découvrir des modèles pour expliquer des grands jeux de données. Ces praticiens éteignent aléatoirement une petite fraction des neurones de temps à autre, et testent la fonctionnalité de leurs réseaux de neurones malgré cela. Cette technique est appelée le *dropout*. Elle s'est révélée redoutablement efficace pour lutter contre l'*overfitting*.

La régularisation et l'optimisation robuste permettent donc toutes deux de lutter contre l'*overfitting*. Mais quel est le lien entre ces deux techniques ? Il se trouve qu'elles sont équivalentes. Dans de nombreux problèmes, on peut montrer que toute solution obtenue par régularisation peut être obtenue par le choix d'un certain ensemble d'incertitudes et l'application de l'optimisation robuste à cet ensemble. À l'inverse, pour un ensemble d'incertitudes donné, il est souvent possible de déterminer une régularisation équivalente. Autrement dit, l'efficacité de la régularisation peut s'expliquer comme une façon d'adresser le bruit dans les données¹⁹.

Mais il y a mieux. Beaucoup mieux.

¹⁸En dimension 1, ces ensembles d'incertitudes correspondent à des intervalles de confiance.

¹⁹Les démonstrations de ceci reposent souvent sur la théorie de la dualité en optimisation.

Bayes au secours de l'*overfitting**

La régularisation a une interprétation naturelle en termes bayésiens. Souvenez-vous de la formule de Bayes traduite dans le monde additif à l'aide des logarithmes. Celle-ci s'écrit comme suit :

$$\log \mathbb{P}[T|D] = \log \mathbb{P}[D|T] + \log \mathbb{P}[T] - \log \mathbb{P}[D].$$

Les méthodes de *machine learning* et d'optimisation robuste consistent généralement à sélectionner la théorie T la plus crédible, étant donné les données. Cette théorie est appelée Maximum-A-Posteriori (MAP). Elle maximise $\mathbb{P}[T|D]$, ce qui est équivalent à maximiser $\log \mathbb{P}[T|D]$.

Dès lors, la quantité $-\log \mathbb{P}[D]$ n'est pas pertinente car elle ne dépend pas de T . Calculer le MAP revient donc à maximiser la somme $\log \mathbb{P}[D|T] + \log \mathbb{P}[T]$. Le premier de ces deux termes est la log-vraisemblance. Il mesure la capacité de la théorie ou du modèle à expliquer les données. Le second de ces termes est le logarithme de l'*a priori*.

Cet *a priori* est équivalent à un terme de régularisation. Mieux encore, en exigeant le fait que la somme des probabilités *a priori* des paramètres soit égale à 1, on est alors tenté de distribuer ces paramètres selon une loi qui tend exponentiellement vite vers 0 quand les paramètres prennent de grandes valeurs. Voilà qui revient à des formes usuelles de régularisation ! En particulier, la régularisation est une conséquence de la formule de Bayes²⁰ !

Mieux encore, tous les hyperparamètres qui semblaient arbitraires, de l'ensemble d'incertitude à la régularisation, sont en fait autant d'éléments qui montrent l'inévitabilité — ou l'efficacité — des *a priori* dans la quête de modèles crédibles ! La régularisation marche, parce qu'elle nous force à introduire un préjugé. Or, on l'a vu, les préjugés forment l'un des piliers de la rationalité.

La *pure bayésienne* voit cependant une défaillance dans la manière dont la régularisation et l'optimisation robuste sont appliquées. La plupart des algorithmes de *machine learning* concluent avec un unique modèle, un seul choix de théorie T . Or, les méthodes d'*ensembling* ou de *bagging* nous invitent à combiner différents algorithmes de *machine learning*, notamment lorsqu'on les combine à l'aide de techniques comme *Adaboost*. En effet, elles montrent que la moyenne de bonnes théories donne souvent de meilleurs résultats que la meilleure de ces théories, notamment parce qu'il s'agit là d'une excellente façon de combattre l'*overfitting*. *Une forêt de modèles incompatibles est plus sage que chacun de ses arbres*²¹.

²⁰En particulier, LASSO revient à supposer un *a priori* distribué selon la loi de Laplace qu'on a vue au chapitre 8 !

²¹  *La sagesse des forêts* | IA 17 | Science4All | L.N. Hoang (2018)

Par exemple, lorsque Netflix organisa une compétition de *machine learning* avec 1 million de dollars en jeu, les grands vainqueurs avaient pris la moyenne de 800 modèles différents²² ! Or, prendre la moyenne des meilleures théories est précisément ce qu'impose la formule de Bayes !

De nombreux chercheurs se sont rendus compte de cela. En particulier, en 2016, Yarin Gal publie sa thèse *Uncertainty in Deep Learning*, dans laquelle Gal montre que de nombreuses techniques usuelles de *machine learning* peuvent se réinterpréter en termes bayésiens. C'est le cas notamment du *dropout* dont on vient de parler ! En effet, chaque faille d'un ensemble de neurones correspond à un modèle. La prédiction du réseau de neurones s'obtient alors en prenant la moyenne des prédictions des différents modèles, chacun d'entre eux se déduisant des failles d'un sous-ensemble de neurones.

Seules les inférences bayésiennes sont admissibles*

Il y a même un théorème qui insiste sur l'importance des préjugés : le *no-free-lunch theorem*. De façon grossière, ce théorème dit qu'il n'y a pas de meilleur algorithme d'apprentissage. Plus précisément, peu importe votre méthode pour choisir un modèle, il existera des problèmes pour lesquels votre méthode sera surpassée par d'autres, qui exploiteront typiquement des *a priori* adéquats.

Le théorème complémentaire à ce *no-free-lunch theorem* est l'admissibilité des inférences bayésiennes en théorie statistique de la décision. Imaginez qu'il existe une donnée fondamentale θ que vous ne connaissez pas. Cependant, vous recevez une information x qui est corrélée²³ avec θ . Il vous faut maintenant prendre une décision dont l'optimalité dépend de θ . Bien entendu, votre décision peut dépendre de x . Mais vous ne connaissez toujours pas θ . On suppose toutefois qu'étant donné θ , vous savez à quelle information x possible vous attendre. Comment décider ?

L'approche bayésienne consiste alors à d'abord remarquer que vous connaissez $\mathbb{P}[x|\theta]$. Cependant, vous ne connaissez pas θ . Que faire ? Utiliser ses préjugés bien sûr ! Le bayésien va donc considérer un *a priori* $\mathbb{P}[\theta]$, puis effectuer une inférence bayésienne pour déterminer $\mathbb{P}[\theta|x]$. Maintenant qu'il connaît les valeurs crédibles de θ , il peut optimiser sa prise de décision.

Le théorème d'admissibilité des inférences bayésiennes affirme alors que, quel que soit votre mécanisme de prise de décision, et quel que soit le préjugé du bayésien, il existera une valeur θ de l'information inconnue pour laquelle le bayésien s'en sortira mieux que vous²⁴. On dit que l'approche bayésienne est *admissible*. Bien sûr, ça ne veut pas dire qu'elle est meilleure que votre approche ; tout dépend de la valeur de θ .

²²  *The Netflix Prize | ZettaBytes | A.M. Kermarrec (2017)*

²³ En fait, ce n'est pas une hypothèse nécessaire.

²⁴ à moins que vous ne fassiez toujours aussi bien que le bayésien.

Mais ce n'est pas là l'aspect le plus intrigant du théorème d'admissibilité. Celui-ci prouve aussi que, sous certaines hypothèses additionnelles raisonnables, quel que soit votre mécanisme de décision, il existe un bayésien avec un certain préjugé $\mathbb{P}[\theta]$ dont les décisions seront *toujours* aussi bien ou meilleures que la vôtre, peu importe la valeur de θ ! Ou dit autrement, l'ensemble des mécanismes de décisions admissibles est exactement l'ensemble des approches bayésiennes²⁵. Toute alternative non-bayésienne sera en tout point inférieure à une méthode bayésienne !

Le rasoir d'Ockham déduit du bayésianisme !

Venons-en enfin à l'un de mes plus grands instants de jouissance dans mes méditations de la formule de Bayes. J'entrai dans le bureau de collègues à l'École Polytechnique Fédérale de Lausanne (EPFL) à l'heure du déjeuner. Deux de mes collègues discutaient du concept d'*Ockham learning*, qui est intimement lié au rasoir d'Ockham. Je me posai alors la question de l'interprétation bayésienne du rasoir d'Ockham. Se pourrait-il que la formule de Bayes implique le rasoir d'Ockham ?

Considérons un langage dans lequel nos théories seront décrites. Ce langage peut être le français, la logique formelle ou un langage de programmation informatique. Chaque théorie est alors décrite par une phrase (potentiellement très longue) dans ce langage, c'est-à-dire une suite finie de symboles du langage. Appelons T_n l'ensemble des théories décrites par une phrase à n symboles. Pour être conforme au bayésianisme, l'*a priori* sur ces théories doit être tel que la somme des crédences $\mathbb{P}[T_n]$ en les théories à n symboles est égale à 1. Autrement dit, le bayésianisme impose la condition suivante :

$$\mathbb{P}[T_1] + \mathbb{P}[T_2] + \mathbb{P}[T_3] + \mathbb{P}[T_4] + \dots = 1.$$

Or chaque quantité $\mathbb{P}[T_n]$ est positive, et il y a une infinité de telles quantités. La théorie des sommes infinies nous dit alors que, si la somme infinie de ces termes positifs est finie, c'est que, nécessairement, les termes $\mathbb{P}[T_n]$ de la somme deviennent arbitrairement petits pour de grandes valeurs de n . Alors que cette pensée traversa tout à coup mon esprit, je me jetai au tableau et écrivit :

$$\sum_{n=1}^{\infty} \mathbb{P}[T_n] < \infty \implies \lim_{n \rightarrow \infty} \mathbb{P}[T_n] = 0.$$

Or, écrire cela, c'est exactement dire que les théories qui nécessitent plus de symboles pour être décrites sont les moins crédibles *a priori*. Incroyable ! La formule de Bayes implique le rasoir d'Ockham !

²⁵  *Admissibility and complete classes* | P. Hoff (2013)

La formule de Bayes va même plus loin et nous précise à quel point les théories plus longues à décrire sont moins crédibles. En effet, le nombre de théories à n symboles est exponentiel en n . On en déduit que la crédence *a priori* d'une théorie à n symboles décroît exponentiellement en n ! Autrement dit, les théories plus sophistiquées ne sont donc pas juste moins crédibles ; elles sont exponentiellement moins crédibles !

Je fus vivement saisi par cette découverte délicieuse — d'autant que je n'avais pas encore rencontré le démon de Solomonoff à ce moment-là. Non seulement cette découverte confortait-elle encore plus la formule de Bayes, mais elle permettait aussi de lever le voile sur la mystérieuse acceptation commune du rasoir d'Ockham. Pour la *pure bayésienne*, le rasoir d'Ockham n'est pas un principe philosophique qu'il faut s'efforcer d'accepter ; le rasoir d'Ockham est un théorème mathématique du paradigme bayésien.

Références en français

- ▶ *Informatique et jeux* | Passe-Science | T. Cabaret (2016)
- ▶ *Jeu de go et intelligence artificielle* | À chaud | Science Étonnante | D. Louapre (2016)
- ▶ *Sommes-nous des simulations ? L'argument de la simulation de Nick Bostrom* | Argument Frappant | Monsieur Phi | T. Giraud (2016)
- ▶ *Deux (deux ?) minutes pour l'éléphant de Fermi & Neumann* | El Jj | J. Cottanceau (2018)

- ▶ *La loi de la chute des corps* | Relativité 13 | L.N. Hoang (2016)
- ▶ *La Terre est-elle au centre du monde ?* Relativité 14 | L.N. Hoang (2016)
- ▶ *Êtes-vous un hooligan politique ?* Démocratie 10 | Science4All | L.N. Hoang (2017)
- ▶ *Le poker résolu ! (ou non)* | Démocratie 15 | Science4All | L.N. Hoang (2017)
- ▶ *Les learning machines de Turing* | IA 7 | Science4All | L.N. Hoang (2018)
- ▶ *La sur-interprétation (overfitting)* | IA 11 | Science4All | C. Michel et L.N. Hoang (2018)
- ▶ *Gros Tony et Dr. John (dilemme biais-variance)* | IA 12 | Science4All | G. Mitteau et L.N. Hoang (2018)
- ▶ *La validation croisée* | IA 13 | Science4All | La statistique expliquée à mon chat et L.N. Hoang (2018)
- ▶ *Les explications ad hoc (dimension VC)* | IA 14 | Science4All | L.N. Hoang (2018)
- ▶ *Le théorème fondamental de l'apprentissage statistique* | IA 15 | Science4All | L.N. Hoang (2018)
- ▶ *La sagesse des forêts* | IA 17 | Science4All | L.N. Hoang (2018)
- ▶ *Régularisation et robustesse* | IA 18 | Science4All | L.N. Hoang (2018)

Références en anglais

- ➲ *The Righteous Mind: Why Good People are Divided by Politics and Religion* | Vintage | J. Haidt (2013)
- ➲ *Understanding Machine Learning: From Theory to Algorithms* | Cambridge University Press | S. Shalev-Shwartz and S. Ben-David (2016)
- ➲ *Uncertainty in deep learning* | PhD Thesis | University of Cambridge | Y. Gal (2016)
- ➲ *Regression shrinkage and selection via the lasso* | Journal of the Royal Statistical Society | R. Tibshirani (1996)

- ➲ *Spurious Correlations* | Tyler Vigen
- ➲ *Hack your way to scientific glory* | FiveThirtyEight (2015)
- ➲ *Admissibility and complete classes* | P. Hoff (2013)
- ➲ *Is anything real?* VSauce | M. Stevens (2013)
- ➲ *The fundamental theorem of statistical learning* | Wanda | L.N. Hoang (2017)
- ➲ *The Netflix Prize* | ZettaBytes | A.M. Kermarrec (2017)

Il y a trois types de mensonges : les mensonges, les sacrés mensonges et les statistiques.

Benjamin Disraeli (1804-1881)

Les politiciens utilisent les statistiques comme les ivrognes utilisent les lampadaires : pas pour l'illumination, mais pour le support.

Hans Kuhn (1919-2012)

13

Les faits sont trompeurs

Hôpital ou clinique ?

Vous êtes gravement malade. Vous faites vos recherches et découvrez que, pour cette maladie, l'hôpital possède un taux de survie de 50 %, alors que la clinique possède un taux de survie 80 %. Il n'y a pas photo. Il faut aller à la clinique plutôt qu'à l'hôpital... non ?

Clairement !

Pas si vite. Après des recherches supplémentaires, vous découvrez des statistiques qui discriminent deux types de patients : les peu malades et les gravement malades. À la clinique, les peu malades ont un taux de survie de 90 %. Pas mal. Cependant, à l'hôpital, le taux de survie pour ces mêmes malades est de 100 %. À l'inverse, les malades en phase critique meurent en grand nombre. L'hôpital parvient néanmoins à en sauver 40 %. C'est beaucoup mieux que la clinique qui n'en sauve que 10 %.

Mais, réfléchissez-y trente secondes. Il se passe quelque chose d'extrêmement étrange. L'hôpital soigne mieux que la clinique les patients peu malades *et* les patients très malades. Cependant, sur le total, c'est la clinique qui s'en sort mieux ! Comment est-ce possible ? Comme se fait-il que chacun des malades semble mieux s'en sortir à l'hôpital qu'à la clinique, alors même que le taux global de survie à la clinique est supérieur à celui de l'hôpital ? Et où aller se soigner ? Je vous invite à arrêter la lecture et à longuement y réfléchir.

Si vous êtes un peu perdu, sachez que c'est parfaitement normal. Ce que je viens de présenter, avec des données fictives, est le *paradoxe de Simpson*. Ce paradoxe est dévastateur. Il montre mieux que tout autre que les statistiques sont étonnamment trompeuses, et que les analyser requiert un énorme effort intellectuel et une grande expertise. Malheureusement, cette expertise est extrêmement rare ; et l'effort intellectuel moyen dans l'interprétation de statistiques est quasi-nul.

Dans son livre de cours de statistiques, Larry Wasserman écrit ainsi que ce paradoxe est « très déroutant pour de nombreuses personnes, y compris des statisticiens bien éduqués ». En introduction de sa vidéo sur le sujet, David Louapre¹, de la chaîne Science Étonnante, parle : « une fois que vous aurez vu la vidéo, je suis sûr que vous ne regarderez plus les chiffres de la même manière quand on vous montre des statistiques. »

Si je me débrouille bien, ce chapitre devrait bouleverser votre façon d'interpréter les résultats statistiques. Retenue, prudence et humilité devraient émerger comme étant les maîtres mots — et j'espère que les chapitres précédents ont déjà encouragé de tels réflexes. En particulier, ce qu'il faudra retenir, c'est surtout que des statistiques en apparence bonne et due forme sont en fait quasiment tout le temps très largement inconclusives. Beaucoup, beaucoup, beaucoup plus inconclusives que ce que l'on pourrait intuitivement croire.

La clé pour comprendre le paradoxe de Simpson est la notion de facteur de confusion. Dans notre cas, le facteur de confusion est la santé des patients au moment de la prise en charge. Si la clinique possède un meilleur taux de survie que l'hôpital, c'est simplement parce que ses patients sont en meilleur santé au moment de cette prise en charge. Le chiffre de 80 % de taux de survie à la clinique est donc essentiellement celui pour les patients peu malades. À l'inverse, le taux de 50 % de l'hôpital est si faible car il correspond essentiellement à des patients gravement malades.

J'ai longtemps eu l'impression que le paradoxe de Simpson n'en était pas un. Voir qu'il s'agissait d'une trivialité. Une fois le tableau de données bien rempli, il n'est pas difficile de voir que pour chaque type de patient, l'hôpital s'en sort mieux, et pourquoi la clinique a malgré tout de meilleures statistiques globales. Cependant, résoudre mathématiquement le problème une fois le tableau rempli n'est pas la difficulté posée par le paradoxe de Simpson. La vraie difficulté, c'est qu'en pratique, on n'a souvent accès qu'aux chiffres de 50 % et 80 %. On a alors tellement envie de conclure ! Pire encore, même si l'on prend le temps de la réflexion, il est souvent très difficile de penser aux bons facteurs de confusion pour échapper au piège du paradoxe de Simpson².

Dans tous les cas, il faut absolument résister à la tentation de tirer des conclusions. « Il faut se hâter de ne pas conclure », répète Étienne Klein.

¹ *Le paradoxe de Simpson* | Science Étonnante | D. Louapre (2015)

² *Satanés facteurs de confusion* | IA 17 | Science4All | L.N. Hoang (2018)

Corrélation n'est pas causalité

En 2012, la revue *The New England Journal of Medicine* publia un court article intitulé « *Chocolate Consumption, Cognitive Function, and Nobel Laureates* ». L'article suggérait que la consommation de chocolat avait des effets bénéfiques sur les capacités intellectuelles. Cette affirmation stupéfiante s'appuyait sur une corrélation très nette entre les consommations de chocolat (par habitant) de divers pays et les nombres de prix Nobel (par habitant) que ces pays ont gagnés. Rapidement, le graphe illustrant cette corrélation devint virale et fit le tour du Web. « *Croquez du chocolat pour avoir le Nobel* », titra *Le Figaro*.

Cependant, « il faut se hâter de ne pas conclure ». Une corrélation ne prouve absolument pas un lien de cause à effet. En particulier, si je ne peux pas exclure l'effet du chocolat sur les capacités intellectuelles, je suis certain qu'il y a d'autres explications bien plus crédibles de la corrélation entre les consommations de chocolat des pays et leur nombre de prix Nobel. Et je vous invite à y réfléchir.

Les statistiques deviennent particulièrement problématiques quand elles sont choisies par des politiques, des militants ou des avocats pour leurs intérêts personnels. En effet, en jouant avec les facteurs de confusion, il sera généralement possible de dénicher des statistiques qui, à première vue, semblent défendre telle ou telle position politique. Comme le dirait Winston Churchill : « quand je demande des statistiques sur le taux de mortalité infantile, ce que je veux est une preuve que moins de bébés sont morts quand j'étais premier Ministre que quand quelqu'un d'autre était premier Ministre. Ça, c'est la statistique politique. »

Par exemple, un phénomène étrange et récurrent est l'augmentation des chiffres de la criminalité juste après l'augmentation des effectifs policiers, comme si lutter contre la criminalité favorisait inéluctablement la criminalité. Une telle corrélation pourrait suggérer qu'investir dans les forces de l'ordre est une mauvaise idée. *La punition, ça ne marche pas*, rapporte-t-on ensuite à la télévision. Cependant, il faut bien se rendre compte que cette conclusion repose sur une interprétation fallacieuse des statistiques. En effet, il y a une autre façon beaucoup plus simple d'expliquer notre corrélation : l'augmentation des effectifs policiers augmente la fréquence des contrôles policiers. Il n'y a sans doute pas plus de criminels. Mais il y aura certainement plus de criminels arrêtés par la police. C'est pour cela que les chiffres de la criminalité augmentent inéluctablement.

De la même façon, si l'on rend les diagnostics médicaux plus abordables pour une grande population, on sera voué à détecter plus de patients malades. L'amélioration des moyens médicaux conduit donc souvent à une augmentation du nombre de malades ! C'est ainsi que l'on peut aussi expliquer une corrélation entre les enfants vaccinés et l'autisme. Ces enfants vaccinés étant mieux suivis médicalement, il y a de grandes chances que, s'ils sont atteints d'autisme, cet autisme soit diagnostiquée. À l'inverse, les enfants non vaccinés atteints d'autisme sont souvent peu suivis médicalement, et leur autisme a donc de bonnes chances de ne pas être diagnostiquée.

On a là des cas de biais de sélection, de survivants ou d'élimination — tous ces cas sont en fait des instances d'une espèce d'évolution darwinienne. Les chiffres révèlent alors davantage la manière dont ils ont été obtenus qu'un lien de cause à effet. Pour éviter tout contre-sens, il est très important de bien comprendre les chiffres qui nous sont présentés. Les chiffres de la criminalité ne sont pas les nombres de criminels, mais les nombres de criminels dont on a détecté l'existence. De même, les chiffres de l'autisme sont les nombres de cas diagnostiqués — pas le vrai nombre d'autistes. Attention aux ambiguïtés !

Ceci étant dit, la corrélation entre chocolat et prix Nobel ne semble pas être un cas de biais de sélection. Cherchons une autre explication plausible d'une corrélation. Confronté à une corrélation entre A et B, il est tentant de penser que c'est A qui implique B. Mais en fait, la notion de corrélation entre A et B est parfaitement symétrique. Si A est corrélé avec B, alors B est corrélé avec A. L'explication d'une corrélation peut donc simplement consister à inverser le lien de cause à effet que l'on voulait conclure.

Par exemple, les sportifs de très haut niveau sont souvent des personnes qui adorent la compétition. On pourrait croire que le fait d'évoluer à très haut niveau crée une émulation et stimule l'envie de relever des défis — c'est sans doute le cas. Cependant, l'explication la plus simple est sans doute plutôt que les sportifs n'ayant pas suffisamment l'esprit de compétition n'ont pas fourni les efforts suffisants pour accéder au très haut niveau. C'est l'esprit de compétition qui a permis l'accès au très haut niveau.

On retrouve cet effet dans de nombreuses situations. Par exemple, les politiciens au pouvoir ont souvent une soif d'être au pouvoir. Les mathématiciens de premier rang ont une appréciation profonde pour l'élégance mathématique. Les grands titres de l'actualité sont particulièrement spectaculaires et dramatiques. Dans tous ces cas, l'explication des corrélations réside dans la manière dont les politiciens, les mathématiciens et les grands titres ont été sélectionnés ou éliminés de manière systémique.

Un autre exemple est la corrélation entre le grand nombre de policiers dans les grands événements et le grand nombre d'incidents qui s'y produisent. Ce n'est alors pas la présence des policiers qui cause les incidents ; ce sont les risques d'incidents qui causent la présence de policiers. Dans tous ces exemples, la corrélation ne révèle pas tant le fait que A implique B, mais peut-être davantage le fait que B implique A.

Cependant, il n'est pas tout à fait clair que la possession de prix Nobel augmente la consommation de chocolat. Pour comprendre cette corrélation, intéressons-nous à une autre corrélation étrange : prendre des pauses au travail à l'extérieur réduit l'espérance de vie. Même si je n'en ai pas fait l'étude empirique, je suis prêt à parier que cette corrélation est statistiquement significative. Pourquoi ? Je vous invite à y réfléchir quelques instants avant de poursuivre la lecture.

L'explication de cette corrélation vient en fait d'une cause commune aux deux variables en jeu : le tabac. En effet, il est courant pour les fumeurs de prendre

bien plus de pauses au travail à l'extérieur que les non-fumeurs. Or, le tabac est une cause majeure du cancer du poumon. Par conséquent, ceux qui prennent des pauses à l'extérieur sont aussi ceux qui fument, qui sont aussi ceux qui ont davantage de risque de contracter le cancer du poumon et de mourir plus jeune. La corrélation s'explique donc là par l'existence d'une cause commune.

Se pourrait-il que la corrélation entre chocolat et prix Nobel en soit de même ? Y a-t-il quelque chose qui fait que certains pays consomment du chocolat, et qui fait aussi que ces mêmes pays gagnent des prix Nobel ? Très probablement, oui. En effet, les pays consommant du chocolat et recevant des prix Nobel sont tous des pays très développés. Les habitants de ces pays jouissent d'une très grande qualité de vie, d'une grande consommation de produits de luxe et de grandes universités. La corrélation s'explique donc là aussi par l'existence d'une cause commune : la richesse. On dit que la richesse est un facteur de confusion qui explique la corrélation entre chocolat et prix Nobel³.

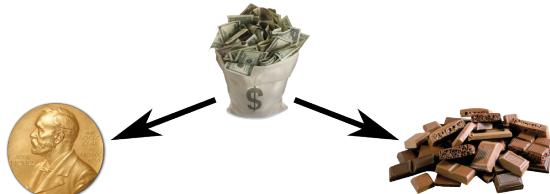


Figure 13.1. La corrélation entre chocolat et prix Nobel s'explique par un facteur de confusion, la richesse, qui cause à la fois la consommation de chocolat et le gain de prix Nobel. Cette représentation graphique correspond d'ailleurs à un réseau bayésien, dont on reparlera au chapitre 17.

Penser aux bons facteurs de confusion est peut-être la tâche la plus ardue des statistiques. Jusque-là, on n'en a vu que des cas relativement simples. Mais les facteurs de confusion sont parfois bien plus subtiles à dénicher.

Cherchez les facteurs de confusion

Par exemple, d'après une excellente vidéo Ted-Ed⁴, une étude anglaise a montré que, sur une période de 20 ans, les fumeurs avaient plus survécu que les non-fumeurs. Une telle étude pourrait être mise en avant, encore et encore, par des politiciens ou des avocats financés par l'industrie du tabac. Il n'y aurait pourtant pas grand-chose à redire aux statistiques. Le problème, c'est leur interprétation. En particulier, il ne faut absolument pas en déduire que le tabac

³ Chocolat, corrélation et moustache de chat | La statistique expliquée à mon chat | L. Maugeri, G. Grisi et N. Uyttendaele (2016)

⁴ How statistics can be misleading | Ted-Ed | M. Lidell (2016)

est bénéfique pour la santé. Pourquoi ? Parce qu'il y a un important facteur de confusion. Le voyez-vous ?

Un autre exemple de cette même vidéo est celui de la peine de mort en Floride. Alors que le débat sur le racisme contre les noirs fait rage outre-Atlantique — en témoigne le mouvement *Black Lives Matter* — une étude statistique de la peine de mort en Floride a été réalisée. Les suspects noirs sont-ils plus susceptibles d'être condamnés à la peine capitale ? Les statistiques indiquaient que non, ce qu'un candidat politique niant l'inégalité raciale mettra rapidement en avant.

Cependant, il y a là un facteur de confusion : la couleur de peau de la victime. Comme les coupables sont très souvent de la même couleur de peau que leurs victimes, il se trouve que les suspects noirs sont plus souvent jugés pour un meurtre d'une victime noire, alors que les suspects blancs sont plus souvent jugés pour celui d'une victime blanche. Or, à couleurs de peau du suspect égales, les juges sont plus cléments lorsque la victime est noire. À l'inverse donc, à couleurs de peau des victimes égales, les suspects noirs sont en fait significativement plus souvent condamnés à mort que les suspects blancs. Telles seront les statistiques mises en avant par les candidats qui se plaignent de l'inégalité raciale devant la loi.

De la même façon, dans un épisode de Podcast Science⁵, le chercheur en criminologie André Kuhn affirme que le fait que, dans la très grande majorité des pays, la proportion de criminels chez les étrangers est supérieure à celle chez les locaux est là encore une statistique biaisée par le paradoxe de Simpson. André Kuhn affirme ainsi qu'à âge, sexe et niveau socio-économique égaux, un étranger est en fait tout aussi probable d'être criminel *a priori* qu'un local.

La cause de la différence statistique sur l'ensemble des populations étrangères et locales ne vient en fait pas d'une différence fondamentale de nature entre étrangers et locaux, comme le laissent souvent entendre certains politiciens. L'analyse de Kuhn montre que cette différence vient en fait de la différence démographique entre les deux populations : les étrangers sont plus souvent de jeunes hommes peu fortunés, comparés à des locaux qui sont comparativement plus souvent de riches femmes âgées. À la lumière de cette réflexion, la statistique qui semblait devoir incriminer les étrangers perd complètement sa pertinence !

Maintenant que l'on a vu divers exemples, revenons-en aux fumeurs. Avez-vous su déterminer le facteur de confusion qui explique le meilleur taux de survie chez les fumeurs ?

Je vous invite à longuement y réfléchir dès que vous aurez effectué une pause dans la lecture de ce livre. Profitez-en pour ressentir l'étendue de votre ignorance et pour vous familiariser avec l'extrême difficulté de l'interprétation des statistiques que cause le paradoxe de Simpson.

⁵ Interview d'André Kuhn : les sciences criminelles | Podcast Science (2011)

La régression à la moyenne

Continuons notre quête de facteurs de confusion peu évidents. Faut-il réprimander ou encourager ? Des instructeurs de l'armée de l'air israélienne constatèrent que les pilotes ayant été réprimandés progressaient nettement juste après. Cependant, à leur grande stupéfaction, ceux qui avaient reçu des louanges ne progressaient pas. Pire encore, les encouragements ou félicitations semblaient les faire régresser, comme s'ils se reposaient alors sur leurs lauriers !

Cependant, diverses études scientifiques suggèrent au contraire que les récompenses sont plus efficaces que les punitions pour enseigner. En particulier, tout enseignant considérera sans doute qu'il est faux de penser que les encouragements sont néfastes à l'apprentissage. Qu'en est-il donc ? La communauté scientifique aurait-elle tort ? Ou n'y aurait-il pas un facteur de confusion dans l'expérience proposée par les instructeurs militaires ?

L'ingrédient incontournable pour éviter le piège du paradoxe de Simpson est le préjugé de la *pure bayésienne*⁶. Raisonner sans préjugé, ou sans faire appel à des modèles extérieurs aux données, c'est sauter à pieds joints dans le piège des facteurs de confusion ! On a ainsi vu qu'il y avait une différence *a priori* entre l'hôpital et la clinique (l'état des patients au moment de la prise en charge), entre étrangers et locaux (l'âge, le sexe et le niveau socio-économique) et entre les suspects noirs et blancs (la couleur de peau de la victime). Y aurait-il une différence *a priori* entre les pilotes réprimandés et les pilotes encouragés ?

Oui ! Les pilotes réprimandés l'ont été parce qu'ils ont été particulièrement mauvais. Tandis que ceux qui ont été félicités l'ont été parce qu'ils ont été particulièrement bons. Cependant, un pilote qui a fait une bêtise un jour ne la fera sans doute pas le lendemain, qu'il ait été réprimandé ou non. De même, celui qui aura réussi une prouesse exceptionnelle un jour aura bien du mal à la répéter le lendemain⁷.

Le phénomène dont on vient de révéler l'existence est donc un cas particulier du paradoxe de Simpson, que certains appellent la *régression à la moyenne*. On peut le résumer par la maxime « après la pluie le beau temps ». S'il fait particulièrement moche aujourd'hui, la probabilité que demain soit encore plus moche est faible, parce que ce à quoi demain va être comparé est déjà exceptionnellement moche. Ou à l'inverse, les fils de Zinédine Zidane auront peu de chance de dépasser le niveau de leur père — même si je le leur souhaite ! — parce que ce à quoi on les compare est déjà un joueur d'exception. Ce phénomène explique aussi pourquoi les ministres dont la mission est de rectifier une situation anormalement mauvaise ont de bonnes chances de s'en sortir la tête haute — qu'ils aient agi ou non.

⁶En particulier, pour expliquer $\mathbb{P}[B|A] \neq \mathbb{P}[B|\text{non } A]$ on va chercher à déterminer les caractéristiques Z à invoquer telles que $\mathbb{P}[Z|A]$ et $\mathbb{P}[Z|\text{non } A]$ diffèrent grandement, et telles que $\mathbb{P}[B|A, Z] \approx \mathbb{P}[B|\text{non } A, Z]$.

⁷  *Is Punishment or Reward More Effective?* Veritasium | D. Muller (2013)

Le paradoxe de Stein

En 1955, Charles Stein découvrit une mystérieuse solution au problème de la *régression à la moyenne*. Imaginez qu'il faille estimer les niveaux des pilotes en fonction de leurs performances. De façon intuitive, on risque de surestimer les niveaux des pilotes performants et de sous-estimer ceux des pilotes maladroits lors de leurs performances. Peut-on éviter le piège du paradoxe de Simpson ?

Oui, répond Stein. Au lieu d'estimer naïvement le niveau d'un pilote à partir uniquement de sa performance, on peut faire strictement mieux en l'estimant à l'aide de sa performance individuelle *et* de la performance du groupe. C'est ce que l'on appelle le *paradoxe de Stein*⁸. Et il a de quoi surprendre. Comment est-ce possible qu'en faisant appel aux performances des autres, il soit possible d'améliorer les prédictions concernant un pilote donné ?

En fait, le paradoxe de Stein est beaucoup plus étrange que cela. Il montre aussi que l'on peut strictement améliorer les estimations du niveau d'un pilote, de la consommation de chocolat d'un pays et du taux de survie d'un hôpital, en invoquant une combinaison des estimations naïves des niveau, consommation et taux de survie ! Et ce qu'il y a de très mystérieux, c'est que l'amélioration des estimateurs sera garantie, même si les unités de mesures des niveau, consommation et taux de survie sont incompatibles !

C'est extrêmement troublant et contre-intuitif. Même s'il n'y a absolument aucun lien causal entre ces trois paramètres, et même si les échelles de ces paramètres n'ont rien à voir, le paradoxe de Stein montre que, pour estimer chacun des paramètres, il y aura *toujours* un gain à invoquer les deux autres paramètres. Autrement dit, en un sens rigoureux, s'il permet l'interprétabilité des modèles, *le morcellement du savoir est statistiquement inadmissible*.

Comme souvent, il y a une manière bayésienne d'élucider le mystérieux paradoxe de Stein. L'astuce est de rajouter des concepts abstraits explicatifs qui vont lier les différentes quantités à estimer. Cependant, ces concepts ne sont pas vraiment des facteurs de confusion ; il s'agit davantage de facteurs de concision. En particulier, ce que montre le paradoxe de Stein, c'est que ces facteurs de concision sont *indispensables* pour identifier les modèles les plus crédibles.

⁸Formellement, pour $1 \leq i \leq n$, on tire indépendamment $x_i \leftarrow N(\theta_i, 1)$ selon une loi normale dont la moyenne θ_i est inconnue (et peut très bien être la consommation de chocolat et le taux de survie d'un hôpital). L'estimateur naïf (du moindre carré) consiste à estimer $\hat{\theta}_i^{naif} = x_i$. Mais on peut faire mieux, par exemple avec l'estimateur de James-Stein $\hat{\theta}_i^{JS} = \left(1 - \frac{n-2}{\|x\|_2^2}\right)x_i$. En effet, pour $n \geq 3$, quelle que soit la valeur de θ , l'erreur quadratique espérée de $\hat{\theta}^{JS}$ sera inférieure à celle de θ^{naif} , c'est-à-dire

$$\forall \theta, \quad \mathbb{E} \left[\|\theta^{JS} - \theta\|_2^2 \right] \leq \mathbb{E} \left[\|\theta^{naif} - \theta\|_2^2 \right].$$

On dit que l'estimateur naïf est inadmissible car il est strictement dominé par un autre estimateur. Il se trouve que l'estimateur de James-Stein est aussi inadmissible, alors que tous les estimateurs bayésiens sont admissibles.

De façon plus générale, on a tendance à vouloir séparer les champs de la connaissance. Qu'on laisse la philosophie aux philosophes, l'économie aux économistes, la physique aux physiciens et les mathématiques aux mathématiciens, non ? Non. D'après le paradoxe de Stein, l'unification de toutes les théories n'est pas juste une chimère de théoriciens. C'est une étape obligée dans la quête de modèles crédibles⁹ — d'où l'importance du fait que le bayésianisme est une philosophie *unifiée* du savoir.

Dans le cas des pilotes uniquement, un exemple de facteur de concision est le niveau moyen de tout pilote. Un modèle bayésien possible consiste ensuite à supposer que le niveau d'un pilote est égal au niveau moyen plus une certaine fluctuation aléatoire. La performance du pilote est ensuite une seconde fluctuation aléatoire du niveau du pilote. En fait, ce que l'on vient de construire là est un réseau bayésien, dont l'étude est au cœur de nombreuses recherches modernes en intelligence artificielle. On y reviendra.

Quoi qu'il en soit, il est important de noter que ce modèle possède nettement moins de paramètres qu'un modèle qui traiterait les cas des différents pilotes séparément. Il est donc *a priori* beaucoup plus crédible. Mais surtout, en ajoutant un *a priori* sur le niveau moyen d'un pilote, ce modèle se prête à l'inférence bayésienne. En particulier, en postulant un certain *a priori* raisonnable, cette inférence bayésienne nous conduit à quelque chose de similaire à l'estimation statistique de Stein. Autrement dit, l'étrange paradoxe de Stein disparaît quand on cherche à rendre le problème conforme aux principes du bayésianisme !

L'échec de la stratification endogène

Toutefois, aujourd'hui encore, l'approche bayésienne n'a pas toujours les faveurs des praticiens. Beaucoup préfèrent des méthodes dites de *stratification*. Ces méthodes consistent à distinguer des sous-populations comparables. Par exemple, mieux vaut comparer les étrangers et les locaux de même âge, sexe et niveau socio-économique, les suspects noirs et blancs dont les victimes ont la même couleur de peau, et les fumeurs et les non-fumeurs de même âge.

Autour des années 2010, certains statisticiens ont toutefois souligné les difficultés posées par le choix « à la main » des strates considérées. Ce choix est souvent arbitraire, et donne l'impression d'un manque d'objectivité — même si un bayésien n'attribue aucune valeur à cette objection ! Il peut aussi être injustifié ou insuffisant, ce qui conduirait à de mauvaises conclusions. Enfin, il nécessite une intervention humaine, et est donc coûteux en temps et en travail. Ne serait-il pas possible d'automatiser le choix des strates ? Ainsi naquit la *stratification endogène*.

⁹Même s'il est crucial pour chacun de prendre la mesure de l'étendue de son ignorance, surtout dans des domaines qui ne sont pas son champ d'expertise.

En 2015, j'ai eu l'occasion d'assister à un séminaire de statistiques au MIT. Plusieurs des statisticiens les plus reconnus du monde étaient dans la salle. Le présentateur, Alberto Abadie, s'attarda alors sur un article de la presse¹⁰ qui présentait les résultats de la stratification endogène réalisée par Sara Goldrick-Gab et ses collaborateurs¹¹, appliquée à des données concernant des étudiants du Wisconsin. La stratification endogène divisa les étudiants en trois catégories : ceux dont les chances de finir le collège semblaient faibles au moment d'intégrer l'université, ceux dont les chances étaient moyennes, et ceux dont les chances étaient grandes. Elle montra que, pour les étudiants du premier groupe, le fait de percevoir des bourses avait eu un net effet bénéfique. Les boursiers qui semblaient peu à même de finir l'université se sont beaucoup mieux débrouillés que les non-boursiers peu à même de finir l'université.

Jusque-là, rien de troublant. Cependant, cette même stratification endogène montrait aussi que, pour les étudiants du troisième groupe, ceux dont les données au moment d'entrer à l'université prédisaient une grande probabilité de succès, le fait de percevoir des bourses avait un effet négatif ! Comme si, lorsqu'on est bon et qu'on nous donne en plus de l'argent, c'est là que les chevilles enflent...

Ou pas. À la surprise de plusieurs personnes dans la salle, Abadie montra que les conclusions de la stratification endogène étaient infondées. L'automatisation de la stratification avait créé sa propre *régression à la moyenne* ! La conclusion de la stratification endogène n'était pas due aux données ; elle était un artefact de la stratification endogène !

Incroyable ! On était en 2015, et les meilleurs statisticiens de la planète découvraient encore que des modèles statistiques pourtant relativement simples, utilisés nonchalamment par d'autres statisticiens de premier rang, étaient fondamentalement erronés !

Quelques semaines plus tard, je pris quelques jours de congés et je partis voir un ami dans la Silicon Valley. Mon ami travaillait alors chez l'un des géants du web. Je partageai ma fascination naissante pour les difficultés piégeuses et subtiles des statistiques. Quelques mois plus tôt, Ramesh Johari, un professeur de Stanford, m'avait déjà stupéfait quand, dans un séminaire, il prouva que la méthode de la *p-value* finirait toujours par rejeter une hypothèse, pourvu que l'on collecte suffisamment de données pour pouvoir conclure — on en a déjà parlé au chapitre 5 !

Mon ami fut plus intrigué par la critique de la stratification endogène par Abadie que par celle de la *p-value* par Johari. Il me fit répéter mes explications une fois. Puis une deuxième fois. Puis une troisième fois. Et puis, tout à coup, il me lança : « Mais je crois que c'est ce qu'on vient de faire pour tester notre nouveau produit ! »

¹⁰ *Research Student Aid Before You Reform* | Chronicle of Higher Education | A. Kelly (2012)

¹¹ *Need-based financial aid and college persistence: Experimental evidence from Wisconsin* | S. Goldrick-Rab, D. Harris, J. Benson et R. Kelchen (2012)

En effet, pour ce faire, mon ami avait mesuré des nombres de clics d'utilisateurs. Il les avait ensuite comparés avec des nombres de clics pour l'ancien produit. Il n'avait pas mesuré de différence statistiquement significative — Johari dirait qu'il n'a pas attendu suffisamment longtemps ! Cependant, la stratification endogène lui avait permis de conclure.

D'après son analyse, le nouveau produit fonctionnait mieux que l'ancien pour les utilisateurs dont le taux de clics était originalement faible, et moins bien pour ceux dont le taux de clics était originalement grand. Mon ami avait même trouvé des explications *a posteriori*, notamment fondées sur les origines géographiques des différents utilisateurs. Voilà qui le confortait dans la conclusion pourtant fondamentalement biaisée de la stratification endogène.

Quelques jours plus tard, je lui envoyai la publication d'Abadie et il me répondit : « j'ai analysé une des expériences, et j'ai [vu] qu'on a utilisé une variable endogène [...] pour classifier les résultats. L'expérience était bien, mais la façon d'analyser les groupes après était incorrecte [...]. La solution est d'utiliser une variable exogène qui ne modifie pas le regroupement. Merci de me faire réfléchir à ça ! J'ai partagé cette analyse, donc les prochaines expériences seront bien [analysées]... »

Quel bonheur pour un théoricien comme moi de voir de tels effets immédiats dans la pratique ! Malheureusement, je n'ai jamais reçu de remerciements financiers de la part de son entreprise...

Randomisons !

Le paradoxe de Simpson montre qu'une bonne analyse des données doit étudier des facteurs de confusion qui ne sont pas dans ces données. Mais si nous ne disposons que des données, comment trouver ces facteurs de confusion ? Comment lutter contre le paradoxe de Simpson ?

J'ai beau avoir peint un portrait maléfique de Ronald Fisher, dont le dogmatisme de sa position et l'acharnement contre les avis contraires ont certainement été dommageables pour l'avancée des statistiques, il y a une expression bien à lui qu'il faut souligner, répéter et célébrer : « Randomisons. »

En effet, pour tester un produit sur une population, il faut absolument comparer la sous-population exposée à la variable ou au produit, à celle qui ne l'a pas été. La deuxième sous-population est ainsi appelée groupe de contrôle ou groupe témoin. Cependant, pour combattre tout facteur de confusion, il ne faut pas laisser l'affectation des sujets dans les sous-populations au hasard. Ou plutôt si, justement ! Il faut la laisser au pur hasard. Car si l'affectation se fait de manière systémique, c'est-à-dire déterminée par un système choisi par le scientifique ou conséquent du contexte environnemental, on peut être sûr qu'il y aura des facteurs de confusion. Et même si ces facteurs de confusion ont un

rôle négligeable, on ne pourra jamais être sûr que ce n'est pas le cas, et on ne pourra donc jamais avoir pleinement confiance en les résultats de l'expérience.

Dans la tradition fishérienne, le standard de la médecine pour le test de nouveaux médicaments est le test randomisé en double aveugle. Dans ce test, chaque patient est traité par un médecin avec le nouveau médicament ou un faux médicament. De façon cruciale, le patient *et* le médecin ne savent pas si le médicament administré est le nouveau médicament ou le faux médicament. Il s'agit là d'un point crucial. En effet, d'un côté, il y a l'effet placebo. Si un patient croit qu'il a eu un médicament alors qu'il a eu un faux, alors il y aura des effets physiologiques positifs sur sa santé. Le patient ira mieux que s'il savait qu'il avait eu un faux médicament.

Mais de l'autre côté, il faut aussi absolument que le médecin ne sache pas s'il a administré le nouveau médicament ou un faux. En effet, l'expérience a montré que les médecins qui administraient volontairement un faux médicament faisaient signe de moins de confiance et d'enthousiasme, si bien que l'effet placebo sur le patient était moindre. Le test randomisé en double aveugle permet de contrôler de tels facteurs de confusion, et ainsi d'aligner les résultats de l'expérience avec le phénomène que l'on veut vraiment étudier : l'effet intrinsèque du nouveau médicament¹².

Plus généralement, ce contrôle des facteurs de confusion est la raison d'être des bonnes vieilles expériences scientifiques, celles réalisées dans la tradition initiée par Galilée. Dans l'idéal, ces expériences scientifiques répliqueraient un grand nombre de fois deux types d'expériences quasi-identiques. La seule distinction entre les deux types d'expériences serait uniquement la variable dont on cherche à déterminer l'effet. On étudie bel et bien la variable, « toutes choses égales par ailleurs¹³ ».

Cependant, de telles expériences scientifiques ne représentent qu'une infime portion de nos expériences quotidiennes. Même les scientifiques se satisfont de répliquer des expériences à des moments et des lieux différents, avec du matériel différent. Pire encore, l'avènement du *Big Data* annonce la collecte de données à tout va, eu égard à la randomisation préconisée par Ronald Fisher. Or, s'il y a bien un point sur lequel le paradoxe de Simpson nous amène à insister, c'est

¹²Techniquement, la randomisation revient à comparer $\mathbb{E}_Z[\mathbb{P}[B|A, Z]]$ et $\mathbb{E}_Z[P[B|\text{non } A, Z]]$, par opposition à comparer $\mathbb{P}[B|A] = \mathbb{E}_Z[\mathbb{P}[B|A, Z]|A]$ et $\mathbb{P}[B|\text{non } A] = \mathbb{E}_Z[P[B|\text{non } A, Z]|\text{non } A]$. Et contrairement à la critique qu'on a soulevée au chapitre 5, ici A est bien le fait de se mettre à prendre un médicament, pas le fait de l'avoir pris dans le passé. Néanmoins, il reste alors possible de conclure que A n'a pas d'effet, alors qu'il est peut-être salvateur pour certaines valeurs de Z , et catastrophique pour d'autres. Pire encore toute quantité $\mathbb{E}_Z[\mathbb{P}[B|A, Z]]$ dépendra de la distribution de Z considérée. Or on insiste rarement sur le biais de cette distribution.

¹³Il est intéressant de noter que cette notion presuppose d'avoir précisé toutes les choses susceptibles de varier. Ainsi, la chute des corps dépend à la fois de la hauteur de l'objet, mais aussi de son énergie potentielle. Si l'on en croit la physique classique, il est alors impossible de tester uniquement la variation des hauteurs, toutes énergies potentielles égales par ailleurs. Il semble en fait que la notion « toutes choses égales par ailleurs » presuppose l'utilisation d'un modèle, et est donc, comme tout autre aspect de la connaissance, fondamentalement subjectif.

le fait que les données brutes ne disent pas tout. Le contexte dans lequel les données ont été prélevées est crucial à l'analyse des données. Quel est l'âge, le sexe et le niveau socio-économique des étrangers ? Quelles sont les victimes des criminels blancs ? Qui sont les patients de la clinique ? Le patient savait-il que c'était un placebo ? Pourquoi le pilote a-t-il été réprimandé ?

Les statistiques collectées sans randomisation à la Fisher sont donc minées de pièges contextuels. *Sans conteste, sans contexte, c'est la mauvaise probabilité qu'on teste.* Il ne faut absolument pas y faire confiance aveuglément. Il faut absolument les interpréter avec une prudence extrême. Il faut absolument s'en méfier. Pour la *pure bayésienne*, il faut absolument les interpréter à la lumière de (plusieurs) modèles *a priori* crédibles, lesquels suggéreront les facteurs de confusion adéquats pour l'analyse des statistiques. Même là, il ne faut pas perdre de vue la crédibilité limitée de ces modèles. « Tous les modèles sont faux. » En particulier, « il faut se hâter de ne pas conclure ».

Cependant, ce scepticisme envers les statistiques ne doit sûrement pas être interprété comme une acceptation des alternatives. Au contraire. Si même les statistiques sont capables à tout moment de nous induire gravement en erreur, toute conviction dont on est incapable de trouver des justifications statistiques doit être surveillée d'encore plus près !

Le retour du mouton noir d'Écosse

Voilà qui nous permet d'en revenir à l'histoire du biologiste, du physicien et du mathématicien voyageant en Écosse et y découvrant un mouton noir, dont on a parlé au chapitre 4. Souvenez-vous. Le mathématicien s'était alors moqué de la généralisation excessive du physicien, car le physicien avait conclu que l'autre moitié du mouton, celle que l'on ne voit pas, devait être noire.

Pourtant, il semble que l'explication du physicien ne soit pas complètement farfelu. Après tout, il y a une très nette corrélation entre la couleur d'un animal d'un côté et sa couleur de l'autre côté. Rares sont les chats dont exactement une moitié est blanche, l'autre est noire. Et plus faible encore serait la probabilité de voir ce chat sous un angle dont la partie visible est d'une seule et unique couleur. La généralisation du physicien ne semble pas déraisonnablement excessive.

Cependant, la couleur de la moitié d'un animal ne semble pas causer la couleur de l'autre moitié, et vice-versa. Comment expliquer le fait que les couleurs d'un animal d'un côté sont presque toujours les mêmes que celles de l'autre ? Là encore, il y a une cause commune : les gènes de l'animal en question. En fait, si on remonte dans le temps, l'animal était à l'origine une seule cellule, qui contenait notamment une molécule appelée ADN. Cette cellule s'est ensuite répliquée, copiant sa molécule d'ADN à l'identique. C'est ainsi que le même ADN se retrouve dans toutes les cellules de l'animal, et qu'il détermine, via l'expression de ses gènes, les couleurs de l'animal de ses deux côtés.

Ce qui est amusant, toutefois, c'est que cette explication est une explication très moderne. Après tout, la structure de l'ADN n'a été découverte qu'en 1953. Pendant des millénaires, l'explication à cause commune que nous avons avancée ci-dessus était hors de portée. Toutefois, il est difficile d'imaginer quiconque penser que la corrélation entre les couleurs des animaux des deux côtés pouvait s'expliquer par un lien de cause à effet. Comment cette corrélation était-elle alors expliquée ? Comment expliquer le fait que les animaux sont habillés d'une couleur homogène, sans faire appel à la molécule d'ADN ?

Qu'est-ce qu'un chat ?

En 2012, Google faisait les titres de l'actualité avec une annonce étrange et perturbante : l'intelligence artificielle de Google, disaient les titres, avait découvert le concept de chat¹⁴ ! Beaucoup ont sans doute trouvé la nouvelle banale. Mais pour moi, il s'agissait là d'une percée monumentale et stupéfiante du *machine learning*, peut-être encore plus impressionnante et inattendue que la victoire d'AlphaGo contre Lee Sedol quatre ans plus tard.

Pour comprendre, il faut dire que l'intelligence artificielle de Google est un réseau de neurones artificiel, muni de capteurs qui lui permettent de « voir » des images numérisées. Google montra dix millions d'images, sans contexte, à son réseau de neurones artificiel. Puis, Google fit passer une espèce d'imagerie par résonance magnétique (IRM) pour mesurer les activations en temps réel de son réseau de neurones artificiels lorsque ce réseau de neurones fut exposé à d'autres images. Google se rendit alors compte que certains de ces neurones s'activaient *grossost modo*, si et seulement si, l'image qui lui fut montrée contenait l'image d'un chat !

Ce qui est vraiment remarquable, c'est que ce ne fut pas dans ce but que l'intelligence artificielle de Google avait été conçue. L'objectif de cette intelligence artificielle, c'était d'analyser, synthétiser et expliquer le contenu des images de la manière la plus pertinente et la plus efficace qui soit. L'intelligence artificielle devait concevoir un modèle des images qu'elle voyait. Pour ce faire, il lui fallait, entre autres, expliquer des corrélations récurrentes entre les couleurs de certains pixels de l'image. Par exemple, quand certains pixels de l'image étaient arrangés en forme d'œil, il y avait très souvent une copie semblable de ces pixels légèrement à gauche ou à droite. Une image d'œil est souvent accompagnée d'une autre image d'œil. Comment expliquer qu'un œil vient rarement seul ?

Ce qui est stupéfiant, c'est que l'explication semble grandement correspondre à celle que la plupart d'entre nous donnerait : parce que les animaux pris en photo ont souvent deux yeux. Le fait que les hommes, les chiens et les

¹⁴  *Google's Artificial Brain Learns to Find Cat Videos* | Wired | Liat Clark (2012)

chats ont (presque) tous deux yeux est une affirmation qui nous paraît désormais si anodine qu'on ne prend guère le temps de la méditer. Notre quotidien nous y habite tellement qu'on n'en cherche même plus d'explications — alors qu'il s'agit d'une question biologique fascinante qui requiert notamment la compréhension mathématique de la parallaxe !

Mais ce qu'il y a d'encore plus fascinant à mon sens, c'est que nous puissions être (globalement) d'accord sur ce que sont les hommes, les chiens et les chats. Au point que l'on ne se rend même plus compte qu'il s'agit là de concepts à la fois abstraits et flous ! Qu'est-ce qu'un chat ? Comment définir le concept de chat ? Et surtout, est-ce qu'un chat *existe* ?

Pour comprendre l'étrangeté du concept de chat, on peut commencer par insister sur l'imprécision de la définition. On pourrait typiquement penser qu'un chat, c'est ce qu'on obtient après accouplement de deux chats. Ou dit autrement, les parents d'un chat sont des chats. Jusque-là, rien de choquant. Il s'agit même sans doute là d'une évidence béante. Et pourtant.

Prenons un chat d'aujourd'hui. Ses parents sont donc des chats. Mais les parents de ses parents le sont donc aussi. Ainsi que les parents des parents de ses parents. Et ainsi de suite. Sauf qu'il n'y a pas de limite à cette régression dans le temps ! Selon ce raisonnement, on est forcé de remonter l'arbre phylogénétique de la vie jusqu'à des époques où il n'y avait pas de chat ! En effet, si on remonte le temps de quelques centaines de millions, voire de quelques milliards d'années, on tombe dans une époque où il n'y avait pas de mammifère, pas de vertébré, voire pas de cellule eucaryote ! Dire que les parents d'un chat sont des chats, ou dire que les chats sont le résultat d'accouplements entre chats sont donc des affirmations logiquement incohérentes¹⁵.

J'en vois certains parmi vous souffler le concept d'ADN pour définir la notion de chat. Cependant, cela pose plusieurs problèmes. Premièrement, on n'a pas séquencé tous les génomes de chat, et on n'aura jamais séquencé les génomes de tous les chats (puisque les chats du futur ne sont pas encore nés !). Il n'est donc pas évident de définir un ensemble de codes ADN qui correspondrait à des chats. En deuxième lieu, même si tel fut le cas, devoir séquencer le génome d'un animal pour déterminer s'il s'agit d'un chat est une solution tout à fait inadéquate en pratique. En troisième lieu, on peut sérieusement questionner le fait qu'une cellule isolée de chat, par exemple ses poils, contenant l'ADN d'un chat est vraiment un chat. Enfin, et surtout, la notion d'ADN était complètement absente il y a un demi-siècle. Au mieux, ceci veut dire que Schrödinger, Darwin et Aristote ne savaient pas de quoi ils parlaient quand ils parlaient de chat.

Il faut se rendre à l'évidence : il n'y a pas de définition rigoureuse satisfaisante du mot « chat ». Si je vous demandais ce qu'était un chat, vous ne pourriez pas fournir de définition universelle et indiscutable. Et ce n'est pas si surprenant.

¹⁵  *Sommes-nous humains ?* Dirty Biology | L. Grasset (2015)

Après tout, ce n'est pas ainsi que vous avez appris à reconnaître et utiliser le concept de chat ! Si vous savez à peu près ce qu'est un chat, c'est parce que vous avez vu des milliers, voire des millions d'images de chat, vous les avez entendu et vous avez lu à leur sujet. Vous avez appris ce qu'est un chat en observant un très grand nombre de données. Mais vous n'avez jamais eu une définition formelle de ce que c'était ! En fait, dans l'histoire de l'humanité, personne n'a jamais su (ou eu à) donner de définition formelle de ce qu'est un chat.

Mieux encore. Il y a bien dû y avoir un premier humain à penser le concept de chat. Comme cet humain fut le premier à y arriver, personne n'a pu le lui enseigner. Cet humain a alors inventé de lui-même le concept de chat. Pourquoi ? Et comment y est-il arrivé ? D'où viennent les nouveaux concepts qu'introduisent les humains ? Est-ce spécifique à l'intelligence humaine ?

Je trouve la découverte de l'intelligence artificielle de Google fascinante car elle nous donne la réponse à ces questions. Non, ce n'est pas une spécificité de l'intelligence humaine, puisque le réseau de neurones de Google y est arrivé. Et il y est arrivé, parce qu'il voulait analyser, synthétiser et expliquer les corrélations entre les couleurs des pixels d'images numérisées. Il voulait avoir un modèle de ce qu'il voit. Et le modèle qu'il a trouvé conduisait naturellement à la création du concept de chat !

Le naturalisme poétique

Voilà qui nous mène à la question la plus fascinante de ce chapitre. Le chat est un concept d'un modèle abstrait. Mais alors, est-ce qu'un chat *existe* ? La question peut paraître stupide. Vous avez sans doute envie d'hurler : *mais, bien sûr, puisqu'on en voit tous les jours* ! Pourtant, si l'on considère la théorie de la physique la plus crédible à ce jour — à savoir le modèle standard de la physique des particules — le monde n'est composé que de champs quantiques dont les excitations sont quantifiées et forment les électrons, quarks, photons et autres constituants physiques. Nulle part dans ce modèle, et nulle part dans la physique, trouve-t-on le concept de chat. Les théories physiques rejettent même l'existence d'objets qui ne sont pas des champs quantiques. En particulier, le modèle standard de la physique réfute l'existence des chats. Au mieux, il s'agit d'un tas d'électrons, de protons et de neutrons arrangés d'une certaine manière.

Dans son excellent livre *The Big Picture*, le physicien Sean Carroll se demande ainsi si, lorsqu'on voit un chat courir derrière une souris, il est scientifiquement correct de dire que le chat veut manger la souris. A-t-on le droit d'accepter l'existence du chat, de la souris et de l'intention que pourrait avoir un chat ?

Pour Sean Carroll, même si la physique théorique rejette la réalité de toutes ces notions, parler de chats, de souris et d'intentions reste néanmoins la *bonne* façon de parler de la situation qu'on vient de présenter. En effet, ce faisant, on considère un autre modèle de la réalité, certes en contradiction avec les notions

de la physique théorique, mais qui n'en reste pas moins compatible en termes prédictifs. En particulier, le phénomène dit d'*émergence*, qui fait le pont entre les descriptions plus fines mais trop complexes de la réalité et les descriptions plus grossières mais plus utiles, est en fait un phénomène bien connu au sein même de la physique, où des notions comme la température et la pression émergent du point de vue macroscopique — même si la physique des particules en rejette l'existence.

Sean Carroll défend ainsi une nouvelle position épistémologique qu'il nomme le *naturalisme poétique*. Il affirme en particulier que toute théorie de la réalité est une sorte de poésie, qui introduit ses propres concepts et permet ses propres prédictions. De tels concepts utiles réfèrent alors, selon Carroll, à une forme de réalité. Cette notion rejoint d'ailleurs le concept de réalisme modèle-dépendant de Stephen Hawking et Leonard Mlodinow, selon lequel toute théorie définit sa propre réalité. Le chat n'est alors pas réel selon la physique des particules. Cependant, le chat *existe* dans le modèle de la réalité qui nous est le plus familier — celui qui dit qu'un chat courant derrière une souris a très probablement envie de la manger.

Les positions de Carroll, Hawking et Mlodinow semblent au moins partiellement bayésiennes. La *pure bayésienne* se moque éperdument de l'existence de quoi que ce soit en dehors de ses modèles. Pour le démon de Solomonoff, il y a des données empiriques mesurées par des capteurs, et le savoir se résume à déterminer les modèles les plus crédibles, étant donné ces données. C'est en tout cas ainsi que fonctionne l'intelligence artificielle de Google. Les concepts abstraits, comme la notion de chat, sont alors des composants de modèles crédibles pour comprendre ces corrélations, ou des étapes intermédiaires de calculs nécessaires à la prédiction.

D'ailleurs, ces concepts abstraits, ce sont ni plus ni moins les facteurs de confusion (ou de concision) qui permettent d'expliquer des corrélations sans faire appel à des liens de cause à effet — et l'on verra plus tard qu'ils jouent un rôle clé dans de nombreux modèles de *machine learning*. Ce sont ces concepts abstraits qui expliquent pourquoi l'hôpital est meilleur que la clinique, et pourquoi il est raisonnable de penser avec grande crédence qu'un mouton noir d'un côté est aussi noir de l'autre.

Et au final, l'existence ou le réalisme de ces concepts importe alors peu. « Tous les modèles sont faux. » Ce qui importe, c'est que ces concepts soient utiles pour expliquer les données auxquelles la *pure bayésienne* est exposée, et pour l'aider à effectuer des prédictions.

Références en français

- ⌚ Interview d'André Kuhn : les sciences criminelles | Podcast Science (2011)
- ▶ Le paradoxe de Simpson | Science Étonnante | D. Louapre (2015)

- ▶ *Chocolat, corrélation et moustache de chat* | La statistique expliquée à mon chat | L. Maugeri, G. Grisi et N. Uyttendaele (2016)
- ▶ *Tu bois du light ? T'es foutu !* La statistique expliquée à mon chat | L. Maugeri, G. Grisi et N. Uyttendaele (2017)
- ▶ *Sommes-nous humains ?* Dirty Biology | L. Grasset (2015)
- ▶ *James Lind - L'essai clinique* | Risque Alpha | T. Le Magoarou (2017)

- ▶ *Satanés facteurs de confusion* | IA 17 | Science4All | L.N. Hoang (2018)
- ▶ *Solution du paradoxe de Simpson* | Science4All | L.N. Hoang (2018)

Références en anglais

- 📘 *All of Statistics: A Concise Course in Statistical Inference* | Springer Science & Business Media | L. Wasserman (2013)
- 📘 *The Big Picture: On the Origin of Life, Meaning and the Universe Itself* | Dutton | S. Carroll (2016)
- 📘 *The Grand Design* | Bantam Books | S. Hawking and L. Mlodinow (2010)

- 📖 *Inadmissibility of the usual estimator for the mean of a multivariate distribution* | Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability | C. Stein (1956)
- 📖 *Chocolate consumption, cognitive function, and Nobel laureates* | The New England Journal of Medicine | F. Messerli (2012)
- 📖 *Endogenous stratification in randomized experiments* | National Bureau of Economic Research | A. Abadie, M. Chingos and M. West (2013)
- 📖 *Need-based financial aid and college persistence: Experimental evidence from Wisconsin* | S. Goldrick-Rab, D. Harris, J. Benson and R. Kelchen (2012)
- 📖 *Research Student Aid Before You Reform* | Chronicle of Higher Education | A. Kelly (2012)
- 📖 *Building high-level features using large scale unsupervised learning* | Q.V. Le, M.A. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, A. Ng (2012)

- 🌐 *Google's Artificial Brain Learns to Find Cat Videos* | Wired | Liat Clark (2012)
- ▶ *Maths: Simpson's Paradox* | singing banana | J. Grime (2010)
- ▶ *Is Punishment or Reward More Effective?* Veritasium | D. Muller (2013)
- ▶ *How statistics can be misleading* | Ted-Ed | M. Lidell (2016)
- ▶ *Simpson's Paradox* | Minute Physics | H. Reich (2017)
- ▶ *Are University Admissions Biased? Simpson's Paradox Part 2* | Minute Physics | H. Reich (2017)

Une réponse approximative à la bonne question a beaucoup plus de valeur qu'une réponse précise à la mauvaise question.

John Tukey (1915-2000)

La vérité [...] est beaucoup trop compliquée pour permettre autre chose que des approximations.

John von Neumann (1903-1957)

14

Vite et (assez) bien

Le mystère des nombres premiers

Le 11 mars 2016, Robert Lemke Oliver et Kannan Soundararajan découvrirent expérimentalement que les derniers chiffres des nombres premiers ne se comportent pas de manière aléatoire. Les derniers chiffres ont tendance à ne pas se répéter. Le nombre premier succédant à un nombre premier qui finit par un 3, comme 23 ou 43, aura un dernier chiffre qui sera plus souvent un 7 qu'un 3 aussi. Ce fut une surprise monumentale pour la communauté mathématique. Mais pour les autres, le plus surprenant est peut-être le fait que cette découverte puisse être surprenante...

Les nombres premiers sont ceux dont les seuls diviseurs sont 1 et eux-mêmes. Les premiers nombres premiers sont 2, 3, 5, 7, 11, 13, 17, et ainsi de suite. Leur étude a fasciné des générations de mathématiciens depuis des millénaires. Il y a plus de 2 000 ans, Euclide a prouvé qu'il y avait une infinité d'entre eux. En 2002, Agrawal, Kayal et Saxena ont trouvé un algorithme polynomial pour déterminer si un nombre est premier. Et en 2012, Yitang Zhang a prouvé qu'une infinité de nombres premiers consécutifs sont séparés de 70 millions ou moins. Beaucoup a été découvert au sujet de ces briques élémentaires de la structure multiplicative des nombres entiers.

Cependant, de nombreuses questions élémentaires demeurent ouvertes. La conjecture de Goldbach postule que tout nombre pair est la somme de deux nombres premiers. La conjecture des nombres premiers jumeaux postule qu'il existe

une infinité de nombres premiers à distance 2, comme 3 et 5, 41 et 43, ou encore 137 et 139. L’hypothèse de Riemann, elle, postule que la distribution des nombres premiers peut se déduire de certaines propriétés mathématiques d’une mystérieuse fonction appelée fonction zêta de Riemann. On y reviendra.

L’un des problèmes ouverts les plus difficiles consiste à déterminer un algorithme rapide pour calculer le n -ième nombre premier. Un problème quasiment équivalent consiste à déterminer le nombre de nombres premiers inférieurs à n avec un algorithme rapide. Celui qui résoudra ces problèmes se couvrira de gloire ! Cependant, il n’est pas dit que ces problèmes puissent un jour être résolus. À ce jour et à ma connaissance, le meilleur algorithme est celui de Deléglise et Rivat, publié en 1996. Mais le temps de calcul de cet algorithme demeure exponentiel en le nombre de chiffres de n .

Voilà qui soulève une autre difficulté au problème de la prédiction. Le cas des nombres premiers n’est ni fondamentalement aléatoire, ni épistémiquement incertain, ni même chaotique. Si les nombres premiers pourraient néanmoins demeurer fondamentalement imprévisibles, c’est parce que la quantité de calcul nécessaire pour prédire le googolième¹ nombre premier pourrait être nécessairement trop grande, pour nos cerveaux et nos machines à calculer !

De la même manière, le théorème de Ramsey soulève des problèmes dont on connaît des méthodes de résolution, mais ces méthodes de résolutions requièrent toutes des calculs déraisonnables. Par exemple, considérons le problème qui consiste à déterminer le nombre minimal de sommets qu’un graphe complet doit avoir pour être sûr que, quelle que soit la coloration des arêtes en rouge ou en bleu, le graphe admette au moins un sous-graphe à monochrome à n sommet. Si vous n’avez rien compris, ce n’est pas grave. Les détails importent peu.

Toujours est-il que pour $n = 3$, on sait que la réponse est 6. La preuve est facile. Pour $n = 4$, on sait que la réponse est 14. Mais « ce n’est plus si simple », rajoute le mathématicien Paul Erdős. Que se passe-t-il pour $n = 5$? « Personne ne sait. Quelque chose entre² 41 et 55. »

« Supposons qu’un être maléfique dise à l’humanité : “donnez-moi la réponse pour $[n =]5$ ou j’exterminate la race humaine.” J’aime plaisanter et dire que le mieux dans ce cas serait de calculer la réponse, à l’aide des mathématiques et des ordinateurs », rajoute Erdős. « S’il nous demande $[n =]6$ [...], la meilleure chose à faire, c’est de le détruire avant qu’il nous détruise, parce qu’on ne pourra pas résoudre le cas $[n =]6$. » Certains problèmes sont impossibles, non pas parce qu’on ne sait pas comment les attaquer, mais parce que la puissance de calcul requise pour les résoudre excède de loin ce que la physique nous permet de calculer.

¹Un googol est égal à 10^{100} .

²En 2018, on sait maintenant que la réponse est entre 43 et 48.

Le théorème des nombres premiers

À défaut de court-circuiter les calculs nécessaires, les chercheurs en théorie des nombres se sont alors naturellement tournés vers des calculs approchés. À commencer par Carl Friedrich Gauss. Autour de 1800, Gauss s'est amusé à étudier les écarts entre nombres premiers successifs. Entre 3 et 5, l'écart est de 2. Entre 7 et 11, il est de 4. Gauss calcula qu'en moyenne, les nombres premiers successifs inférieurs à 100 étaient espacés de 4. Ceux inférieurs à 1 000 étaient, en moyenne, espacés de 6 ; ceux inférieurs à 10 000 de 8,1 ; ceux inférieurs à 100 000 de 10,4. Autrement dit, à chaque fois que l'on décuplait les nombres premiers considérés, on ajoutait par là même un peu plus de 2 à l'écart moyen entre nombres premiers successifs (2,3 pour être un peu plus précis).

Ça ne vous rappelle rien ? Il s'agit là d'une transformation de la multiplication (par 10) en une addition (par environ 2,3). L'écart moyen entre nombres premiers successifs semble être une fonction logarithmique de ces nombres premiers. Gauss en vint à l'intuition suivante : le nombre de nombres premiers inférieurs à n , qu'il nota $\pi(n)$, semblait pouvoir être bien approché par $n / \ln n$, où $\ln n$ est le logarithme dit *népérien*³. Le logarithme népérien, c'est celui dont la base est la constante d'Euler $e \approx 2,718$.

Formellement, Gauss conjectura le fait que l'erreur relative de l'approximation de $\pi(n)$ par $n / \ln n$ s'annule quand n tend vers l'infini ! On dit qu'à l'infini $\pi(n)$ est équivalent à $n / \ln n$. Cette conjecture est devenue le théorème des nombres premiers en 1896, lorsque, de façon indépendante, les mathématiciens Jacques Hadamard et Charles Jean de la Vallée Poussin parvinrent à fournir une preuve de cette remarquable description approchée de la répartition exacte des nombres premiers. Même s'il ne s'agit que d'une approximation qui ne nous donne pas la localisation exacte des nombres premiers, ce théorème remarquable est devenu l'un des monuments de la théorie des nombres !

En 1854, l'étudiant de Gauss, le brillant Bernhard Riemann, est allé plus loin encore dans l'approximation de $\pi(n)$. Il réussit à déterminer une formule exacte pour $\pi(n)$ à l'aide d'une autre fonction tout aussi mystérieuse appelée fonction zéta et notée ζ . En particulier, certains nombres, appelés zéros de la fonction zéta, permettent de calculer exactement $\pi(n)$.

Bien entendu, il y a un hic. Il y en a même deux. Le premier est que l'on ne sait pas exactement où sont ces zéros (et la fameuse hypothèse de Riemann⁴ postule justement que les zéros non triviaux sont en fait alignés sur une droite verticale du plan complexe). L'autre problème est qu'il y en a une infinité : le calcul exact de Riemann requiert un calcul infini.

³On peut ainsi constater que $\ln(10) \approx 2,3$, conformément à l'observation de Gauss.

⁴  Deux (deux ?) minutes pour l'hypothèse de Riemann | El Jj | J. Cottanceau (2016)

Les approximations de τ

Ceci étant dit, les mathématiques sont en fait remplies de calculs infinis. Le plus célèbre de ces calculs est sans doute celui de la fameuse constante du cercle τ (qui est reliée à son « imposture » historique⁵ $\pi = \tau/2$). Au XIV^e siècle, le génie indien Madhava découvrit la stupéfiante égalité $\tau = 8 - 8/3 + 8/5 - 8/7 + \dots$. Autrement dit, la constante du cercle τ n'est autre que 8 fois la somme alternée des inverses des entiers impairs !

On pourrait croire que ces calculs infinis ne servent à rien. En fait, on les retrouve omniprésents dans les modèles préliminaires des mathématiques appliquées, par exemple dans les équations de la dynamique des fluides (une dérivée ou une intégrale est un calcul infini). En effet, si ces calculs infinis représentent un idéal incalculable, ils suggèrent en général une excellente façon d'effectuer des calculs approchés. L'équation de Madhava permet ainsi de déterminer des approximations de τ , et donc de la circonférence de cercles dont on connaît le rayon. Et ce sont d'ailleurs de telles approximations que les ingénieurs seront amenés à utiliser en pratique.

Si vous demandez à votre calculatrice préférée ou à Google de vous donner la valeur de τ , il y a de bonnes chances qu'elle vous mente et n'en donne qu'une approximation à 13 décimales. Ceci ne se restreint d'ailleurs pas à τ . Votre calculatrice gère très mal tous les nombres dont l'écriture binaire n'est pas finie⁶, que ce soient des constantes irrationnelles comme e ou $\sqrt{2}$ ou des nombres rationnels comme $1/3$ ou $0,2$. En particulier, parce qu'elle ne travaille qu'avec des nombres approchés, votre calculatrice peut trouver des résultats aberrants, comme $1/3 \cdot 3 \neq 1$, ou $(x+y)-x=0$, même si y est strictement positif, et dès lors que x est beaucoup plus grand que y .

Il arrive souvent que des mathématiciens considèrent alors que les calculs informatiques sont des approximations de la théorie mathématique. La position du démon de Solomonoff est cependant l'inverse. Pour le démon de Solomonoff, les nombres réels, par exemple, sont des modèles qui permettent de structurer et mieux penser les algorithmes. En particulier, pour un informaticien qui chercherait à suivre les pas de Solomonoff, le modèle dont il cherche à mesurer les crédences n'est pas celui dont les paramètres sont des nombres réels, mais celui qui sera implémenté dans un fichier informatique avec une troncature des nombres réels du modèle mathématique. C'est ainsi que, contrairement à son idéalisation mathématique, la taille d'un réseau de neurones artificiel se compte en octets⁷.

⁵ π est une fraude | Science4All | L.N. Hoang (2017)

⁶ Why Computers are Bad at Algebra | Infinite Series | K. Houston Edwards (2017)

⁷ D'ailleurs les propriétés des réseaux de neurones dépendent fortement de cette troncature du modèle mathématique. Ainsi, un réseau de neurones « mathématique » a une dimension VC d'au moins $\Omega(|\text{aretes}|^2)$, tandis que toute troncature finie de ses poids réels lui confère une dimension VC de seulement $O(|\text{aretes}|)$.

Les développements limités

Si approximer des nombres comme τ avec grande précision n'a finalement qu'un intérêt limité, approximer des courbes, des fonctions et des comportements physiques, biologiques ou mathématiques peut avoir un très grand nombre d'applications. C'est le cas de l'approximation de la fonction qui compte les nombres premiers via le théorème des nombres premiers.

Dans un cadre général, il existe un outil qui, à partir de tout modèle, calcule une version approchée beaucoup plus simple de ce modèle. Cet outil est le développement limité. Typiquement, un développement limité va se permettre d'approcher un petit morceau de cercle par une droite, ou une petite région à la surface (pourtant arrondie) de la Terre par un plan. En termes algébriques, ceci revient à remplacer des équations dites *non-linéaires* par des équations affines de la forme $y = ax + b$. Pour des phénomènes aux variations raisonnablement faibles, ces approximations seront parfaitement acceptables. Voilà qui explique pourquoi, à l'école, on passe autant de temps à étudier ces équations simplistes, et pourquoi on les retrouve à travers toutes les sciences⁸.

L'exemple le plus spectaculaire de développement limité indispensable aux physiciens est peut-être celui des équations de la relativité générale d'Einstein. Ce développement limité conduit à la physique newtonienne ! En d'autres termes, les lois de Newton, notamment les lois de la gravité, ne sont autres qu'une approximation de celle d'Einstein, laquelle est parfaitement acceptable dans un contexte limité. Ce contexte est celui des cas de « faible » gravité⁹.

Il arrive que certains défenseurs des sciences persistent à affirmer que les lois de Newton sont « vraies », dans leur champ d'applicabilité. Notre *pure bayésienne* n'est toutefois pas d'accord. « Tous les modèles sont faux. » Ou dit en termes bayésiens, il ne faut jamais mettre toutes ses crédences sur un seul modèle.

Mais ce n'est pas tout. La *pure bayésienne* n'attache en fait presque aucune crédence aux lois de Newton, puisque ces lois expliquent strictement moins de phénomènes qu'une combinaison adéquate (même bancale) de la relativité générale et la mécanique quantique, sans être nettement plus succincte à décrire.

⁸Formellement, ceci correspond à l'approximation de Taylor-Lagrange. Considérons par exemple une fonction $f : \mathbb{R} \rightarrow \mathbb{R}$ infiniment dérivable. La « linéarisation » de f au point x_0 correspond à l'approximation de f autour de x_0 par

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0).$$

On peut obtenir de meilleures approximations avec des termes d'ordre supérieurs comme suit :

$$f(x) \approx \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k.$$

Le théorème de Taylor-Lagrange quantifie la petitesse des erreurs des approximations.

⁹Plus précisément, il s'agit des cas à faibles courbures de l'espace-temps.

Les contraintes du pragmatisme

Le problème de la *pure bayésienne*, c'est qu'elle n'est pas pragmatique. Souvenez-vous, dans sa forme la plus pure qui est le démon de Solomonoff, elle viole les lois de la physique ! En particulier, la *pure bayésienne* n'est pas limitée par des temps de calculs pourtant inévitables, et peut instantanément résoudre les équations de la théorie quantique des champs et de la courbure de l'espace-temps, ou déterminer le googolème nombre premier.

En pratique toutefois, les puissances de calculs sont limitées. On l'a vu. Il est illusoire d'espérer effectuer plus de 10^{70} calculs dans notre système solaire. Voilà qui réduit drastiquement nos capacités à calculer la formule de Bayes. Si elle veut étudier toutes les théories descriptibles en 1 000 caractères (soit deux ou trois pages), la *pure bayésienne* devra calculer des termes de vraisemblance $\mathbb{P}[D|T]$ pour chacune de ces théories ! Or le nombre de ces théories est énorme. En supposant qu'on n'utilise que les 26 lettres de l'alphabet, cela correspond à un corpus de 26^{1000} théories. Du coup, la *pure bayésienne* devra effectuer au moins 26^{1000} calculs. C'est physiquement largement impossible.

Or, souvenez-vous, il ne s'agit là que du cas où la *pure bayésienne* se restreint aux théories de 1 000 caractères. Le cerveau humain, à titre de comparaison, contient environ 10^{15} synapses, ce qui signifie que décrire exactement le cerveau nécessiterait environ 10^{15} bits d'information. Étudier toutes les théories avec autant de caractères représenterait au moins $2^{10^{15}}$ calculs ! Appliquer la formule de Bayes à ces théories est donc complètement illusoire, même avec des googols et des googols d'univers.

Les *learning machines* de Turing

En 1950, Alan Turing transpose avec merveille ce raisonnement sur la complexité inévitable des algorithmes à l'intelligence artificielle. Dans un incroyable article intitulé *Computing Machinery and Intelligence* publié dans la revue *Mind, a Quarterly Review of Psychology and Philosophy*, Turing se pose d'abord une question très distante de celle de la longueur nécessaire des algorithmes. « Les machines peuvent-elles penser ? » Telle est la question à laquelle Turing cherche à répondre. Cependant, l'ambiguïté du mot « penser » le pousse à préciser le problème. Au lieu de cela, Turing se demande si une machine peut agir comme un homme.

Plus précisément, Turing propose le test suivant : demandez à un humain A d'échanger des messages écrits avec deux autres entités X et Y. Parmi ces deux entités X et Y, il y a un autre humain B et une machine. La machine aura alors passé le test si l'humain A est incapable de déterminer laquelle des entités X et Y est l'humain B, et laquelle est la machine. Autrement dit, la machine aura réussi le test, si elle aura su tellement bien imiter un humain, qu'aucun

humain ne pourra déceler son inhumanité. C'est ce que Turing a appelé le jeu de l'imitation, et que l'on appelle aujourd'hui le *test de Turing*¹⁰.

Dans les sections 3 à 5 de son article, Turing reprend ses travaux de 1936 et définit rigoureusement ce qu'est une machine — ce qui aura conduit à l'invention de l'ordinateur, puis à l'avènement de l'ère numérique ! Puis, dans la section 6, Turing réfute 9 arguments classiques qui cherchent à démontrer que les machines sont incapables de penser. Mais surtout, dans la section 7, Turing anticipe la difficulté de son test et sa résolution. Alors que les ordinateurs n'existent pas encore vraiment, Turing prévoit d'ores-et-déjà non seulement leur future existence, mais aussi le fait que leurs performances seront largement suffisantes pour gagner le jeu de l'imitation. « Cela semble improbable que [les progrès d'ingénierie] ne seront pas adéquats pour [passer le test]. » Pour Turing, « le problème est surtout un problème de programmation. »

En particulier, avec une clairvoyance bluffante, Turing anticipe le fait que le code d'un programme qui réussit le test de Turing ferait nécessairement autour de 10^9 caractères. Autrement dit, en réutilisant la terminologie que l'on a introduite au chapitre 7, Turing parie que le test de Turing a une complexité de Solomonoff qui se compte en giga-octets. Pour en arriver à cette estimation, Turing s'appuie sur la seule machine qu'il connaissait et qui était capable de passer le test de Turing. Je parle du cerveau humain ! Et oui, qui de mieux qu'un homme pour imiter un homme ? Fort heureusement pour Turing, la neuroscience avait déjà fourni une estimation de la complexité du cerveau humain, en avançant un chiffre entre 10^{11} et 10^{15} synapses entre nos neurones — les estimations récentes varient entre 10^{14} et $5 \cdot 10^{14}$.

Turing postula qu'une petite fraction de ces synapses était absolument indispensable à qui chercherait à passer le test de Turing, d'où le chiffre de 10^9 caractères. Turing rajoute : « au rythme actuel de travail, j'écris environ mille caractères de programme par jour, si bien qu'avec soixante travailleurs, en travaillant ainsi pendant 50 ans, on pourrait accomplir la tâche [de programmer un algorithme qui passera le test de Turing], en supposant que rien ne finit à la poubelle à papier. Des méthodes plus expéditives sont désirables. »

Cet éclair de génie de Turing ne s'applique d'ailleurs pas qu'au test de Turing. Il y a fort à parier que de nombreuses tâches requièrent elles aussi des programmes très longs pour être résolues, à commencer par la maîtrise du langage naturel, la possession d'un « bon sens » ou l'art de l'empathie. Pire encore, certaines tâches, notamment en biologie et en sciences sociales, pourraient nécessiter des algorithmes dont la longueur excède la taille du cerveau humain. Dès lors, non seulement des algorithmes que nous écrivons échoueraient, mais nos cerveaux aussi seraient nécessairement incapables de résoudre ces tâches.

Prédire la prochaine crise financière pourrait être au-delà des capacités cognitives de nos cerveaux limités. « Si les gens ne croient pas que les mathématiques sont simples, c'est parce qu'ils ne se rendent pas compte à quel point la

¹⁰  *Le test de Turing* | IA 3 | Science4All | A. Gelaude et L.N. Hoang (2017)

vie est compliquée », disait John von Neumann. Cette affirmation peut ainsi être rendue rigoureuse, en mesurant la complexité des disciplines à l'aide de la complexité de Solomonoff¹¹ — ou mieux encore, la sophistication de Solomonoff qu'on introduira au chapitre 18 !

Revenons-en à Turing. Toujours avec son génie inégalable, Turing fait remarquer que le cerveau humain est capable de passer le test de Turing. Il postule alors que ce sera davantage en imitant la manière dont le cerveau humain est parvenu à passer le test de Turing qu'on permettra aux machines d'y arriver à leur tour. Or, Turing constate que l'éducation de l'enfant est une part cruciale de la manière dont son cerveau a fini par devenir ce qu'il est devenu. « Le cerveau d'un enfant est probablement un peu comme un carnet de note que l'on achète chez le papetier », écrit Turing. « Peu de mécanisme, et beaucoup de pages blanches¹². » Turing propose alors d'aider la machine à remplir ses propres pages blanches en lui permettant d'apprendre de données. Ainsi naquit le concept de *learning machines*, c'est-à-dire de machines capables d'apprendre.

L'idée des *learning machines* est donc de permettre à une machine d'écrire elle-même un programme de plusieurs milliards de caractères — ou plus si nécessaire. En termes plus algorithmiques, ceci revient à dire que le *machine learning* permet enfin l'étude et l'exploration de l'ensemble des algorithmes dont la description excède le giga-octet. Et de façon cruciale, ce qui guide cette exploration n'est pas les doigts d'un programmeur ; ce sont des données brutes, comme celles dont jouissent les enfants aussi.

En particulier, le raisonnement de Turing vient contredire ce que de nombreux intellectuels, y compris certains experts, répètent pourtant si souvent. « Le *machine learning* marche bien — les mathématiciens ne savent pas pourquoi », titrait Wired en 2015. Pourtant, dès 1950, Alan Turing avait prédit le succès à venir du *machine learning*, en précisant même que son émergence se ferait à la fin du XX^e siècle ! Mieux encore, Turing avait souligné la raison précise pour laquelle le *machine learning* surpasserait les programmes écrits par l'homme pour de nombreuses tâches : seul le *machine learning* est capable d'explorer l'espace des algorithmes dont la longueur excède le giga-octet¹³.

De même, certains experts reprochent souvent aux gros réseaux de neurones d'être impossibles à interpréter. Il n'est pourtant pas étonnant qu'un gros réseau de neurones performant ne soit pas descriptible en peu de caractères. En tout cas de manière exacte. En effet, s'il l'était, cette description en peu de caractères serait un algorithme court et capable de générer un autre algorithme (le réseau de neurones), lequel résoudrait des problèmes comme le test de Turing. Du coup, l'algorithme court aurait résolu le test de Turing. Or, on a justement postulé que ce test ne peut pas être résolu par des algorithmes courts.

¹¹Malheureusement, le problème de l'arrêt a pour conséquence l'incalculabilité de la complexité de Solomonoff.

¹²On verra au chapitre 19 que ceci est aujourd'hui rejeté par les neurosciences, en vertu de principes bayésiens !

¹³  *Les learning machines de Turing* | IA 7 | Science4All | L.N. Hoang (2018)

Reste à spécifier comment explorer l'espace des longs algorithmes. Alan Turing n'a cependant pas spécifié *quelle méthode* utiliser pour ce faire. Il a simplement affirmé que la capacité à apprendre allait être clé. Au début du XXI^e siècle, de nombreuses approches ont alors été proposées. On a déjà parlé de plusieurs d'entre elles. En vrac, on peut citer la régression linéaire, la régression logistique, les arbres de décision, les forêts de décision, les *support vector machines*, les réseaux de neurones, les réseaux bayésiens ou encore les champs de Markov. On reviendra longuement sur ces trois derniers algorithmes de *machine learning* dans ce chapitre et les suivants.

Pour l'heure, je veux rappeler ce que propose la *pure bayésienne*. Quitte à se restreindre aux algorithmes dont la longueur est au plus 10^9 caractères, la *pure bayésienne* va vouloir tous les tester et tous les comparer. Voilà qui nécessitera une quantité de calculs (largement) supérieur à 2^{10^9} , ce qui demeure complètement illusoire, même avec des googols d'univers en un temps très supérieur à l'âge de ces univers. En particulier, si Turing a vu juste concernant la complexité de Solomonoff de son test, alors l'approche purement bayésienne pour résoudre le test de Turing est vouée à l'échec en pratique.

Le bayésianisme pragmatique

Pour résoudre des problèmes à forte complexité de Solomonoff comme le test de Turing, le *bayésien pragmatique* est contraint de renoncer à la formule de Bayes. Il doit lui préférer des méthodes qui ne nécessitent pas des temps de calculs irréalistes. Voilà qui nous amène à modifier la notion d'utilité. Pour la *pure bayésienne*, étaient utiles les théories à grandes crédences. Pour le *bayésien pragmatique*, les théories à grandes crédences, mais qui requièrent des temps de calculs déraisonnables, sont en fait inutiles. L'attention du *bayésien pragmatique* se tournera alors vers les théories les plus crédibles, parmi l'ensemble des théories à faibles temps de calculs.

Voilà qui nous permet de justifier nos crédences pragmatiques en l'approximation donnée par le théorème des nombres premiers et en les lois de Newton. Pour la *pure bayésienne*, ces deux descriptions sont complètement inutiles, puisqu'elles sont toutes deux largement dominées par le calcul exact des nombres premiers d'un côté, et la relativité générale d'Einstein de l'autre. Cependant, le *bayésien pragmatique* va embrasser avec joie ces deux approximations, car leurs besoins en calcul sont très nettement inférieurs. Calculer une bonne approximation $n/\ln n$ se fait en temps logarithmique en n , alors que les calculs vectoriels et différentiels de la théorie de Newton sont grandement plus rapides que ceux des tenseurs de la théorie d'Einstein. Pourvu que l'on se place dans un contexte où ces deux approximations sont suffisamment acceptables (grandes valeurs de n et faible gravité), ces deux approximations seront, pour le *bayésien pragmatique*, beaucoup plus *utiles* que leurs homologues exacts !

Cette notion d'utilité relative aux temps de calculs pourrait également être l'explication première du succès stupéfiant des réseaux de neurones. En effet, contrairement à des algorithmes plus sophistiqués, les réseaux de neurones, ou du moins ceux dits *feedforward*, ont des temps de calculs nécessairement limités et faibles. En fait, si l'on considère une définition suffisamment large des réseaux de neurones *feedforward*, alors l'ensemble de ces réseaux est exactement l'ensemble des algorithmes rapides, une fois ces algorithmes parallélisés. Effectuer du *machine learning* sur des réseaux de neurones *feedforward*, c'est donc chercher à expliquer au mieux les données (avec un *a priori* bayésien adéquat si possible) à l'aide d'algorithmes rapides. C'est donc un premier pas vers le *bayesianisme pragmatique*.

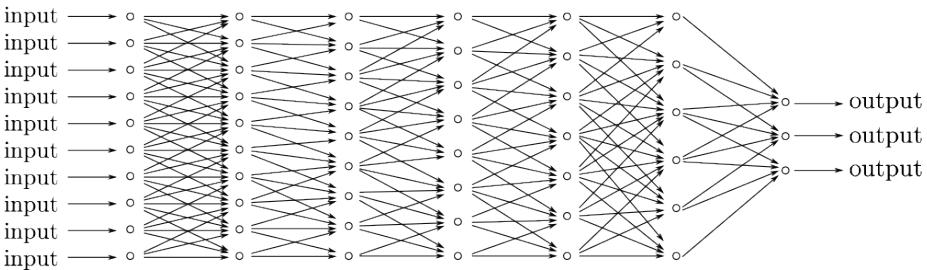


Figure 14.1. Un réseau de neurones est une suite de communications entre neurones et de calculs élémentaires par les neurones. Il est dit *feedforward* si la communication est acyclique, c'est-à-dire, intuitivement, si elle ne va que dans un seul sens et ne boucle jamais.

Les réseaux de neurones ne sont pas les seuls à être rapides. De même, calculer la prédiction d'une régression linéaire est également un calcul rapide. Mais sans doute trop. Dans un article que j'ai co-écrit avec Rachid Guerraoui¹⁴, on affirme même que la faiblesse de nombreuses approches de *machine learning* est de ne se restreindre qu'à l'étude d'algorithmes dont le temps de calcul parallélisé n'est composé que d'une poignée d'étapes. Le problème, c'est qu'on ignore alors tous les algorithmes plus lents. Or, pour des raisons sur lesquelles on reviendra au chapitre 18, il y a fort à parier que nombreux d'algorithmes pertinents pour analyser les données brutes ou pour résoudre les problèmes qui nous importent sont en fait nécessairement des algorithmes relativement lents. Cette remarque nous aura conduit à une justification théorique du succès du *deep learning*. On y reviendra.

¹⁴ Deep Learning Works in Practice. But Does it Work in Theory? R. Guerraoui and L.N. Hoang (2018)

Les algorithmes sous-linéaires

Jusque-là, je me suis beaucoup intéressé aux propriétés des algorithmes candidats pour gagner nos crédences. J'ai aussi affirmé que l'accumulation d'une grande quantité de données était indispensable pour nous guider dans cette quête aux meilleurs algorithmes prédictifs.

Cependant, toutes les données ne se valent pas. En particulier, à l'instar du flux continu d'une caméra de surveillance installée au fin fond d'un désert, il arrive que la quasi-totalité des données collectées soient sans intérêt. Le problème, c'est qu'à l'heure du *Big Data*, simplement lire ces données sans intérêt pour vérifier qu'elles sont en effet sans intérêt peut prendre un temps déraisonnable — c'est ce que je me dis pour me donner bonne conscience quand je ne lis pas tous les messages qui me sont envoyés !

Pour résoudre ce problème, les informaticiens théoriciens ont tourné leur attention vers les algorithmes dits *sous-linéaires*. Ces algorithmes ont la particularité d'extraire la substantifique moelle de jeux de données sans prendre le temps de lire toutes les données. Autrement dit, ces algorithmes parviennent à lire « très en diagonale » les données auxquelles ils s'appliquent.

L'archétype historique d'un tel exemple est Google. Une recherche sur Google se doit de fournir des résultats de manière quasi-instantanée. Or, la base de données de Google contient une grosse portion des millions de milliards de pages du web. Il est hors de question pour Google d'explorer toute sa base de données avant de retourner des réponses pour l'utilisateur. Ceci prendrait des jours ! L'algorithme de recherche Google se doit d'être sous-linéaire.

L'astuce de Google, à l'instar de l'astuce des bibliothèques et des dictionnaires, est de trier et d'arranger au préalable leur base de données, de sorte qu'y faire des requêtes puisse se faire rapidement. Par exemple, en organisant les mots par ordre alphabétique, les dictionnaires permettent aux utilisateurs de rapidement savoir vers où chercher un mot dont ils cherchent la définition. Mieux encore, grâce à l'ordre alphabétique, étant donné une page et un mot à chercher, l'utilisateur sera capable de savoir si le mot à chercher précède la page en question, ou s'il apparaîtra plus tard dans le dictionnaire. Plus généralement, rechercher une donnée dans une base de données triée (par un ordre total) peut se faire très rapidement. L'algorithme dit *dichotomique* (qui est sans doute celui que vous utiliseriez pour chercher un mot dans un dictionnaire) a un temps de calcul logarithmique en la taille de la base de données.

Cependant, les algorithmes de Google, des bibliothécaires et des dictionnaires requièrent de gros calculs en amont. Organiser et ordonner tous les sites web ne peut pas se faire sans les avoir tous visités. Peut-on néanmoins extraire de l'information utile de jeux de données sans tout explorer, et sans organiser ces jeux de données au préalable ? De façon surprenante, pour certaines sortes d'informations utiles et de données, la réponse est oui. C'est le cas notamment de la transformée de Fourier.

La transformée de Fourier, étudiée au début du XIX^e siècle entre autres par Joseph Fourier, est une façon de changer la description de signaux sonores, d'images¹⁵, de vidéos ou autres cours de la bourse. Arrêtons-nous sur le son.

Le son peut ainsi être décrit par les variations de pression de l'air au niveau de vos tympans. Mais de façon équivalente, et certainement plus simple notamment pour écrire la musique, on peut décrire un son via les fréquences qui le composent. La transformée de Fourier est alors une sorte de dictionnaire bilingue, qui traduit la description du son par ses variations en volume, en une description par ses fréquences. Or, de telles traductions ont un très grand nombre d'applications, notamment parce qu'il est souvent plus pertinent de modifier les fréquences d'un son pour l'améliorer que de modifier ses variations de volume, mais aussi parce qu'un son est souvent composé uniquement d'une poignée de fréquences.

En fait, la transformée de Fourier a envahi notre quotidien, si bien que, selon le professeur Richard Baraniuk, nos ordinateurs et téléphones calculent des milliards de transformées de Fourier par jour. À chaque fois qu'on écoute de la musique, qu'on contemple une image digitale, ou qu'on regarde une vidéo, on fait faire des transformées de Fourier à nos machines.

Dès lors, toute accélération potentielle des algorithmes de calcul de transformées de Fourier représente des milliards d'euros en termes de calculs informatiques ou d'électricité requise pour ces calculs, sans compter le temps d'attente des utilisateur. En 1964, James Cooley et John Tukey réussirent le tour de force d'accélérer significativement le temps de calcul de la transformée de Fourier (discrète). Alors que l'approche naïve requiert un temps quadratique en la taille des données, l'algorithme de Cooley et Tuckey, appelé *fast Fourier transform*, tourne en temps quasi-linéaire¹⁶. Autrement dit, pour calculer la transformée de Fourier d'un mégaoctet de données, la *fast Fourier transform* ne requiert que quelques millions d'opérations (soit quelques millisecondes pour un ordinateur moderne), par opposition aux millions de millions d'opérations (près d'une minute de calcul) pour l'approche naïve.

Cependant, à l'âge du *Big Data*, la *fast Fourier transform* est encore trop lente, notamment si elle doit traiter des gigaoctets, voire teraoctets de données. Peut-on rendre la transformée de Fourier encore plus rapide ? En 2012, Hassanieh, Indyk, Katabi et Price¹⁷ découvrirent que la réponse est oui. Ou plutôt, ils proposèrent un algorithme astucieux qui calcule k fréquences approximativement principales d'un signal, en un temps de l'ordre de k fois le logarithme de la taille des données¹⁸. En particulier, cet algorithme remarquable, appelé *sparse Fourier transform*, tourne en un temps sous-linéaire, ce qui signifie qu'il arrive à connaître le signal à traiter en l'ignorant presque entièrement.

¹⁵  Deux (deux ?) minutes pour l'éléphant de Fermi & Neumann | El Jj | J. Cottanceau (2018)

¹⁶ La complexité de la transformée de Fourier discrète est alors passé de $O(n^2)$ à $O(n \log n)$.

¹⁷  Faster than the Fast Fourier Transform | ZettaBytes | M. Kapralov (2017)

¹⁸ La complexité est en fait en $O(k \log^2 n)$.

Plusieurs modes de réflexion

S'ils ne sont pas encore la norme en informatique, les algorithmes sous-linéaires semblent être ceux que nos cerveaux préfèrent. Qui n'a jamais parcouru en diagonales des documents écrits, écouté d'une oreille le fil d'une discussion ou mangé sans délester le goût de sa nourriture ? Étrangement, il arrive parfois que quelque chose dans le texte, la discussion ou la nourriture attire notre attention. Si tel est le cas, on semble basculer de mode de réflexion. On se met à utiliser des algorithmes plus lents et plus précis pour décortiquer la signification du texte, la subtilité de la discussion ou la saveur particulière d'une spécialité locale.

Ce que je décris là est le cœur de l'excellent livre *Thinking Fast and Slow* du prix Nobel d'économie Daniel Kahneman. Kahneman y distingue deux systèmes de pensées : le système 1 et le système 2. Le système 1 est telle une *sparse Fourier transform*. Il est rapide, efficace, affairé et probablement très erroné. Le système 2 est telle une transformée de Fourier exacte. Il est lent, gourmand en énergie, paresseux et plus correct.

Selon Kahneman, on a tendance à s'identifier à notre système 2, et on a tendance à être inconscient de notre système 1. Pourtant, la quasi-totalité du temps, c'est bien le système 1 qui est au commande. Et cela nous conduit à souvent nous tromper. Testons cela : une batte et une balle coûtent ensemble 1,10 \$. La batte coûte 1 \$ de plus que la balle. Combien coûte la balle ?

Il y a de bonnes chances qu'une réponse vous saute aux yeux ; mais que cette réponse soit fausse. Si c'est le cas, selon Kahneman, c'est parce que votre système 1 s'est précipité et a donné la première réponse qui lui venait à l'esprit ; ce n'est qu'ensuite, peut-être après m'avoir lu, que votre système 2 a commencé à questionner le système 1.

On pourrait croire que Kahneman souhaite nous convaincre d'abandonner le système 1, et de faire travailler aussi souvent que possible le système 2. Ce n'est pas entièrement le cas. Après tout, même s'il est plus juste, le système 2 est aussi plus fatigant pour le cerveau. Réfléchir a un coût. Or, comme tout acteur, pianiste ou surfeur le sait, on ne peut pas se permettre de penser tous nos faits et gestes. Il faut que les gestes finissent par être exécutés de manière naturelle et spontanée, sans effort intellectuel conscient majeur.

Pour arriver à effectuer des gestes compétents sans avoir à y réfléchir longuement, acteurs, pianistes et surfeurs utilisent tous la même astuce : leur système 2 forcent leur système 1 à apprendre le geste en question. Et ceci est sans doute la remarque la plus importante à retenir pour qui veut apprendre ou enseigner. L'apprentissage, ce n'est pas tant faire rentrer des informations dans notre cerveau ; c'est davantage amener le système 2 à éduquer le système 1, de sorte que ce système 1 découvre une heuristique qui résout rapidement le problème que le système 2 sait déjà résoudre, mais avec beaucoup de temps et d'énergie. Autrement dit, *apprendre, c'est surtout découvrir des heuristiques capables de faire vite et (assez) bien.*

Devenez post-rigoureux !

Il semble que ce soit aussi ainsi que fonctionne l'apprentissage des mathématiques. Aujourd'hui, si l'on me pose des questions sur l'addition, l'exponentielle ou la machine de Turing, il n'est pas improbable que j'arrive à répondre à ces questions sans faire appel au système 2. C'est parce que, au fil des années, mon système 2 a fini par réussir à enseigner à mon système 1 comment résoudre ces problèmes sans effort intellectuel ! Je dispose dans mon cerveau d'un très grand nombre d'heuristiques pour résoudre un très grand nombre de problèmes de mathématiques ; des problèmes qu'un jeune étudiant pourrait trouver difficiles. Certains appellent cela des facultés mathématiques. D'autres appellent cela une intuition mathématique. Je préfère appeler cela des heuristiques performantes découvertes grâce au dur labeur du système 2.

Mais ce n'est pas là l'aspect le plus important de l'entraînement mathématique du système 1. Sur son blog¹⁹, le mathématicien et médaille Fields Terence Tao distingue trois phases dans l'apprentissage mathématique, qu'il appelle pré-rigoureuse, rigoureuse et post-rigoureuse. En bref, les étudiants commencent d'abord par jouer avec les nombres et les concepts mathématiques élémentaires sans se soucier de la validité de leurs manipulations algébriques. Puis, avec le temps et grâce à l'éducation mathématique, vient le moment de la rigueur, qui vire rapidement au purisme. Tout doit alors être justifié dans un langage formel. Enfin, la dernière phase post-rigoureuse est celle des chercheurs, qui passent le plus clair de leur temps à combiner des arguments heuristiques pour avoir les grandes lignes des preuves de leurs théorèmes, et se détachent alors du formalisme appris plus tôt. De façon cruciale, la seconde phase semble être un passage obligé pour accéder à la dernière.

Le raisonnement de Tao peut se comprendre en termes de systèmes 1 et 2. La phase pré-rigoureuse correspond à un cas où ni le système 1, ni le système 2 n'ont appris la rigueur. La phase rigoureuse est amorcée au moment où le système 2 découvre la rigueur et son importance. Pendant cette phase, le système 2 va alors éduquer le système 1 pour lui faire prendre conscience des limites de ses intuitions. Mais surtout, le système 1 va alors apprendre à mesurer la crédence qu'il a en lui-même, pour mieux déterminer s'il peut se débrouiller seul ou s'il lui faut faire appel au système 2. Ce n'est qu'après ce dur apprentissage que le système 1 sera un parfait allié du système 2, et ne le dérangera qu'à des moments nécessaires. En particulier, il semblerait que devenir un bon mathématicien consiste essentiellement à disposer d'un système 1 compétent — mais que la phase rigoureuse est une étape obligée pour atteindre cet état.

Ceci étant dit, y compris en phase post-rigoureuse, le système 2 ne cesse pas d'éduquer le système 1 pour lui permettre de progresser. C'est ce qui s'est passé dans le cerveau de tous les théoriciens des nombres suite à la publication de Lemke Oliver et Soundararajan sur les derniers chiffres des nombres

¹⁹  *There's more to mathematics than rigour and proofs* | Terence Tao (2009)

premiers. Cette découverte surprit les systèmes 1 des théoriciens des nombres, dont l'intuition sur les nombres premiers se fondait sur l'heuristique du théorème des nombres premiers. Cette heuristique prédisait que les derniers chiffres des nombres premiers successifs étaient grossièrement indépendants. Cette heuristique était erronée dans ce cas. Depuis, les théoriciens des nombres ont sans doute effectué une inférence bayésienne approchée pour réduire leurs crédences en l'applicabilité de l'heuristique du théorème des nombres premiers aux nombres premiers consécutifs.

Le modèle à deux systèmes de pensée de Daniel Kahneman est assurément un modèle intéressant pour le *bayésien pragmatique*. Bien entendu, il est faux. Mais « tous les modèles sont faux ». Le modèle de Kahneman semble cependant *utile* — même si le chapitre 17 suggérera la présence d'un troisième système responsable du processus créatif.

En pratique, notamment pour affronter le *Big Data*, la plupart des algorithmes à utiliser sont sans doute des heuristiques rapides, peut-être même sous-linéaires. Il demeure néanmoins crucial de disposer d'algorithmes plus lents et plus justes. Mieux encore, si l'on dispose de quelques algorithmes lents dont on sait qu'ils sont assez justes, on peut alors les utiliser pour entraîner les heuristiques rapides.

Si je devais parier (et en bon bayésien, j'aime parier !), je dirais que c'est à cela que ressemblera l'intelligence artificielle du futur.

Les approximations de Bayes

Le *bayésien pragmatique* va donc devoir effectuer de nombreuses approximations de phénomènes qu'il étudie. Mais l'approximation la plus importante qu'il devra faire est surtout celle de l'équation qui lui permet d'apprendre : il lui faut trouver des heuristiques pour approcher la formule de Bayes. On peut grossièrement distinguer cinq approches pour ce faire.

Une première approche consiste à considérer uniquement un nombre restreint de modèles candidats. Ce nombre restreint peut être grand — on peut choisir de considérer des millions ou des milliards de modèles candidats. Cependant, pour que le calcul exact de la formule de Bayes soit faisable, il ne peut pas être exponentiellement grand, ni infini, comme cela serait le cas pour les modèles paramétrés. Cette approche pourra être typiquement complémentée par l'algorithme par *multiplicative weights update* dont on a parlé précédemment.

Une deuxième approche consiste à ne calculer qu'un modèle à forte crédence, voire à identifier le modèle le plus crédible. C'est ce que l'on appelle la maximisation de l'*a posteriori* (MAP), comme on l'a vu au chapitre 12. Autrement dit, le MAP va chercher une théorie prédictive T telle que $\mathbb{P}[T|D]$ est maximisée. Or, nous disposons de nombreux algorithmes pour maximiser ces quantités, comme la descente de gradient, l'algorithme *expectation-maximization* (EM) ou

les *generative adversarial networks* (GANs). Cependant, il ne faut pas perdre de vue que le MAP est une approximation grossière de la formule de Bayes ; en particulier, en s'arrêtant sur un seul modèle, il tombe dans le régime de l'*overfitting*.

Une troisième approche consiste à ignorer la fonction de partition, ce dénominateur de la formule de Bayes qui requiert la comparaison de tous les modèles imaginables. En particulier, dès lors, la somme des poids des différents modèles peut ne pas être égale à 1. On verra notamment au chapitre 17 comment, malgré cela, à l'aide d'algorithmes comme MCMC ou la divergence contrastive, il demeure néanmoins possible d'effectuer des prédictions.

La quatrième approche que j'aimerais mentionner est la plus étrange. Elle consiste à s'autoriser de violer les lois des probabilités. C'est par exemple ce qu'a proposé Samuel Rodrigues en 2014, en modifiant la définition des probabilités conditionnelles, et en autorisant ainsi la manipulation de probabilités qui ne satisfont pas la formule de Bayes²⁰. Dans un genre similaire, l'algorithme *Sum of Squares*, que certains chercheurs comme Boaz Barak et David Steurer pensent être un candidat pour être, en un sens, un algorithme *optimal*²¹, introduit la notion de pseudo-probabilité. Ces pseudo-probabilités généralisent celles que l'on connaît bien, mais peuvent prendre des valeurs négatives ! Cette largesse semble conduire à des résultats qui ont l'avantage d'être rapides à calculer, et qui, interprétés adéquatement, demeurent très utiles.

Enfin, une cinquième et dernière approche consiste à considérer un ensemble restreint de lois de probabilité dont la manipulation peut être faite de manière rapide ; quitte à ne pas tout à fait coller avec les données. Cette approche connaît plusieurs sous-variantes, dont les *gaussian mixture models*, les méthodes bayésiennes variationnelles et l'*expectation-propagation*. De façon cruciale, il est alors important de disposer d'une mesure de similarité entre des lois de probabilité. Ce n'est pas tâche aisée ! Quantifier adéquatement l'incertitude n'a rien d'intuitif. Ça tombe bien. Il s'agit du sujet du prochain chapitre.

Références en français

- ▶ *Les nombres premiers* | Science Étonnante | D. Louapre (2016)
- ▶ *Deux (deux ?) minutes pour l'hypothèse de Riemann* | El Jj | J. Cottanceau (2016)
- ▶ *Deux (deux ?) minutes pour l'éléphant de Fermi & Neumann* | El Jj | J. Cottanceau (2018)
- ▶ *Les développements limités* | Relativité 3 | Science4All | L.N. Hoang (2016)
- ▶ *Les nombres sont-ils (presque) aléatoires ?* | Science4All | L.N. Hoang (2016)
- ▶ *π est une fraude* | Science4All | L.N. Hoang (2017)

²⁰ *Probability Theory without Bayes' Rule* | S. Rodrigues (2014)

²¹ *Sum of Squares: An Optimal Algorithm?* ZettaBytes | B. Barak (2017)

- ▶ *L'intuition : à prendre ou à laisser ?* My4Cents (Toubkal) | Science4All | L.N. Hoang (2017)
- ▶ *Le test de Turing* | IA 3 | Science4All | A. Gelaude et L.N. Hoang (2017)
- ▶ *Les learning machines de Turing* | IA 7 | Science4All | L.N. Hoang (2018)

Références en anglais

- 📘 *Quantum Computing since Democritus* | Cambridge University Press | S. Aaronson (2013)
- 📘 *Deep Learning* | MIT Press | I. Goodfellow, Y. Bengio et A. Courville (2016)
- 📖 *Computing $\pi(x)$: the Meissel, Lehmer, Lagarias, Miller, Odlyzko method* | Mathematics of Computation of the AMS | M. Deléglise and J. Rivat (1996)
- 📖 *Unexpected biases in the distribution of consecutive primes* | Proceedings of the National Academy of Sciences | R. Lemke Oliver et K. Soundararajan (2016)
- 📖 *Computing Machinery and Intelligence* | Mind | A. Turing (1950)
- 📖 *Why Philosophers Should Care About Computational Complexity* | S. Aaronson (2011)
- 📖 *Nearly optimal sparse fourier transform* | Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of computing | H. Hassanieh, P. Indyk, D. Katabi and E. Price (2012)
- 📖 *Deep Learning Works it Practice. But Does it Work in Theory?* | R. Guerraoui and L.N. Hoang (2018)
- 🌐 *Mathematicians discover prime conspiracy* | Quanta Magazine | E. Klarreich (2016)
- 🌐 *There's more to mathematics than rigour and proofs* | T. Tao (2009)
- ▶ *Primes are like Weeds (PNT)* | Numberphile | J. Grime (2013)
- ▶ *The Riemann Hypothesis* | Singingbanana | J. Grime (2013)
- ▶ *The Science of Thinking* | Veritasium | D. Muller (2017)
- ▶ *Faster than the Fast Fourier Transform* | ZettaBytes | M. Kapralov (2017)
- ▶ *Why Computers are Bad at Algebra* | Infinite Series | K. Houston Edwards (2017)
- ▶ *Alan Turing's lost radio broadcast rerecorded* | Singingbanana | J. Grime (2017)

La certitude absolue est un privilège des esprits incultes et des fanatiques. C'est, pour la communauté scientifique, un idéal inaccessible.

Cassius J. Keyser (1862-1947)

Les légers succès peuvent être expliqués par la compétence et le travail. Les succès stupéfiant sont dus à la variance.

Nassim N. Taleb (1960-)

15

La faute à pas de chance

FiveThirtyEight et l'élection présidentielle de 2016

Coup de tonnerre. Contre toute attente, en ce 8 novembre 2016, le candidat Donald Trump est élu Président des États-Unis d'Amérique. Et si ce fut une surprise monumentale, c'est parce que tous les sondages avaient prédit la défaite de Trump. Le lendemain, mes collègues se sont empressés de me taquiner au sujet de l'échec des modèles bayésiens. En particulier, Nate Silver et son équipe à FiveThirtyEight¹ ne donnèrent que 28,6 % de chance pour l'élection de Trump, contre les écrasants 71,4 % pour Hilary Clinton. Le bayésianisme avait échoué.

Ou peut-être pas. Mon pari est que beaucoup de ceux qui ont vu ces chiffres les ont confondus avec les résultats d'une élection ; à savoir les pourcentages de voix reçus par les différents candidats. Mais ce n'est pas cela que mesurent les chiffres de FiveThirtyEight. Les chiffres de FiveThirtyEight sont des crédences bayésiennes au sujet de l'identité du futur président.

Revenons-en à l'énigme des deux enfants. Si vous n'en savez pas plus, il sera raisonnable de parier que ces deux enfants ne sont pas tout deux des garçons. En bons bayésiens, vous associerez à ce cas une probabilité de 25 % (en supposant pour simplifier qu'un enfant a une chance sur deux d'être un garçon). Imaginons que vous découvrez que les deux enfants sont effectivement deux garçons. Est-ce vraiment une raison suffisante pour rejeter l'approche bayésienne ?

¹  Who Will Win the Presidency? Election Forecast | Five Thirty Eight (2016)

J'irai même jusqu'à prétendre que la prédition de FiveThirtyEight est non pas une défaillance, mais un succès de l'approche bayésienne. Alors que de nombreux experts anticipaient déjà les décisions présidentielles de Clinton, une bonne interprétation des résultats de FiveThirtyEight aurait dû interpeler et exiger une grande prudence. En particulier, une probabilité de 28,6 % n'a rien de négligeable. Si un événement avec une telle probabilité survient, il n'y a pas lieu de le qualifier de surprise générale. Avoir deux garçons de suite, ce n'est pas défier les lois des probabilités.

Qui plus est, la *pure bayésienne* ne cherche jamais à juger la validité d'un modèle de manière isolée — si ce n'est pour souligner, encore et encore, que « tous les modèles sont faux ». La *pure bayésienne* va constamment juger la validité d'un modèle relativement à d'autres modèles. Or, dans le cas de l'élection de Donald Trump, nombreux étaient les modèles qui clamaient haut et fort sa défaite inéluctable. Comparativement à cette piètre concurrence, le modèle bayésien de FiveThirtyEight a clairement brillé, que ce soit dans le cas de l'élection de 2016, ou dans le cas des élections présidentielles précédentes.

Vous vous dites peut-être maintenant que le bayésien est avant tout un prophète qui ne se mouille jamais, voire un charlatan qui annonce à la fois la pluie et le beau temps. Les sciences, dit-on parfois, ne font pas de prédictions approximatives que les expériences ne sauront pas rejeter. Ce serait mal connaître l'histoire des sciences. « Les gens cherchent la certitude. Mais il n'y a pas de certitude », prévient Richard Feynman. Et il y a un domaine en particulier où le rôle central des probabilités a fini par faire consensus : la mécanique quantique.

La mécanique quantique est-elle probabiliste ?

L'incertitude de la mécanique quantique est habituellement illustrée par le chat de Schrödinger. Dans cette expérience de pensée, un chat est emprisonné dans une boîte, avec un dispositif qui libère un poison en cas de désintégration d'un atome radioactif. La mécanique quantique prédit alors que, tant que la boîte est fermée, le chat est dans une étrange superposition quantique, qui fait de lui un chat vivant *et* mort — qui est un cas mathématiquement distinct d'un chat vivant *ou* mort.

Mais mettons ces bizarries quantiques de côté. Concentrons-nous sur un aspect moins étrange de la mécanique quantique. Quand on ouvre la boîte, le chat a une certaine probabilité d'être vivant ; et une autre probabilité d'être mort. Il devient vivant *ou* mort. Mais surtout, l'état quantique du chat après observation est imprévisible. Il semble fondamentalement aléatoire.

Le physicien Erwin Schrödinger détestait cette conclusion. « Je n'aime pas ça, et je suis désolé d'avoir quelque chose à voir avec cela », écrivit-il. Il n'était pas le seul. « Dieu ne joue pas aux dés », ajouta Albert Einstein. Ce à quoi Niels Bohr aurait répondu : « Einstein, arrêtez de dire à Dieu ce qu'il doit faire. »

En 1935, Einstein, Podolski et Rosen publient un article stupéfiant, dans lequel ils prouvent que la mécanique quantique ne peut pas être une théorie locale. Autrement dit, ils prouvent que la mécanique quantique implique que deux particules très éloignées peuvent instantanément s'influencer, ce qu'Einstein surnomma « *spooky action at distance* ». Einstein rejette cette absurdité qui semblait violer le postulat de la relativité restreinte selon lequel la vitesse de la lumière est une vitesse maximale. Il en déduit que la mécanique quantique était incomplète, et, en particulier, que l'utilisation des probabilités était un pur artefact de notre ignorance.

Cependant, en 1982, Alain Aspect et ses collaborateurs ont montré expérimentalement l'existence de cette *spooky action at distance*. Voilà qui a amené Stephen Hawking à plaisanter : « non seulement Dieu joue aux dés, il les lance parfois là où on ne peut pas les trouver. »

De nos jours, une grande proportion des physiciens quantiques adhèrent à l'interprétation dite *de Copenhague*, en référence au physicien danois Niels Bohr. Selon cette interprétation, l'état quantique du chat au moment de l'ouverture de la boîte est probabiliste. Mais pas uniquement au sens épistémique. Selon cette interprétation, l'imprévisibilité de l'état quantique du chat n'est pas qu'un manquement dans nos connaissances. C'est un mécanisme fondamentalement encodé dans les lois de l'univers.

Cependant, on est loin d'un consensus scientifique quant à cette interprétation². Il existe ainsi de nombreuses explications alternatives. L'une d'elles est la théorie des *multivers quantiques* de Hugh Everett. Cependant, la proposition d'Everett fut rejetée par Niels Bohr. Dégouté, Everett arrêta la physique, généralisa l'utilisation des multiplicateurs de Lagrange en optimisation, et devint multimillionnaire. Triste sort.

Si le multivers d'Everett semble être une absurdité, notamment pour les adeptes de Popper, le bayésien Eliezer Yudkowsky le défend sur le blog Less Wrong³, en vertu notamment d'une version algorithmique du rasoir d'Ockham. Même si ses implications peuvent dépasser notre imagination limitée, l'interprétation d'Everett repose sur un unique principe très simple : et si la seule loi de l'univers était l'équation de Schrödinger, celle qui prédit l'évolution des états quantiques en l'absence d'observations ?

Cette proposition d'Everett a tout pour plaire à celui qui admire l'élégance et la simplicité des équations de la physique. Adieu à l'apparition d'un phénomène probabiliste dont le déclenchement coïncide avec l'obscur et ambiguë notion d'observation (qui peut très bien être l'observation d'un état quantique par une machine et n'a donc rien à voir avec la conscience). Selon l'interprétation d'Everett, l'incertitude de l'observation est en fait une intrication quantique entre les objets qui sont entrés en interaction au moment de l'observation. Étant nous-mêmes intriqués avec tout ce qui nous entoure, nous n'observons que les

²  Quantum Mechanics (an embarrassment) | Sixty Symbols | S. Carroll (2013)

³  If Many-Worlds Had Come First | Less Wrong | E. Yudkowsky (2008)

états quantiques de l'univers avec lesquels nous sommes intriqués. On dit parfois que l'on est ainsi bloqué dans l'une des nombreuses branches d'univers créées par l'équation de Schrödinger au moment de l'intrication⁴.

En particulier, dès lors, le phénomène probabiliste qui survient au moment de l'observation n'est plus intrinsèque aux lois de l'univers. Il s'agit d'un incertain épistémique, qui est dû à notre appartenance à une et une seule des branches possibles du multivers quantique. Bien entendu, les conséquences du postulat d'Everett sont abracadabantesques. Mais en bons bayésiens, ces conséquences inobservables n'importent pas au moment de juger de la crédence d'une théorie. Ce qui importe, ce sont le terme d'expérience de pensée, à savoir la capacité de la théorie à expliquer les données *observées*, et l'*a priori*, que Solomonoff mesure via la longueur de la plus courte description algorithmique de la théorie. L'interprétation d'Everett égale toute autre interprétation en termes prédictifs. Admettant cependant une description algorithmique nettement plus courte, elle semble devoir mériter les crédences des bayésiens.

Il me semble toutefois qu'il faille mettre des bémols à cet argument de Yudkowsky, surtout si l'on en croit le formalisme de Solomonoff du chapitre 7. En particulier, pour être prédictif, il semble nécessaire de combiner l'équation de Schrödinger à une description du complexe état physique du multivers. En particulier, à cause de l'interférence possible entre les branches quantiques, la description des univers quantiques alternatifs semble nécessaire pour permettre des prédictions. Or cette description semble extrêmement coûteuse en espace mémoire. Voilà qui semble la rendre peu crédible *a priori*...

N'étant pas expert en de telles questions et n'ayant donc qu'une crédence limitée en mon raisonnement, je préfère avouer l'étendue de mon ignorance à ce sujet. D'ailleurs, plusieurs autres interprétations de la mécanique quantique ont été proposées, comme la théorie déterministe mais non-locale de De Broglie-Bohm⁵, le *quantum Bayesianism* (QBism) ou n'importe laquelle des 10 autres variantes présentées sur la page Wikipedia anglophone⁶ en 2018.

Quoi qu'il en soit, quelle que soit votre interprétation préférée de la mécanique quantique, force est de constater que la théorie des probabilités joue un rôle critique en mécanique quantique. Quand il s'agit de prédire l'issue de la collision entre deux protons au grand collisionneur de hadrons du CERN, la meilleure description à ce jour consiste à assigner diverses probabilités à différentes issues possibles, à l'instar de la manière dont Nate Silver prédit l'issue d'une élection. Et si la mécanique quantique a tant de succès malgré l'indétermination de ses prédictions, c'est parce que les probabilités qu'elle assigne aux différentes issues sont remarquablement en phase avec les fréquences observées !

⁴ [The Many Worlds of the Quantum Multiverse | PBS Space Time | M. O'Dowd \(2016\)](#)

⁵ [Do we have to accept Quantum weirdness? De Broglie Bohm Pilot Wave Theory explained | Looking Glass Universe \(2017\)](#)

⁶ [Interpretations of quantum mechanics | Wikipedia \(2018\)](#)

La théorie du chaos

L'utilisation inévitable des probabilités semble loin d'être restreinte à la mécanique quantique et aux prédictions politiques de Nate Silver. En particulier, suite à la découverte du chaos, les mathématiciens ont fini par se convaincre que, dans bien des cas, il est impossible de faire mieux. C'est typiquement le cas de la météorologie. Alors que les équations de la météorologie sont assez bien connues, malgré le nombre croissant de capteurs en tout genre sur Terre et en orbite, les prédictions météorologiques à long terme restent très peu fiables. Cette incertitude inévitable des prédictions météorologiques a été prédite par le mathématicien Edward Lorenz dans les années 1960, au moment où Lorenz posa les fondements de ce que l'on appelle depuis la *théorie du chaos*.

Cette théorie fait le constat suivant : certains systèmes dynamiques simples sont incroyablement sensibles à des variations imperceptibles des conditions initiales. C'est ce qu'illustre parfaitement le *double pendule*, par opposition au pendule simple. Alors que Galilée remarqua avec stupéfaction l'incroyable régularité du pendule simple, dont la période d'oscillation est quasiment indépendante de l'amplitude d'oscillation, j'ai eu la chance d'être invité par le vidéaste Dr Nozman pour contempler l'étonnante imprévisibilité du double pendule⁷.

Ce dispositif est incroyablement simpliste : attachez un pendule à un pendule. Mettez le double pendule en position d'instabilité à la verticale, au-dessus du point d'équilibre. Et lâchez-le. Si le double pendule est suffisamment bien huilé, alors vous pouvez être sûr que la trajectoire qu'il dépeint est unique dans l'histoire de l'univers. Même si vous vouliez le répéter, vous ne pourriez pas obtenir à nouveau cette trajectoire, ni même une trajectoire qui ne serait que partiellement similaire. En effet, une variation imperceptible dans la condition initiale peut complètement bouleverser le comportement du double pendule, après seulement une poignée d'oscillations.

Depuis la découverte de Lorenz, les mathématiciens ont découvert que le chaos était loin d'être l'exception. Il semble même être la norme. Le monde réel est chaotique, et de petites fluctuations imperceptibles peuvent avoir des conséquences spectaculaires peu de temps plus tard. C'est ce que l'on appelle communément l'*effet papillon*, qui répond par l'affirmative à la question rhétorique de Philip Merilees : « Un battement d'ailes d'un papillon au Brésil peut-il causer une tornade au Texas ? » Bien entendu, ceci ne veut pas dire que le papillon est le seul et unique coupable de la tornade. Ce que cela signifie, c'est qu'aucune prédition météorologique à moyen terme ne peut être entièrement fiable, à moins de mesurer tous les mouvements de tous les papillons sur Terre⁸.

Mais l'imprévisibilité n'est pas qu'une propriété des systèmes complexes avec des états potentiellement très complexes.

⁷  *Cet objet est chaotique ! (double pendule)* | Dr. Nozman | L.N. Hoang et G. O'livry (2017)

⁸  *Effet Papillon et Théorie du Chaos* | Science Étonnante | D. Louapre (2018)

Les automates déterministes imprévisibles

Différents mathématiciens et informaticiens se sont tournés vers les automates pour étudier l'émergence de la complexité dans des systèmes dynamiques simplistes. Un automate est un univers virtuel, dont l'état physique évolue selon un temps discret. À chaque pas de temps, un nouvel état physique est calculé à partir de l'état précédent, selon une règle souvent simpliste.

Parmi les exemples classiques d'automates spectaculaires, on trouve les *automates de Wolfram*, qui forment un univers unidimensionnel infini, composé de cellules les unes à côtés des autres. Ces cellules peuvent être allumées ou éteintes. À l'instant initial, une seule cellule est allumée, et toutes les autres sont éteintes. À chaque pas de temps, chaque cellule s'allume ou s'éteint, en fonction de l'état des cellules voisines et de règles d'interaction entre cellules voisines. De façon stupéfiante, les simulations de Wolfram montrent que des règles simplistes peuvent conduire à des phénomènes imprévisibles. Ceci est d'ailleurs particulièrement le cas de la règle 30 de Wolfram, qui dessine des figures fractales surprenantes.

Mieux encore, le mathématicien John Conway a proposé des règles simples pour un automate de dimension 2 appelé le *jeu de la vie*. Là encore, il s'agit d'un ensemble de cellules, cette fois-ci disposées selon un quadrillage infini, et ces cellules s'allument ou s'éteignent en fonction de l'état des cellules voisines, conformément à des règles simples⁹. Néanmoins, ces règles simples ont été prouvées Turing-complètes. Autrement dit, tout calcul effectué par une machine peut être simulé par le jeu de la vie de Conway. En particulier, si l'on admet la thèse de Church-Turing, alors notre univers entier ne serait qu'un calcul, lequel pourrait donc être entièrement simulé par une énorme grille du jeu de la vie (dont la taille monstrueuse est au moins de l'ordre du googol, c'est-à-dire 10^{100}).

Un dernier exemple d'automate très étudié est celui de la *fourmi de Langton*. Une fourmi est placée sur une grille dont les cases sont des cellules allumées ou éteintes. Si la cellule est éteinte, alors la fourmi tourne à droite et avance d'une case. Sinon, elle tourne à gauche et avance d'une case. Quoi qu'il en soit, au moment de quitter sa case, la fourmi inverse l'état de la case. Si la case était allumée, elle s'éteint, et vice-versa. Démarrons l'automate avec une fourmi, tête en haut, sur un quadrillage dont toutes les cases sont éteintes. Sans trop de surprise, les 500 premiers mouvements de la fourmi sont relativement symétriques et structurés. Cependant, ces symétries apparentes semblent complètement détruites après quelques milliers de mouvements. Les mouvements de la fourmi paraissent alors aléatoires.

Mais ce n'est pas là le plus étrange. Après 10 000 mouvements, la fourmi se met tout à coup à suivre une trajectoire régulière et quasi-périodique qui la fait dévier dans la direction diagonale haut-gauche, qui se poursuit alors à l'infini. On appelle cette trajectoire l'*autoroute* de la fourmi de Langton. De

⁹  *Le jeu de la vie* | Science Étonnante | D. Louapre (2017)

façon étonnante, il s'agit encore d'un problème ouvert que de déterminer quelles conditions initiales finiront par tracer une autoroute¹⁰.

Ces automates montrent que des règles simples peuvent aisément conduire à des phénomènes qui semblent imprévisibles. Pour la *pure bayésienne* toutefois, ces règles n'ont en fait rien d'imprévisible. Il suffit de calculer les simulations pour les déterminer. Cependant, pour le commun des mortels, des phénomènes comme l'autoroute de la fourmi de Langton, sont, en pratique, très imprévisibles, ne serait-ce que parce qu'ils semblent nécessiter un grand nombre de calculs pour être prévus. On parle alors d'émergence.

La thermodynamique

L'autoroute semble émerger des lois fondamentales régissant le mouvement de la fourmi de Langton. De la même manière, il semble que, dans certaines conditions de température et de pression, les lois fondamentales d'interactions moléculaires font émerger les équations de la dynamique des fluides. De même, en 1872, Ludwig Boltzmann a fait émerger la seconde loi de la thermodynamique de l'hypothèse atomique. En particulier, Boltzmann a montré que l'irréversibilité du temps était une propriété émergente. Arrêtons-nous sur ce point.

L'un des tours de force de Boltzmann fut d'abord de relier l'hypothèse atomique à la notion d'entropie. Pour comprendre cela, il nous faut commencer par une observation triviale : mélanger le chaud et le froid fait du tiède. Et surtout, l'inverse n'est pas vrai. Versez-vous un verre tiède. Il est improbable que la partie droite du verre gèle pendant que la partie gauche bout. Autrement dit, l'énergie tend à s'homogénéiser, pas à se concentrer en un point.

Cette observation peut sembler triviale. Mais il fallut un génie, le physicien Rudolf Clausius, pour oser la prendre au sérieux. Clausius réussit à mathématiser ce principe, en introduisant une grandeur physique qu'il appela entropie. Dire que mélanger du chaud et du froid fait du tiède revient alors à affirmer que l'entropie d'un système fermé augmente. C'est cette seconde affirmation que Clausius promut au rang de seconde loi de la thermodynamique¹¹.

Cependant, l'entropie de Clausius demeurait obscure et mal comprise. Le génie de Boltzmann fut de définir l'entropie à partir de l'hypothèse atomique, jetant ainsi au passage les premières briques de la physique statistique. Boltzmann fut si fier de sa définition qu'il la fit inscrire sur sa tombe : $S = k \ln W$. Que signifie cette équation ? Il faut d'abord comprendre que nos appareils de mesures thermodynamiques ne sont pas capables de mesurer les positions et vitesses des 10^{26} particules qui nous entourent. Après tout, il faudrait des millions de zettaoctets pour ce faire. Par opposition, les grandeurs thermodynamiques

¹⁰  *La fourmi de Langton* | Science Étonnante | D. Louapre (2015)

¹¹  *L'entropie* | Passe-Science | T. Cabaret (2016)

que nous mesurons, comme la pression, la température, le volume ou la masse, sont des grandeurs qui résument le comportement d'un très grand nombre de particules. On parle de grandeurs *macroscopiques*, par opposition aux grandeurs *microscopiques* qui correspondent directement aux particules.

Le génie de Boltzmann fut de remarquer que l'entropie permet exactement de quantifier l'incertitude microscopique, une fois les grandeurs macroscopiques connues. Plus précisément, Boltzmann montra que l'entropie S étudiée par Clausius n'était autre que le logarithme du nombre W d'états microscopiques qui sont cohérents avec les grandeurs macroscopiques, à un facteur multiplicatif k près que l'on appelle désormais la constante de Boltzmann. L'entropie, cette grandeur dont Clausius avait prédit la croissance inéluctable, n'était autre qu'une quantification de l'incertitude microscopique qui demeure, une fois les mesures macroscopiques effectuées.

L'entropie de Shannon

Quantifier l'incertitude peut sembler inutile, voire absurde. Pourtant, il y a un contexte où la maîtrise de l'incertitude a joué un rôle majeur dans l'histoire de l'humanité : le décodage des codes nazis. Pendant la seconde guerre mondiale, l'Anglais Alan Turing et l'Américain Claude Shannon se sont rencontrés pour échanger leurs connaissances cryptographiques. Il semble que Shannon et Turing n'ont que très peu parlé de cryptographie¹², mais que tous deux avaient compris l'importance de la quantification de l'incertitude. En particulier, Turing introduisit les calculs de *banburismus* pendant la guerre pour inférer une crédence en le fait que différents messages étaient cryptés avec la même configuration de la machine Enigma¹³. Shannon alla plus loin encore.

En 1948, Shannon publie l'un des articles les plus influents de l'histoire de l'humanité. Cet article est intitulé *A Mathematical Theory of Communication*. Cet article sublime propose de modéliser les messages envoyés par une source par une loi de probabilité. En termes bayésiens, ceci revient à considérer un *a priori* sur les messages que cette source sera amenée à énoncer. Un soldat nazi a ainsi de bonnes chances d'insérer « *Heil Hitler* » quelque part dans son message, d'utiliser des mots allemands ou de n'émettre que la traduction allemande de « rien à déclarer ». Les messages nazis étaient aléatoires ; mais ils n'avaient rien d'arbitraire.

Le premier coup de génie de Shannon fut d'identifier la quantité d'informations d'un message avec sa rareté vis-à-vis de la croyance bayésienne. Par exemple, le mot « Lé » contient beaucoup d'information en Europe. Il m'identifie de

¹²  *A Mind at Play: How Claude Shannon Invented the Information Age* | Simon & Schuster | J. Soni and R. Goodman (2017)

¹³  *Maths from the talk "Alan Turing and the Enigma Machine"* | Singing Banana | J. Grime (2013)

manière quasi-sûre. C'est parce que ce mot est très rarement utilisé en Europe. C'est cette rareté qui lui permet de communiquer beaucoup d'informations.

A contrario, ce même mot à Hanoi ne contient presque aucune information. Il ne permet de faire référence qu'aux quelques milliers, voire centaine de milliers de Vietnamiens dont il s'agit du prénom. Le fait que ce mot est très fréquent implique la faible quantité d'informations qui lui est associé. Autrement dit, l'information d'un message ne peut se mesurer que relativement à un contexte, et plus exactement, à une croyance bayésienne qui détermine la probabilité du message. *Sans conteste, sans contexte, c'est la mauvaise probabilité qu'on teste.*

Le second coup de génie de Shannon fut d'utiliser le logarithme pour quantifier la quantité d'informations d'un message. Pourquoi le logarithme ? C'est pour que la quantité d'informations de deux messages indépendants soit la somme des quantités d'informations de chaque message. Or, la probabilité de deux messages indépendants est le produit des probabilités des deux messages. Pour que le produit devienne une somme, il nous faut utiliser l'objet mathématique qui traduit les produits en sommes. Comme on l'a vu au chapitre 11, cet objet est le logarithme.

Plus précisément, Shannon définit la quantité d'information d'un message m de probabilité $p(m)$ par $h(m) = \log_2(1/p(m))$. Autrement dit, la quantité d'informations d'un tel message est l'exposant $h(m)$ tel que $p(m) = 1/2^{h(m)}$. Les messages à très faibles probabilités auront alors une grande quantité d'informations $h(m)$. Enfin, Shannon déduit de cette formule la quantité d'informations espérée H d'une source d'information. Il s'agit de la quantité moyenne d'informations $h(m)$ que cette source émet. Autrement dit, il pose

$$H = \mathbb{E}_m[h(m)] = \sum_m p(m) \log_2(1/p(m)).$$

Cette quantité H , Shannon voulut l'appeler l'information espérée de la source, ou encore fonction d'incertitude. Mais il finit par suivre le conseil que John von Neumann lui donna : « Tu devrais l'appeler entropie, pour deux raisons. En premier lieu, ta fonction d'incertitude a été utilisée en mécanique statistique sous ce nom, et elle a donc déjà un nom. En second lieu, et de façon plus importante, personne ne sait ce que l'entropie est vraiment, donc dans tout débat, tu auras toujours un coup d'avance. »

Mais l'entropie de Shannon est-elle vraiment la même que l'entropie de Boltzmann ? Oui. Il s'agit en fait d'une généralisation. Pour comprendre cela, il faut bien voir qu'étant donné des mesures macroscopiques, on dispose d'un *a priori* sur les différents états microscopiques plausibles. Cependant, Boltzmann a montré que chacun des W états microscopiques cohérents avec les mesures macroscopiques était, à l'équilibre thermodynamique, équiprobable. Par conséquent chaque état microscopique a une probabilité $1/W$ de se produire. En remplaçant $p(m)$ par $1/W$ dans l'équation de Shannon, on en déduit alors que,

dans le cas où W états microscopiques sont équiprobables, l'entropie d'un système thermodynamique est alors $H = \log_2(W)$. Réajuster les unités pour les rendre compatibles avec le système international de la physique force alors à ajouter un coefficient multiplicatif k .

Shannon avait bel et bien généralisé l'entropie de Boltzmann.

La compression optimale de Shannon

Le troisième coup de génie de Shannon fut de comprendre ce que l'entropie mesure vraiment. Aussi étrange que cela puisse paraître, l'entropie mesure la compression optimale des messages. Autrement dit, elle mesure le nombre minimal de bits pour stocker un message sur un disque dur, ou le temps minimal pour envoyer ce message à travers un câble dont le débit est limité. En effet, Shannon a prouvé qu'il sera impossible de faire mieux que la limite fondamentale calculée par son entropie.

Pour comprendre le lien entre l'entropie de Shannon et la compression, étudions le jeu de société « Qui est-ce ? ». Dans ce jeu, chaque joueur doit deviner le personnage de l'autre parmi un ensemble de personnages possibles. Pour ce faire, chaque joueur doit poser une question binaire de la forme suivante : le personnage à deviner est-il un homme ? Porte-t-il des lunettes ? A-t-il des longs cheveux ? Chaque joueur pose une question, l'un après l'autre. Le premier à deviner le personnage de l'autre gagne.

Dans son article de 1948, Shannon prouve que s'il y a n personnages possibles et si l'adversaire choisit uniformément aléatoirement l'un des n personnages, alors il faudra nécessairement, en moyenne, au moins $\log_2(n)$ questions binaires pour déterminer le personnage de l'autre. Mieux encore, supposons que l'on sache que l'adversaire choisit plus souvent un personnage masculin que féminin et qu'il préfère les individus avec des lunettes. Autrement dit, supposons que l'on ait une croyance bayésienne justifiée en ce que l'adversaire pourrait avoir choisi. Alors, l'article de Shannon prouve que le nombre moyen de questions sera alors au moins l'entropie de Shannon.

Mieux encore, l'entropie de Shannon correspond à un cas idéalisé où la suite de réponses fournie par l'adversaire détermine l'encodage optimal des identités des personnages choisis par l'adversaire. Plus précisément, l'encodage proposé par Shannon consiste à étiqueter le personnage par une suite de 0 et de 1, où le 0 correspond au non, et le 1 au oui. Ainsi, si l'adversaire répond oui puis non aux deux premières questions binaires qu'on lui pose, alors l'encodage optimal commence par 1 puis 0. Les identités des personnages sont alors identifiées à des suites de 0 et de 1, que Shannon appela *binary digits*, ou *bits* pour faire plus court¹⁴.

¹⁴  *Entropy as a Fundamental Compression Limit* | ZettaBytes | R. Urbanke (2017)

Plus généralement, Shannon prouva que toute communication se réduisait à une séquence de 0 et de 1, et que la communication avait tout à gagner à être ainsi numérisée. Aujourd’hui, cette observation peut paraître triviale. Ce n’était pas le cas à l’époque, alors que beaucoup misaient encore sur la technologie analogique. Grâce à son article de 1948, Shannon avait déclenché l’ère du numérique.

La redondance de Shannon

Le quatrième coup de génie de Shannon fut de montrer comment communiquer à travers des canaux imparfaits. En pratique, quand un message électrique est envoyé entre A et B, ce signal peut aisément être perturbé par diverses interférences. Des 1 peuvent s’être transformés en 0, et vice-versa. Pour adresser ce cas, Shannon eut l’idée d’introduire une croyance bayésienne sur les perturbations qu’un message peut avoir reçues. Shannon prouva ensuite que, pourvu que la croyance bayésienne soit justifiée, le canal imparfait est dès lors équivalent à un canal parfait dont le débit est égal au débit du canal imparfait moins une sorte d’entropie des imperfections du canal. En particulier, tout message pouvait encore être communiqué via un canal imparfait, à condition d’y ajouter une redondance suffisamment grande. Shannon quantifia même la redondance nécessaire : elle doit être égale, en gros, à l’entropie des perturbations subies par les messages qui traversent le canal imparfait¹⁵.

Tout cela peut paraître bien obscur. Pourtant, la redondance est un phénomène qui nous est très familier, même si l’on ne s’en rend pas toujours compte. Lorsque vous discutez avec vos amis dans un bar bruyant, il y a peu de chance que vous entendiez tout ce qui se dit. Néanmoins, il n’est généralement pas nécessaire d’entendre tout ce qui se dit pour comprendre ce qui se dit. En effet, une grande partie des mots de la langue française ont un rôle mineur dans le sens des phrases. Enlevez tous les articles ou verbes peu importants d’une phrase. Vous verrez il aisément deviner j’essaie dire.

La langue française est très redondante. C’est ce qui explique que les traductions françaises sont souvent plus longues que leurs versions anglaises. Voilà qui explique aussi pourquoi les Français parlent plus vite que les Anglais. Le débit d’information reste à peu près le même, car même si les Français disent plus de syllabes que les Anglais, les syllabes françaises contiennent plus de redondances, et donc moins d’informations, que les syllabes anglaises.

Aujourd’hui, tous les concepts de Shannon, que ce soient les bits, l’entropie, la capacité d’un canal ou la redondance, sont devenus des outils centraux des technologies de l’information. Mais leurs applications transcendent de loin le monde des technologies. On les retrouve bien sûr en physique statistique, pour étudier l’évolution des gaz, mais aussi en linguistique pour comprendre l’évolution du

¹⁵  *Shannon’s Optimal Communication* | ZettaBytes | R. Urbanke (2017)

langage, et même en (exo)-biologie pour détecter de la vie intelligente, dont les communications possèdent sans doute une redondance similaire aux communications entre humains adultes, ou entre dauphins adultes. C'est typiquement cette redondance du langage qui nous permet de finir les phrases des. Et il en serait sans doute de même pour les phrases de vies intelligentes extra-terrestres, puisque ces vies intelligentes chercheront sans doute à communiquer à travers des canaux imparfaits¹⁶.

La divergence de Kullback-Leibler

Mais ce n'est pas tout ! L'une des applications des concepts de Shannon fut d'enfin proposer une façon de juger la validité de prédictions probabilistes comme celle de FiveThirtyEight. Rappelons que FiveThirtyEight avait attribué une probabilité de 28,6 % à l'élection de Trump. Si une telle probabilité devait être encodée à la Shannon, il faudrait $\log_2(1/0,286) \approx 1,8$ bit. Et bien, on peut considérer là qu'il s'agit du nombre de points perdus par le modèle prédictif de FiveThirtyEight. Plus généralement, on peut ainsi comptabiliser le nombre total de points perdus par un modèle probabiliste prédictif en ajoutant les $\log_2(1/p(m))$, où les $p(m)$ sont les probabilités que le modèle assigne aux différents événements m qui sont survenus.

Pourquoi est-ce une excellente idée de mesurer les performances des prédictions probabilistes à l'aide d'une approche à la Shannon ? C'est parce que Shannon prouva que, si le monde était vraiment probabiliste conformément à une loi q , alors le modèle prédictif qui minimiseraient les points perdus serait celui qui prédit la loi $p = q$. Autrement dit, avec cette quantification de l'incertitude, prédire l'incertitude n'est pas préjudiciable. En particulier, quand il y a lieu d'être incertain, comme c'est le cas dans les systèmes chaotiques par exemple, la meilleure prédition est alors une prédition probabiliste. Le genre de prédition qu'un bayésien est amené à faire.

En particulier, sachant que le nombre de points perdus est minimisé par la prédition probabiliste q , on peut comparer la performance d'une prédition p relativement à la prédition optimale q . Pour cela, on est amené à calculer les différences de scores espérés :

$$\begin{aligned} D_{KL}(q||p) &= \mathbb{E}_{m \leftarrow q} \left[\log_2 \frac{1}{p(m)} - \log_2 \frac{1}{q(m)} \right] \\ &= \sum_m q(m) \log_2 \frac{q(m)}{p(m)}. \end{aligned}$$

¹⁶  *The History of SETI (Search for Extraterrestrial Intelligence)* | Art of the Problem (2014)

Introduite par Solomon Kullback et Richard Leibler en 1951, cette quantité est aujourd’hui connue sous le nom de divergence KL (même si les physiciens vont préférer la calculer avec le logarithme naturel). Elle calcule l’erreur de la prédiction p par rapport à la prédiction optimale q , et est toujours supérieure ou égale à 0. Il s’agit donc d’une mesure d’à quel point la prédiction p diverge de la prédiction optimale.

À bien des égards, la divergence KL est une excellente quantification de la justesse d’une prédiction probabiliste. Ou du moins, il s’agit d’une bien meilleure approche que l’approche intuitive et naïve qui consiste à rejeter la prédiction de FiveThirtyEight, uniquement parce que celle-ci attribuait une probabilité minoritaire à l’événement qui est survenu. Trop souvent, notre réaction consiste à simplifier à outrance les modèles probabilistes et à en tirer une prédiction déterministe. Trop souvent, nous voulons déterminer qui a raison et qui a tort, et omettons ceux dont l’avis était justifiablement partagé. Trop souvent, nous ignorons la prudence dont les modèles probabilistes et le *bayésianisme* nous invitent à faire preuve.

On en vient ainsi bien souvent à reprocher à ceux dont les prédictions sont mitigées de ne pas se mouiller. Pourtant, une prédiction qui minimise la divergence KL n’a aucune raison d’être une telle prédiction. S’il y a une alternative beaucoup plus probable que toute autre, la prédiction optimale est de lui accorder une crédence beaucoup plus grande qu’à toute autre. Autrement dit, la divergence KL permet parfaitement de discriminer les prédictions qui ne prennent pas position parce qu’elles ne connaissent pas le problème (ourtant peu incertain) auquel elles sont confrontées, des prédictions qui ne prennent pas position parce qu’elles savent que le problème est fondamentalement difficile et plein d’incertitudes.

Malheureusement, en pratique, rares sont les fois où la divergence KL est utilisée pour juger de la validité de nos prédictions. En général, parce qu’on cherche à récompenser (financièrement ou socialement) ceux qui ont fait la bonne prédiction déterministe, on est inévitablement poussé à effectuer des prédictions déterministes, et à ignorer la prudence dont il faudrait pourtant faire preuve. L’absence de quantification adéquate de l’incertitude nous pousse à la sur-interprétation et à tomber dans le piège du biais du survivant. Ce n’est pas aux experts sans avis tranchés qu’on tend le micro. Ce ne sont pas eux que l’on diffuse aux heures de grande écoute. Et ce ne sont pas eux qui sont massivement relayés par les réseaux sociaux.

Pire encore, peut-être parce qu’on a tendance à davantage se souvenir de nos triomphes que de nos échecs, on est constamment beaucoup trop sûrs de nous-mêmes. On se met régulièrement en excès de confiance. Modifier la manière dont on juge la validité des prédictions semble être un premier pas indispensable pour combattre cet excès de confiance ; et appliquer la formule de Bayes serait la voie idéale pour juger du degré de confiance adéquat à avoir.

La métrique de Wasserstein

La divergence KL n'est toutefois pas la seule mesure possible des performances des prédictions probabilistes. En fait, il en existe un très grand nombre¹⁷. Cependant, contrairement la divergence KL, beaucoup de ces mesures se prêtent mal à la recherche algorithmique d'une bonne prédiction p .

Il existe toutefois une autre façon de mesurer la distance entre différentes probabilités p et q qui, elle aussi, se prête bien aux calculs algorithmiques. Il s'agit de la métrique de Wasserstein, aussi connue sous le nom de distance du canonnier ou solution optimale du problème du transport. L'avantage de cette métrique, c'est qu'elle prend en compte à quel point un événement m est différent de m' , ce que ne fait pas la divergence KL. Ainsi, si vous prédissez que la couleur d'une tache sera jaune, et si je prédis qu'elle sera bleue et si elle est en fait jaune-orangée, alors la divergence KL dira que l'on s'est tous deux trompés. Cependant, vous étiez plus proche de la bonne réponse que moi. La métrique de Wasserstein permet de donner un sens précis à l'intuition selon laquelle vous aviez plus raison que moi.

Détaillons. Imaginez que vous deviez prédire là où surgiront les taches de café sur le sol d'un entrepôt au cours de l'année à venir. En bons bayésiens, votre prédiction est probabiliste, et elle dit que certains foyers sont plus probables *a priori* que d'autres. Faisons simple. Disons que vous mettiez mille grains de sable noir pour que la densité de sable noir corresponde à votre prédiction probabiliste. Autrement dit, vous mettez beaucoup de grains de sables là où vous pensez que les taches de café ont beaucoup de chance d'apparaître.

L'année passe et les employés ont été particulièrement maladroits. Ils ont fait tomber mille taches de café. Pour chaque tache, on place un grain de sable jaune. Sur le sol de l'entrepôt, il y a maintenant mille grains de sable jaune, et mille grains de sable noir. Vous disposez maintenant de mille fourmis. Chaque fourmi doit prendre un grain de sable noir et l'amener à côté d'un grain de sable jaune, de sorte qu'à la fin, les tas de sable noir et de sable jaune soient identiques. Chaque fourmi peut se balader très rapidement et sans problème, sauf lorsqu'elle porte un grain de sable, auquel cas elle avance extrêmement lentement. Enfin, les fourmis se coordonnent pour résoudre le problème du transport de sable noir le plus vite possible. Et bien, le temps moyen que ces fourmis mettront est justement la distance de Wasserstein entre votre prédiction probabiliste et les données empiriques.

Toutefois, si la métrique de Wasserstein se prête généralement bien aux calculs par ordinateur, elle présuppose une mesure de similarité entre les données. Dans notre cas, on pouvait mesurer la similarité entre deux grains de sable à l'aide de la distance qui sépare ces grains de sable. Cependant, dans de nombreux cas, identifier une mesure de similarité pertinente est une difficulté majeure.

¹⁷  *Statistical distance* | Wikipedia (2018)

Les *Generative Adversarial Networks* (GANs)

Dessinez plusieurs chats sur un bout de papier. Lequel des chats que vous aurez dessiné ressemble-t-il le plus à un « vrai » chat ? Et comment pourriez-vous quantifier la ressemblance entre vos dessins et des chats authentiques ? Comment définir une métrique de similarité entre des images ?

L'intérêt de ces questions ne se restreint d'ailleurs pas qu'au *Pictionary*. Déterminer la similarité entre des objets complexes comme du texte, du son ou des images¹⁸, est devenu l'un des plus grands défis du *bayésien pragmatique*. En effet, les données vraiment intéressantes sont en fait de tels objets complexes, qu'il s'agisse des oeuvres de Laplace, d'électrocardiogrammes ou d'images de tardigrades au microscope.

Prenons l'exemple de la cosmologie. De nos jours, les données de ce domaine sont essentiellement des photographies du ciel dans toutes les longueurs d'onde, des ondes radios aux rayons gammas, en passant par les micro-ondes, les infrarouges, la lumière visible, les ultraviolets et les rayons X. La question qui fascine les astrophysiciens est alors celle de déterminer les paramètres crédibles θ des modèles cosmologiques, sachant les photographies du ciel . Voilà un cas typique d'application de la formule de Bayes !

$$\mathbb{P}[\theta | \square] = \frac{\mathbb{P}[\square | \theta] \mathbb{P}[\theta]}{\mathbb{P}[\square | \theta] \mathbb{P}[\theta] + \sum_{\omega \neq \theta} \mathbb{P}[\square | \omega] \mathbb{P}[\omega]}$$

Ce calcul est ais  pour la *pure bay sienne*. Cependant, il est inaccessible au *bay sien pragmatique*. Comme souvent, le d nominateur est beaucoup trop long   calculer. Mais ce n'est pas tout. Parce que les mod les cosmologiques sont tr s complexes, m me les termes d'exp rience de pens e $\mathbb{P}[\square | \theta]$ n cessitent des temps de calcul surr alistes. En fait, notamment parce qu'il s'agit de r seaux bay siens¹⁹, ces mod les cosmologiques sont surtout con us pour permettre des simulations. Autrement dit, tout ce que l'on peut faire en temps raisonnable, c'est g n rer des images probables ,   supposer que les param tres des mod les cosmologiques sont θ . On parle de *mod les g n ratifs*²⁰.

D s lors, le *bay sien pragmatique* va devoir s'appuyer sur des m thodes qui contournent le calcul direct des termes d'exp rience de pens e. On parle d'*inf rences sans vraisemblance*, ou *likelihood-free methods*. Il en existe de nombreuses variantes, appel es *Approximate Bayesian Computation* (ABC) ou *parametric Bayesian Indirect Likelihood* (pBIL). Mais depuis 2014 et les travaux de Ian

¹⁸Ou mieux encore, entre des lois de probabilit s sur les textes, sons et images.

¹⁹On reparlera de cela au chapitre 17.

²⁰Autrement dit, un mod le g n ratif est une loi de probabilit  sur un ensemble complexe comme l'ensemble des images, et que l'on ne peut  tudier que via des simulations du mod les. Ou dit encore autrement, il s'agit d'un mod le con u pour  tre chantillonn . On en reparlera davantage au chapitre 17.

Goodfellow et de ses collaborateurs²¹, une méthode sans vraisemblance en particulier semble avoir gagné les crédences pragmatiques de nombreux chercheurs, à savoir les *Generative Adversarial Networks* (GANs).

Intuitivement, l'idée des GANs est de remplacer l'humain qui jouerait au *Pictionary* par un algorithme appelé *adversaire*, ou *professeur*. Cet *adversaire* sera en charge de mesurer la ressemblance entre des images générées par un modèle et des images authentiques. Pour cela, on tire une pièce. Si la pièce tombe sur pile, on choisit une image authentique parmi une grande banque d'images authentiques. Sinon, on demande au modèle de générer une image. Ensuite, quelle que soit l'image, on demande à l'*adversaire* d'exprimer une croyance bayésienne p en l'authenticité de l'image. Intuitivement, si le modèle est bon, alors l'*adversaire* devrait être confus. Il assignerait alors une probabilité $p = 1/2$ quelle que soit l'image.

De façon cruciale, pour s'assurer que l'*adversaire* ait tout intérêt à répondre une croyance bayésienne, Goodfellow et ses collaborateurs proposent d'assigner un coût à l'*adversaire* égal à $\log_2(1/p)$ si l'image est authentique, et $\log_2(1/(1-p))$. Autrement dit, le comptage des points se fait conformément à la divergence de Kullback-Leibler.

Par ailleurs, le modèle pourra ensuite chercher à déterminer ses paramètres pour lesquels les données générées seront le plus semblables aux données authentiques. De façon cruciale, il y aura un sens précis à cela. Il s'agira de générer des données pour lesquels les valeurs de p de l'*adversaire* seront constamment le plus proches possibles de $p = 1/2$, qu'il s'agisse d'images authentiques ou d'images générées par le modèle. En fait, la même quantité qui a été utilisée pour comptabiliser les points malus de la croyance bayésienne p de l'*adversaire* peut être utilisée pour mesurer la performance du modèle²² !

Au moment où j'écris ces lignes, les GANs ont particulièrement le vent en poupe. Leurs époustouflantes performances ne cessent de s'améliorer, en profitant notamment du *deep learning*. En effet, le modèle et l'*adversaire* des GANs de 2018 sont des réseaux de neurones²³ (de convolution). Grâce à leur approximation de la formule de Bayes sans vraisemblance et aux outils usuels d'apprentissage pour réseaux de neurones, les GANs ont acquis la capacité à générer des photographies dont l'indiscernabilité avec des photographies authentiques est bluffante²⁴. À bien des égards, les performances spectaculaires de ces intelligences artificielles

²¹  *Generative Adversarial Nets* | NIPS | I. Goodfellow J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, et Y. Bengio (2014)

²² Le modèle et l'*adversaire* jouent alors un jeu à somme nulle !

²³ Le modèle est typiquement un réseau de neurones profond \mathcal{G} combiné à une distribution « simple » qui génère des variables z . La simulation produit alors des données $\mathcal{G}(z)$. L'*adversaire* est un réseau de neurones profond \mathcal{D} . L'avantage de cette structure, c'est que l'on peut appliquer l'algorithme de rétropropagation. Cet algorithme identifie pourquoi \mathcal{D} a choisi $p \neq 1/2$, et remonte cette information au modèle génératif \mathcal{G} pour que celui-ci s'améliore. En cela, \mathcal{D} est davantage un *professeur* qu'un *adversaire*.

²⁴  *How an A.I. ‘Cat-and-Mouse Game’ Generates Believable Fake Photos* | New York Times | C. Metz et K. Collins (2018)

modernes n'ont été permises que par l'avènement de nouvelles façons de mesurer la crédibilité de prédictions probabilistes.

D'ailleurs, pour conclure ce chapitre, j'aimerais insister une fois de plus sur l'importance de la prise en compte des incertitudes dans notre façon de juger la qualité des prédictions. Sous-estimer le rôle du hasard dans l'analyse de phénomènes est malheureusement courant. Il est indispensable de prendre l'habitude de constamment raisonner avec l'incertitude, plutôt que de chercher à imposer des prédictions déterministes à des contextes pourtant fondamentalement imprévisibles. Dans notre complexe univers, aucune connaissance ne peut être certaine, que ce soit à cause d'une nature physique de la réalité, de notre manque de données empiriques, de la présence de phénomènes chaotiques ou de nos limites en puissance de calcul.

La quasi-totalité des questions de prédictions n'ont pas de réponse simple et univoque. Elles requièrent de la prudence. Et cette prudence ne pourra être justifiée qu'à partir du moment où l'on accepte enfin que de nombreux événements sont la faute à pas de chance. Dès lors, le jugement de nos modèles et de nos prédictions doit absolument se donner les moyens de quantifier l'incertitude. *Quantifier l'incertitude est trop important pour être laissé au hasard.*

Références en français

- ▶ *Cet objet est chaotique ! (double pendule)* | Dr. Nozman | L.N. Hoang et G. O'livry (2017)
- ▶ *Automate cellulaire* | Passe-Science | T. Cabaret (2015)
- ▶ *L'Entropie* | Passe-Science | T. Cabaret (2016)
- ▶ *La mécanique quantique en 7 idées* | Science Étonnante | D. Louapre (2015)
- ▶ *La fourmi de Langton* | Science Étonnante | D. Louapre (2015)
- ▶ *Effet Papillon et Théorie du Chaos* | Science Étonnante | D. Louapre (2018)

Références en anglais

- 📘 *The Signal and the Noise: Why So Many Predictions Fail—but Some Don't* | Penguin Books | N. Silver (2015)
- 📘 *The Black Swan: The Impact of the Highly Improbable* | Random House | N.N. Taleb (2007)
- 📘 *A Mind at Play: How Claude Shannon Invented the Information Age* | Simon & Schuster | J. Soni and R. Goodman (2017)
- 📘 *A Mathematical Theory of Communication* | The Bell System Technical Journal | C. Shannon (1948)
- 📘 *On information and sufficiency* | Annals of Mathematical Statistics | S. Kullback and R. Leibler (1951)

🖼 *Generative Adversarial Nets* | NIPS | I. Goodfellow J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, et Y. Bengio (2014)

🖼 *Maths from the talk "Alan Turing and the Enigma Machine"* | Singingbanana | J. Grime (2013)

🌐 *Who Will Win the Presidency?* Election Forecast | Five Thirty Eight (2016)

🌐 *If Many-Worlds Had Come First* | Less Wrong | E. Yudkowsky (2008)

🌐 *Shannon's Information Theory* | Science4All | L.N. Hoang (2013)

🌐 *Entropy and the Second Law of Thermodynamics* | Science4All | L.N. Hoang (2013)

🌐 *Statistical distance* | Wikipedia (2018)

🌐 *How an A.I. 'Cat-and-Mouse Game' Generates Believable Fake Photos* | New York Times | C. Metz et K. Collins (2018)

▶ *Inventing Game of Life* | Numberphile | J. Conway (2014)

▶ *Quantum Mechanics (an embarrassment)* | Sixty Symbols | S. Carroll (2013)

▶ *The Many Worlds of the Quantum Multiverse* | PBS Space Time | M. O'Dowd (2016)

▶ *Do we have to accept Quantum weirdness? De Broglie Bohm Pilot Wave Theory explained* | Looking Glass Universe (2017)

▶ *The Higgs Mechanism Explained* | Space Time | PBS Digital Studios (2015)

▶ *What is Information Entropy? (Shannon's formula)* | Art of the Problem (2013)

▶ *Entropy as a Fundamental Compression Limit* | ZettaBytes | R. Urbanke (2017)

▶ *Shannon's Optimal Communication* | ZettaBytes | R. Urbanke (2017)

Notre mémoire est un monde plus parfait que l'univers : elle rend vie à ce qui n'existe plus !

Guy de Maupassant (1850-1893)

Ne faites pas confiance à votre mémoire. Si quelqu'un vous demande si vous pouvez vous souvenir de quelque chose, dites non.

Julia Shaw (1987-)

16

Trou de mémoire

La valeur des données

Les dernières années ont vu l'émergence d'un nouveau *buzzword* : le *Big Data*. Cependant, pour de nombreux experts du domaine, ce *buzzword* ne révèle pas un changement de phase de nos économies. Ça fait longtemps que le *Big Data* est là. Et ça fait longtemps qu'il n'a cessé d'augmenter à un rythme exponentiel. On parle, après tout, d'*informatique*, et le mot *information* n'est autre qu'un synonyme du mot *data*. Cependant, à défaut de modifier l'état de l'art des technologies de l'information, la notion de *Big Data* semble souligner le rôle central des données dans nos industries, nos économies et dans nos sociétés.

J'en fus particulièrement frappé en emménageant en Suisse. La recherche d'un logement, le dossier de location et le contrat entre bailleur et locataire sont d'effroyables parcours administratifs, qui ont le don de rappeler l'un des douze travaux d'Astérix. Ainsi, pour sonder le marché des locations d'appartements de Lausanne, il m'a fallu m'enregistrer à des mailing lists, des groupes Facebook et des sites d'agences de location. Cette recherche aura représenté des heures et des jours de travail et d'efforts cognitifs, que j'aurais pourtant pu aisément déléguer à un algorithme de recommandation performant.

Pire encore, le dossier de location consistait en une collecte de documents auprès de nombreuses entités comme l'employeur, l'office des poursuites ou la banque, suivie d'un transfert de ces documents à une autre entité comme l'agence immobilière. Étrangement, il est encore aujourd'hui nécessaire que ces documents

transitent par moi, alors qu'il pourrait s'agir de simples requêtes directes entre les bases de données des entités intéressées — modulo un accord que j'aurais pu fournir de manière électronique. L'agence immobilière (voire le propriétaire) aurait pu directement demander à l'employeur, à l'office des poursuites et à la banque d'accéder à l'information qui me concernait.

Conscient de cela, je ne peux que me plaindre de l'inefficacité des procédures actuelles, mais aussi du coût que celle-ci engendre. Ainsi, l'impression d'un document administratif certifiant que j'avais déclenché une procédure de permis de travail me coûta la bagatelle de 20 francs suisses (environ 18 euros), alors même que le coût d'accès à la base de données, qui permettrait de générer un certificat plus fiable, ne dépasse pas le centime.

Enfin, l'élaboration et la signature du contrat requièrent aujourd'hui encore le support papier. J'ai dû remplir une pile de formulaires, les uns après les autres, et signer tout un tas de documents dont je n'ai lu que les grandes lignes. Que de temps perdu ! Pourquoi est-il encore nécessaire d'écrire et de réécrire les noms, prénoms et dates de naissance dans toute une flopée de documents divers et variés ? L'ironie de cela étant que, pendant ce temps, je travaillais sur des vidéos ZettaBytes avec l'EPFL portant sur la signature électronique¹ et la *blockchain*², des solutions modernes pour l'élaboration, la signature et la gestion des données et des contrats.

Le jour où les systèmes de gestion des données seront entièrement informatisés — ce jour approche ! — toutes ces procédures pourront être déclenchées en quelques clics. De nombreuses industries ont d'ailleurs déjà vécu ce jour. L'achat d'une musique, la lecture de livres et le visionnage d'une vidéo sont déjà passés au numérique. Les entreprises à l'origine de la numérisation des services autour de ces industries ont fleuri, au point de devenir des géants du web, qu'ils s'appellent Apple, Amazon ou Netflix. Ces entreprises, comme d'autres de la Silicon Valley, notamment Google, Facebook et Twitter, ont su voir avant tout le monde la valeur des données. Elles investissent désormais des millions, voire des milliards, dans la collection, la gestion et l'analyse de ces données. C'est l'émergence de leurs modèles d'affaire qui a déclenché la vague de folie appelée *Big Data*.

Le déluge de données

Cette vague de folie nous a fait entrer dans une phase intrigante du monde de l'information, puisque la quantité de données produites est en train d'excéder notre capacité à l'analyser, voire à la stocker. Le cas plus parlant est celui du LHC, ce grand collisionneur de particules au CERN. Cette immense structure souterraine, qui fait des dizaines de kilomètres de large, produit des milliards

¹  Bitcoin | ZettaBytes | R. Guerraoui et J. Hamza (2017)

²  The Blockchain | ZettaBytes | R. Guerraoui et J. Hamza (2017)

de collisions entre protons à chaque seconde. La quantité d'information ainsi produite est tellement gigantesque, que la majorité de cette information est immédiatement jetée. Des filtres initiaux permettent ainsi de ne sélectionner que les données qui semblent avoir un intérêt pour les sciences physiques. Mais malgré ce tri sélectif drastique, les données stockées représentent encore des pétaoctets, et remplissent des salles entières de machines à calculer.

Ce que le CERN doit se contenter de faire pourrait être l'avenir de toutes les entreprises confrontées au *Big Data*. En effet, la quantité de *data* croît plus vite encore que les espaces de stockage ! Pour l'instant, il reste possible d'effectuer plusieurs sauvegardes de nos données. Mais ce ne sera pas le cas longtemps. Bientôt, il faudra nécessairement jeter une grosse partie, voire l'écrasante majorité des données collectées. À l'heure où, de plus, on constate la prolifération des capteurs en tout genre et l'émergence de l'*Internet of Things*, la question du choix de la sauvegarde des données s'apprête à devenir une question inédite dans l'histoire de l'information.

En plus de poser de sérieux problèmes de stockage, le *Big Data* pose également de gros problèmes de temps de calcul. Imaginez-vous en train de chercher une donnée parmi des millions de milliards. Même au rythme d'un milliard de données traitées par seconde — la vitesse des microprocesseurs de vos ordinateurs — il faudrait plusieurs jours pour y arriver ! Pour contourner ces problèmes de stockage et de réactivité des calculateurs, de nombreux *data scientists* préfèrent d'ores-et-déjà imaginer comment résoudre des problèmes sans le stockage des données brutes.

Le problème des toilettes

Imaginez-vous à un festival. Vous devez aller aux toilettes. 300 toilettes sont alignées dans une très grande allée. Sauf que ces toilettes sont toutes horriblement sales. Vous cherchez alors à utiliser la meilleure de toutes les toilettes. Cependant, il y a une file derrière vous, de sorte qu'une fois que vous avez fermé la porte d'une toilette et avancé pour tester la toilette suivante, celle que vous venez de visiter ne vous sera plus accessible ; quelqu'un d'autre y aura accédé. Autrement dit, vous devez décider une fois et pour toutes si vous allez utiliser la toilette au moment où vous la visitez. Comment faire pour maximiser vos chances d'avoir utilisé la toilette la plus propre ?

Ce problème est devenu un classique des mathématiques³. Introduit par Martin Gardner en 1960, il est connu sous divers noms, notamment le problème des secrétaires, le jeu du googol, le problème des fiancées ou le problème de la dot du sultan. Toutes ces formulations sont équivalentes et reposent sur le dilemme suivant : vous avez des données qui arrivent séquentiellement, et vous devez

³  *Mathematical Way to Choose a Toilet* | Numberphile | R. Symonds (2014)

arrêter une décision avant que toutes celles-ci soient arrivées, car les opportunités du passé disparaissent avec le temps.

Si ce problème est devenu aussi connu, c'est aussi et surtout parce qu'il dispose d'une jolie réponse très contre-intuitive. En effet, en suivant la stratégie optimale, la probabilité de trouver la meilleure toilette est d'environ⁴ 37 %. Cette stratégie optimale consiste à justement visiter environ 37 % des toilettes que l'on rejettéra toutes, puis à choisir la première toilette visitée meilleure que la meilleure toilette visitée jusque-là. Le plus surprenant, c'est que cette stratégie simpliste garantit toujours une probabilité de 37 % de trouver la meilleure toilette, y compris lorsque l'on considère des milliers, des millions ou des googolplex de toilettes⁵ !

En particulier, ce qui est remarquable dans cette résolution du problème, c'est que l'algorithme que l'on aura utilisé n'aura presque rien à garder en mémoire. La seule donnée qu'il devra mémoriser, c'est la propreté de la meilleure toilette qu'il a visitée. Il peut complètement oublier toutes les autres toilettes visitées.

Ceci étant dit, je déconseille fortement l'utilisation de cette stratégie en pratique. En effet, elle maximise la probabilité de trouver la meilleure toilette ; mais ne dit rien des cas où elle échoue. Du coup, il arrive souvent que vous vous retrouviez à rejeter toutes les toilettes et deviez donc accepter la toute dernière. Ce cas potentiellement catastrophique a même 37 % de chances de survenir !

Traiter rapidement un déluge de données

Le problème des toilettes a inspiré un très grand nombre de variantes qui modélisent un très grand nombre de problèmes, surtout depuis l'avénement du web. Chaque variante a ensuite donné lieu à des algorithmes de résolution différents. Néanmoins, de façon intrigante, les solutions développées par ces algorithmes semblent être des principes généraux pertinents à la prise de décision lors d'un déluge de données, et étant donné une mémoire limitée.

L'une des variantes insiste davantage sur le dilemme entre accepter une opportunité présente ou la laisser passer en attendant mieux, notamment lorsque ces opportunités sont nombreuses, qu'elles possèdent des caractéristiques diverses et variées et que plusieurs peuvent être acceptées à la fois. L'étude de ces problèmes est souvent appelée optimisation en temps réel, ou *online optimization*. Les applications concernent notamment de nombreux problèmes d'allocation de ressources limitées, de la vente des billets de concert à celle des publicités sur Internet. Parmi les solutions proposées, on trouve souvent la quantification de

⁴En fait, elle est égale à $1/e$, où e est la constante d'Euler.

⁵Pour un petit nombre de toilettes, le temps d'arrêt optimal est inférieur, et la probabilité de trouver la meilleure toilette supérieure. Pour 2 ou 3 toilettes, la probabilité de la meilleure stratégie est 1/2. Entre 4 et 10 toilettes, cette probabilité décroît de 46 % à 40 %. Pour 26 toilettes, elle tombe à 38 %. Pour 150 toilettes, elle tombe à 37 %, mais demeure ensuite toujours au-dessus de 36,78 % !

la valeur des ressources à partir des données passées, que ce soit via l'étude des contraintes du problème⁶ ou via des variantes de l'algorithme quasi-bayésien par *multiplicative weights update*.

Une autre variante du problème adresse davantage l'incertitude que l'on a sur la description probabiliste des données du problème. Autrement dit, ces problèmes jouent non pas sur l'incertitude des préjugés, mais sur l'incertitude *sur* nos préjugés. En particulier, ceci donne lieu au dilemme exploitation versus exploration. On parle aussi d'apprentissage en temps réel, ou *online learning*. L'exploration consiste à effectuer davantage de tests potentiellement coûteux pour collecter davantage de données et affiner nos crédences bayésiennes, tandis que l'exploitation consiste à prendre des décisions optimales, étant donné nos crédences. L'algorithme bayésien par échantillonnage de Thompson permet notamment de proposer une solution à ce dilemme. Les applications incluent alors les décisions de poursuite ou d'arrêt prématuré de tests médicaux, ou celles des tests de nouveaux produits sur le web (appelés tests A/B).

Enfin, une troisième variante du problème insiste davantage sur les limitations en temps de calcul — mais s'autorise à réétudier des données passées. Imaginons que vous rentrez de vacances, et qu'il vous faut sélectionner un top 10 des photos de vacances parmi les 2 000 que vous avez prises. Une difficulté s'ajoute alors au problème : il faut maintenant faire attention à éviter les doublons. Une variante de l'algorithme des toilettes qui adresse ce problème et ses variantes est l'algorithme *glouton*, appelé *greedy* en anglais. Cet algorithme va d'abord ne retenir que la meilleure photo qu'il voit, eu égard à ses relations synergétiques avec d'autres photos. Puis il va retenir la meilleure photo compatible avec la photo retenue, puis la meilleure compatible avec les deux photos retenues, et ainsi de suite. Malheureusement, cette approche est en général sous-optimale, car elle n'anticipe pas la synergie d'une photo retenue avec les photos futures. On dit que l'algorithme glouton est myope. Néanmoins, il a été démontré que cette myopie, surtout lorsqu'on la combine à une randomisation⁷, permet de garantir que l'approche gloutonne sera là une bonne heuristique.

Depuis quelques années, les algorithmes s'appuyant sur la quantification des valeurs des ressources, la gestion adéquate de l'incertitude sur les connaissances ou l'approche gloutonne ont gagné beaucoup d'intérêt, notamment chez les géants du web, dont les groupes de recherche se sont activement lancés sur l'étude théorique des propriétés de ces algorithmes. L'avènement du *Big Data* ne fera sans doute qu'accélérer cette tendance.

⁶Ceci correspond aux variables dites *duales* et aux multiplicateurs de Lagrange.

⁷Qui peut consister par exemple à tirer au hasard une photo parmi les 10 meilleures photos qui soient compatibles avec celles déjà retenues.

Le filtre de Kalman

Cependant, les variantes autour du problème des toilettes supposent généralement que le futur sera semblable au passé⁸. Voilà qui explique pourquoi il n'est souvent pas nécessaire de disposer de beaucoup de mémoire pour les résoudre. Pour s'adresser à des contextes dynamiques, l'apprentissage doit anticiper le changement. Le modèle classique pour réussir cette prouesse est le *filtre de Kalman*, du nom du génie Rudolf Kalman, et ses généralisations comme les *Hidden Markov Models*.

Imaginez que votre voiture cherche à connaître sa position et sa vitesse. En bonne bayésienne, elle commence par prendre le soin de décrire l'étendue de son ignorance. Elle peut modéliser cette ignorance à l'aide d'estimations moyennes plus des erreurs. En vertu notamment du théorème centrale limite, Kalman supposa que ces erreurs étaient distribuées selon une loi gaussienne. Et cette hypothèse fut d'une très grande utilité, notamment parce la somme de variables gaussiennes est gaussienne, et parce que le produit de densités gaussiennes est une densité gaussienne. Mais ne vous inquiétez pas, je vais vous épargner les détails des calculs.

Ce qui est malheureux à observer, c'est qu'à chaque pas de temps, de nouvelles incertitudes s'ajoutent. La voiture a peut-être accéléré. Et toute mesure de cette accélération sera accompagnée d'erreur. Cette erreur s'ajoutera alors à l'incertitude que l'on avait déjà. Mais alors, plus le temps passe, plus les erreurs s'ajoutent, et moins on peut être sûr de la position et de la vitesse de la voiture.

Pour réduire ces incertitudes, on peut alors profiter de mesures fournies par d'autres appareils de mesure. Cependant, ces mesures sont également incertaines. Néanmoins, en combinant les positions et les vitesses déduites de l'instant passé avec les différentes mesures des différents appareils⁹, le filtre de Kalman nous permet de déduire la loi gaussienne qui décrit la position et la vitesse de la voiture *a posteriori*, c'est-à-dire étant donné les données des appareils de mesure. Grâce à ses données additionnelles, l'*a posteriori* sera alors plus précis.

Ce filtre de Kalman est aujourd'hui utilisé dans un très grand nombre de domaines. On le retrouve bien sûr dans de nombreux problèmes de navigation ou de contrôle de trajectoire, mais aussi en traitement du signal, en économétrie, en estimation de charge des batteries, en interfaces informatiques, dans les détecteurs de particules, en vision par ordinateur, en tomographie, en sismologie, en suivi médical et en prédictions météorologiques. Et de façon cruciale, il ne s'agit alors de rien d'autre que d'une formule de Bayes, appliquée sous plusieurs hypothèses utiles, comme l'incertitude gaussienne, la linéarité entre les variables et la structure qui relie les différentes variables du problème.

⁸Les données sont typiquement supposées indépendantes et identiquement distribuées, conformément à la tradition fréquentiste !

⁹Et pourvu que ces mesures soient des fonctions affines de la position et de la vitesse de la voiture.

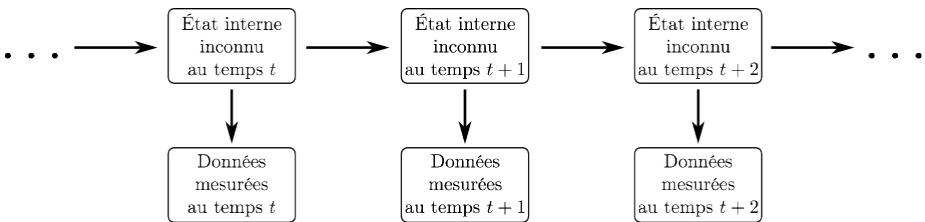


Figure 16.1. Le filtre de Kalman est un cas particulier des *Hidden Markov Models*. Ces modèles décrivent une évolution causale d'un état interne caché. À chaque instant, cet état interne cause néanmoins des données observables.

Cette structure est appelée *Hidden Markov Model*. Elle possède des variables dites *cachées*, qui prennent des valeurs inconnues à chaque instant. Dans le cas de la voiture, ces variables cachées sont la position et la vitesse de la voiture. À chaque instant, de plus, on suppose que des variables observables se déduisent des variables cachées. Dans le cas de la voiture, il s'agit des données des appareils de mesure. Intuitivement, la position et la vitesse de la voiture *causent* ces données mesurées. La formule de Bayes permet alors de déterminer les variables cachées probables à un instant donné, étant données les variables cachées probables de l'instant précédent et les données mesurées par l'appareil à l'instant présent.

À l'instar de nombreux autres modèles graphiques à variables cachées dont on parlera plus longuement dans le prochain chapitre, le *Hidden Markov Model* possède un très grand nombre d'applications ; celles-ci incluant bien entendu les applications du filtre de Kalman qu'il généralise. De façon cruciale surtout, ces modèles permettent de réagir en temps réel à des données massives, sans jamais requérir plus de mémoire. En effet, seule la description de la loi de probabilité de la variable cachée doit être conservée en mémoire. Or, notamment dans le cadre des lois gaussiennes du filtre de Kalman, ou dans un cas où le nombre de valeurs possibles des variables cachées est petit, la description de la loi de probabilité de la variable cachée demeure succincte. Voilà qui fait des *Hidden Markov Model* des modèles prometteurs pour affronter le déluge du *Big Data*.

Nos cerveaux confrontés au *Big Data*

Tous ces algorithmes peuvent vous paraître très distants de votre quotidien. Vous vous dites peut-être que vous êtes capables, vous, de garder en mémoire les informations auxquelles vous êtes exposé. Ce serait là une grave erreur.

En particulier, ce dont on ne se rend pas toujours compte, c'est de la quantité monstrueuse d'informations qui inondent notre cortex cérébral à tout instant, à l'instar des données du CERN. Nos yeux, nos oreilles, notre nez, notre toucher, notre thermoception et nos innombrables autres sens nous envoient près d'un

gigaoctet d'informations à chaque seconde¹⁰. Si l'on cumule toutes les données collectées au cours de décennies, on voit que la quantité totale de données auxquelles nos sens nous ont exposé se compte en exaoctets, soit des millions de téraoctets ! C'est juste monstrueux.

À l'instar des technologies de l'information modernes, notre cerveau ne peut pas stocker toutes les données auxquelles il est exposé. Il ne veut pas tout stocker ; la plupart de ces données étant sans intérêt. Il lui faut oublier la quasi-totalité de ces informations.

C'est typiquement le cas du traitement des données visuelles. Ainsi, selon le neurologue Marcus Raichle¹¹, à chaque seconde et avec une consommation énergétique minime, notre cortex visuel disposerait de l'impressionnante faculté de transformer le gigaoctet de données visuelles en quelques kilooctets d'information pertinente¹².

Plus généralement, le cerveau va davantage retenir les « grandes idées », plutôt que les détails de ces informations. Typiquement, je parie que vous êtes incapable de réciter l'une des phrases des 15 premiers chapitres de ce livre. Pourtant, vous avez retenu, je l'espère, que ces chapitres parlent de la formule de Bayes, de ses fondements logiques, de son application au problème de l'induction, de son histoire, du démon de Solomonoff, de son utilisation en théorie des jeux ou encore de ses liens avec la théorie de l'évolution. Vous avez retenu, je l'espère, des représentations abstraites de divers concepts fondamentaux autour de ce que vous avez lu, même si vous n'avez sans doute retenu presque aucune des phrases que vous avez lues.

Loin d'être une faiblesse, c'est cette capacité à ne retenir qu'une représentation compressée de ce à quoi nos sens nous exposent qui fait la force de nos cerveaux. Nous sommes généralement capables de nous souvenir de choses qui nous importent, et de complètement oublier des choses qui nous sont indifférentes.

Effacer les souvenirs traumatisques

Enfin, ce n'est pas tout à fait vrai. Il nous arrive malheureusement d'oublier des choses dont nous aimerais nous souvenir, ou de nous souvenir de choses que nous aimerais oublier. Les vétérans rentrés de guerre et exposés à des images atroces subissent ainsi souvent des traumatismes psychologiques qui ne cessent de les hanter. À une autre échelle, chacun d'entre nous a ses propres petites et grosses phobies. Et nous souhaiterions souvent simplement les oublier.

¹⁰  *How much bandwidth does each human sense consume relatively speaking?* Quora | R. Rapplean (2012)

¹¹  *Two views of brain function* | Trends in Cognitive Sciences | M. Raichle (2010)

¹² Raichle estime qu'à chaque seconde, environ 10^{10} bits sont collectées par nos rétines, mais seulement 10^4 bits atteignent la couche IV de la région V1 du cortex visuel.

Et si oublier nos phobies était possible ? De façon stupéfiante, la psychologue Merel Kindt a montré que oui ! Le documentaire *Memory Hackers* de NOVA PBS de 2016 montre ainsi Kindt en train de soigner l'arachnophobie de l'un de ses nombreux patients. Pour ce faire, elle le force à regarder des tarentules, pour activer la peur chez ses patients. À ce moment, elle en profite pour donner à ses patients du propranolol. Le propranolol est une molécule qui s'intercale entre les neurones et interrompt la communication entre ces neurones. Selon Kindt, ceci perturbe le souvenir de la peur, au point de le faire disparaître. Et ça marche ! Le lendemain, le sujet se met à caresser la tarentule comme s'il s'agissait d'un hamster inoffensif.

La recherche dans ce domaine n'en est encore qu'à ses premiers pas. Mais les travaux de Kindt et de d'autres chercheurs pourraient permettre d'offrir de nouvelles solutions pour traiter les cas d'addictions à des drogues, voire les troubles de stress post-traumatiques. Ils mettent aussi le doigt sur un fait contre-intuitif désormais bien établi : les souvenirs sont stockés dans les connexions synaptiques.

Contrairement à ce que les neurologues du début du XX^e siècle ont pu penser, la mémoire à long terme ne correspond pas à des bits d'information stockés à l'intérieur de neurones dédiés. Quand on pense à un souvenir, des torrents d'activations neuronales déferlent dans notre cortex cérébral. C'est dans le sillon de ces torrents qu'est gravée notre mémoire. Cette mémoire réside dans la manière dont les neurones sont connectés, pas dans l'état physique des neurones eux-mêmes. En particulier, loin d'être localisée à un endroit précis du cerveau, l'information associée à un souvenir est diffuse à travers la connectivité du réseau neuronal qui constitue notre cerveau.

Dès lors, la relation entre la mémoire active du cerveau et sa mémoire à long terme est similaire à la relation entre votre navigateur web et le web. Pour accéder à la page web qui vous intéresse, il vous faut trouver le lien dit URL de votre page web. Une fois que ce lien est connu, le navigateur va aisément explorer Internet pour récupérer les informations nécessaires à l'affichage de la page à laquelle vous vouliez accéder. Cependant, si vous avez perdu ce lien, le retrouver peut être d'une difficulté atroce — tout comme se remémorer un souvenir, que vous savez que vous avez mais que vous n'arrivez pas à retrouver, peut être une expérience frustrante.

Un corollaire de cette observation est le fait que se remémorer un souvenir réactive le torrent d'activations neuronales qui lui est associé, ce qui permet alors de mieux graver le souvenir. Se souvenir, c'est aider à se souvenir. Cependant, d'une remémoration à l'autre, à l'instar d'un cours d'eau, le torrent d'activations neuronales varie légèrement. Pire encore, le propranolol peut grandement affecter la trajectoire de ce torrent, par exemple en le dissociant de la partie du cerveau associée à la peur. Il semblerait que ce soit cette déviation forcée du torrent d'activations neuronales qui cause une modification du souvenir, et la disparition d'une phobie.

Les faux souvenirs

Mais si un souvenir peut être modifié pour le mieux, il peut aussi l'être pour le pire. C'est ce qu'ont découvert, encore et encore, les chercheurs en psychologie comme Julia Shaw. Shaw en particulier a mis en place une expérience redoutable qui va sans doute grandement questionner le fonctionnement de notre système judiciaire. Cette expérience montre à quel point il est aisé de faire croire à des sujets qu'ils ont commis un délit qu'ils n'ont pas commis !

Pour commencer, Shaw annonce à ses sujets que le but de l'expérience est d'étudier les souvenirs d'enfance. Shaw leur raconte alors une histoire inventée, et affirme que l'histoire a été rapportée par les parents des sujets. L'histoire est celle d'un délit commis par les sujets. Elle n'est toutefois pas absurde. En particulier, elle reprend des éléments du passé du sujet, comme le lieu ou les noms des personnes impliquées dans l'histoire. Les sujets commencent par rétorquer qu'ils n'ont pas de tels souvenirs. Cependant, Shaw leur demande alors de se relaxer et fait ensuite subtilement appel à leur imagination. Shaw rajoute que les autres sujets ont su se rappeler leurs souvenirs d'enfance ainsi, histoire d'encourager les sujets dans cet effort. Ce faisant, Shaw active les neurones des sujets associés aux souvenirs d'enfance ; le mécanisme utile à la consolidation (et à la déformation) des souvenirs a été déclenché ! Shaw demande alors à ses sujets de continuer à y réfléchir, mais de n'en parler à personne.

Une semaine plus tard, Shaw retrouve ses sujets. Les sujets commencent alors à raconter une histoire qui leur semble plausible, avec des mots hésitants. Deux semaines plus tard, les mots hésitants disparaissent et les sujets sont plus sûrs d'eux-mêmes. De façon stupéfiante, dans 70 % des cas, les sujets de Shaw furent eux-mêmes persuadés d'avoir commis un délit ! L'expérience de Shaw fut tellement bouleversante pour les sujets qu'elle dut être arrêtée prématurément.

Les expériences de Kindt et Shaw montrent que l'on ne peut pas se fier à ses souvenirs. Nos souvenirs sont imprécis, flous, mais surtout réajustés, retravaillés et réadaptés à chaque fois que nous y pensons. Le pire, c'est que nous en sommes largement inconscients ! Or, c'est sur les souvenirs qu'ont les juges et les jurés des faits rapportés par les souvenirs des témoins et des accusés que repose notre système judiciaire. Les souvenirs des juges, jurés, témoins et accusés sont ensuite réajustés, retravaillés et réadaptés par les discours persuasifs et émotionnellement prenants des différents avocats. Quelle crédence peut-on réellement attacher à de tels souvenirs ? Sans doute beaucoup moins que la crédence qui leur est accordée en pratique. Ainsi, en s'appuyant sur l'étude de centaines de cas, la psychologue Élizabeth Loftis a montré que dans trois quarts des cas, les personnes disculpées par des tests ADN après avoir été jugées coupables ont été condamnées à cause de témoignages visuels erronés¹³.

On ne peut pas faire confiance aux témoignages visuels, y compris lorsque les témoins semblent sûrs d'eux. En fait, il est même incroyablement facile d'amener

¹³  *Les faux souvenirs* | e-penser | B. Benamran (2016)

des témoins à être convaincus d'avoir vu quelque chose de très erroné. En 1999, dans une expérience devenue célèbre, Daniel Simons et Christopher Chabris proposèrent à des sujets de compter le nombre de passes effectuées par un ensemble de joueurs de basketball. Leur vidéo est disponible sur YouTube¹⁴ et je vous invite fortement à aller y jeter un œil avant de poursuivre la lecture.

À la question du nombre de passes effectuées, les sujets donnèrent souvent la bonne réponse. Puis, Simons et Chabris demandèrent aux sujets s'ils avaient vu le gorille traverser le terrain de basketball. Les sujets répondirent non. Pire, nombreux se dirent convaincus qu'aucun gorille n'avait traversé le terrain. *Ils l'auraient vu*, dirent-ils. Pourtant, le ralenti est formel. Un gorille a bel et bien tranquillement traversé le terrain de jeu, et a même pris le temps d'effectuer quelques pas de danse. Les sujets, occupés par une tâche cognitive exigeante, ont été atteints d'une *cécité d'inattention*. Pire encore, ils sont inconscients de leur inattention. Dans cette expérience comme dans de nombreuses autres¹⁵, ils sont en flagrant excès de confiance vis-à-vis de leurs capacités de perception.

D'autres biais cognitifs empirent notre cas. On l'a vu, le psychologue Jonathan Haidt affirme que nous cherchons constamment à justifier nos intuitions premières à l'aide de notre raison, ce qui signifie que nous ajustons volontiers nos souvenirs. Pire encore, Derek Muller a montré qu'exposer des étudiants à des vidéos qui ne font qu'expliquer rigoureusement un phénomène scientifique contre-intuitif a tendance à augmenter la confiance que ces étudiants ont en leurs croyances non-scientifiques erronées. Voilà qui explique pourquoi la communauté scientifique a fini par rejeter tout témoignage et toute expérience personnelle, y compris lorsque ces témoignages sont racontés avec grande conviction.

Ces défauts majeurs de nos mémoires montrent les limites de nos cerveaux, et nous invitent à fortement diminuer la crédence que l'on a en nos souvenirs, et en les convictions qui se fondent sur ces souvenirs. Nous vivons dans un monde rempli d'incertitudes. Ces incertitudes portent sur le futur, bien sûr, mais également sur le passé — et même le présent. Voilà qui aura conduit Descartes à douter méthodiquement de tout ce dont il pouvait douter, avant de conclure qu'une seule chose était indubitable : le fait qu'il était en train de penser. *Cogito ergo sum*, dit-il.

Cependant, prise à son extrême, cette approche radicale conduit à douter aussi des sciences et de consensus pourtant bien établis par la communauté scientifique comme l'efficacité des vaccins ou l'origine anthropique du réchauffement climatique. Pour la *pure bayésienne*, le sophisme des sceptiques extrêmes est leur quête de vérités indubitables. « Tous les modèles sont faux. » Le savoir ne consiste donc pas à caractériser des faits ou des théories indubitablement vraies. Il s'agit davantage du calcul des degrés de crédence en différents faits, théories et souvenirs. Savoir, c'est déterminer les niveaux d'incertitude adéquats. Dès

¹⁴  Selective attention test | D. Simons (2010)

¹⁵  Test Your Awareness : Whodunnit? | dothetest (2008)

lors, le bon langage pour adresser cette incertitude n'est pas celui de la logique formelle, du vrai et du faux ; c'est celui des probabilités — et l'outil incontournable pour raisonner avec ces probabilités est la formule de Bayes.

Bayes au secours de la mémoire

On a vu que l'une des prouesses du cerveau humain était sa faculté à compresser une très grande quantité de données brutes en une poignée d'idées. S'inspirant de cela, les chercheurs en intelligences artificielles ont inventé l'architecture de l'auto-encodeur.

Le rôle d'un auto-encodeur est de condenser de grandes quantités d'information, comme une image haute définition ou tout un film, pour n'en retenir que la substantifique moelle. Autrement dit, ces réseaux de neurones cherchent à faire exactement ce que l'on vous a demandé de faire en cours de français : effectuer un résumé. Et pour tester la qualité du résumé, on demande aussi à l'auto-encodeur de décompresser son résumé, et d'imaginer quelle image haute définition ou quelle autre donnée brute est associée à ce résumé.

L'une des clés pour ce faire est une approche bayésienne. Reconstituer les données consiste ainsi à déterminer quelles données ont pu conduire à ce résumé. Voilà le cas typique d'application de la formule de Bayes ! Il nous faut déterminer les causes probables de ce résumé. On a ainsi

$$\mathbb{P}[\text{data}|\text{resume}] = \frac{\mathbb{P}[\text{resume}|\text{data}]\mathbb{P}[\text{data}]}{\mathbb{P}[\text{resume}]}.$$

On verra d'ailleurs dans le prochain chapitre comment, en s'appuyant sur cette formule et en introduisant des résumés inhabituels, les chercheurs en intelligence artificielle ont su doter leurs machines de pouvoirs créatifs.

Mais plus encore que de permettre de décoder de la mémoire encodée, la formule de Bayes peut permettre de déterminer un codage adéquat des souvenirs. Voilà qui permettra de ne retenir que les éléments essentiels et de les compresser en peu d'espace mémoire. En particulier, en gardant uniquement en mémoire les quelques modèles les plus probables *a posteriori*, la mémoire pourra résumer efficacement les données massives. Et ainsi optimiser la gestion de sa mémoire.

Ce que je dis là peut paraître abstrait et réservé aux intelligences artificielles. Pourtant, il s'agit du noeud de la frustration d'un théoricien confronté à une collection exhaustive de faits difficiles à lier — et donc à garder en mémoire. Ainsi Claude Shannon disait-il de la chimie qu'il étudia à l'école qu'elle « [lui] a toujours semblé un peu ennuyeuse ». « Trop de faits isolés et trop peu de principes généraux à mon goût. »

Une approche bayésienne de la connaissance cherchera, au contraire, à trouver des modèles capables de résumer un grand nombre de faits isolés par une

poignée de grands principes. Avec l'espérance que garder en mémoire les quelques principes généraux suffira à inférer, via la formule de Bayes ci-dessus, une grande partie des faits isolés. Voilà qui justifie grandement l'utilité du bayésianisme pour l'étude des sciences du passé, qu'il s'agisse de l'Histoire, de la théorie de l'évolution ou de la cosmologie. À en croire le bayésianisme, le but de ces sciences ne serait pas tant distinguer le vrai du faux. Il s'agirait davantage de la construction de modèles simples, structurés, faciles à retenir, et néanmoins capables d'assez bien expliquer la multitude d'observations réalisées — même si on peut aussi arguer que ces sciences permettent aussi de prédire les vestiges non-observés probables.

Des mémoires à plus ou moins long terme

Les études de Kindt et Julia concernent la mémoire à long terme. Il s'agit d'encoder de l'information dans la structure topologique du réseau de neurones qu'est le cortex cérébral. Le problème, c'est qu'accéder à cette information peut être difficile. Il faut tester différentes propagations de signaux possibles à travers le réseau. C'est ce qu'il nous arrive typiquement lorsque l'on cherche à se remémorer les paroles d'une chanson connue. La première fois qu'on y pense, on butte souvent après quelques mots, comme si le flot de signaux électriques dans notre cerveau était arrêté, ou dévié dans la mauvaise direction. Cependant, à forcer de répéter ce flot, on finit par trouver la bonne voie, et retrouver les paroles de la chanson.

On rencontre d'ailleurs le même problème pour retrouver de l'information dans des grandes bases de données. Ce qui a fait la fortune de Google est d'ailleurs, entre autres, d'avoir su organiser les données du web au préalable pour accélérer la recherche d'information. Mais il ne s'agit pas là que d'une question d'organisation de l'information. Le support de l'information pose ainsi souvent un compromis entre la capacité de stockage et la vitesse d'accès à l'information stockée. Certains supports sont remarquablement rapides, comme la mémoire vive ou, mieux encore, les registres des microprocesseurs, mais d'une capacité limitée. D'autres sont lents, mais d'une capacité énorme, comme les disques durs, le stockage sur bande utilisé au CERN ou encore le stockage sur brins d'ADN¹⁶.

À cela s'ajoute la finitude de la vitesse de la lumière qu'impose la relativité d'Einstein ! Si l'information est stockée à distance, comme c'est de plus en plus le cas avec le *cloud*, il y a alors un délai nécessaire à l'accès à l'information. Rappelons que la vitesse de la lumière est d'environ 10^5 kilomètres par seconde. Accéder depuis Lausanne à une information qui se trouve à l'autre bout de l'Europe, à quelques milliers de kilomètres, prend donc nécessairement au moins $10^3/10^5 \approx 10^{-2}$ secondes, soit quelques dizaines de millisecondes.

¹⁶  All of Humanity's Data in a Backpack | ZettaBytes | C. Dessimoz (2017)

Ceci peut vous paraître rapide. Pourtant, ceci empêche déjà plusieurs allers-retours sans que l'utilisateur s'impatiente. En particulier, si un serveur à New York doit aller demander l'information chez un serveur à Tokyo, lequel doit s'adresser à un serveur à Berlin avant que l'information soit retransmise de Berlin à Tokyo, puis de Tokyo à New York, puis de New York à un utilisateur à Lausanne, alors l'application de l'utilisateur manquera nécessairement de réactivité.

Pour la quasi-totalité des utilisateurs, ce délai n'est qu'un petit inconfort. Mais dans le monde de la finance, et en particulier du *trading* à haute fréquence, de tels délais peuvent faire toute la différence et représenter des millions de dollars. C'est ainsi que de nombreuses entreprises ont découvert pouvoir gagner des millions en communiquant entre New York et Chicago via des micro-ondes plutôt que via la fibre optique. Après tout, via la fibre optique, la vitesse du signal n'est que celle de la lumière dans la fibre, laquelle est légèrement inférieure à sa vitesse dans l'air ou dans le vide. Quelques précieuses millisecondes peuvent ainsi être gagnées en communiquant via des micro-ondes.

La lenteur de l'accès aux données des supports de grandes capacités et les délais de communications inévitables ont conduit les ingénieurs en informatique à utiliser des caches. Un cache est une mémoire rapide proche de l'unité de calcul. Dans votre ordinateur, il s'agit de la mémoire vive, ou mieux encore, des registres L1, L2 et L3, dont la mémoire ne contient que quelques octets, mais dont le temps d'accès se compte en microsecondes, voire en nanosecondes.

Le principe du cache a d'autres applications. Par exemple, quand vous naviguez sur le web avec Mozilla Firefox, Google Chrome, Safari ou autres navigateurs web, votre navigateur va sauvegarder en cache, dans la mémoire de l'ordinateur, les informations que vous téléchargez de manière récurrente. Du coup, ces informations vous seront directement accessibles. Vous n'aurez donc pas à subir les délais de communications que requiert Internet.

De la même manière, certains chercheurs pensent que notre cerveau traite de manière différente la mémoire à court terme et la mémoire à long terme. Alors que la mémoire à long terme est gravée dans la connectivité du réseau neuronal, la mémoire à court terme pourrait davantage être contrôlée par les neurotransmetteurs — même s'il semblerait qu'il s'agisse là d'une hypothèse peu corroborée expérimentalement et qui ne mérite donc pas toutes nos crédences.

Les réseaux de neurones récurrents

Quoi qu'il en soit, la recherche en réseau de neurones artificiels a opté pour une troisième alternative : des boucles de propagation des signaux neuronaux. L'architecture qui inclut de telles boucles est appelée réseau de neurones récurrent. Elle permet ainsi de rendre disponible une partie de l'information du passé immédiat pour traiter les données présentes. Voilà qui est devenu l'état de l'art

pour traiter des données dont la structure est fondamentalement séquentielle, comme le texte d'un livre ou le son d'un discours.

L'astuce des réseaux de neurones récurrents appartient à une famille plus large d'algorithmes avec états internes. Cette famille inclut notamment les *filtres de Kalman* et les *Hidden Markov Models* dont on a parlé précédemment. L'objectif de l'état interne est d'extraire de manière synthétique l'information du passé immédiat, de façon à mieux comprendre la dernière donnée analysée. Cependant, se pose alors la question de la vitesse à laquelle l'état interne évolue, en comparaison avec la lecture des données.

C'est quelque chose que j'ai eu l'occasion d'observer dans mon apprentissage des mathématiques comme dans celui d'élèves que j'ai pu avoir à encadrer. En première lecture d'un document (comme le livre que vous lisez !), il est utile de ne pas trop s'attarder sur les détails pour ne pas perdre le fil de la lecture. C'est parce que trop ralentir la lecture, ou, de façon équivalente, trop modifier l'état interne de notre mémoire à court terme, désynchronise là où le texte veut nous amener de là où notre esprit est allé. C'est ainsi qu'il peut être utile de lire d'abord un texte assez rapidement, puis le relire en s'attardant longuement aux détails, avant d'effectuer une nouvelle lecture rapide. Chaque lecture apportera quelque chose de nouveau, car chaque lecture sera associée à une dynamique différente de l'état interne de notre mémoire à court terme.

Jusque-là, on a parlé de lecture linéaire des données. Mais la manière dont on lit et on écoute n'est pas tout à fait linéaire. En effet, quand une phrase est très alambiquée, on a tendance à la lire plusieurs fois avant de poursuivre la lecture linéaire. De même, il nous arrive de ne comprendre autrui qu'une fois sa phrase finie. C'est particulièrement le cas en sanskrit, en hindi et en japonais où le verbe est en fin de phrase. Dans la même veine, il arrive souvent dans les articles de mathématiques que l'explication des calculs soit donnée après le calcul. Ou encore que ce soit la chute d'une blague, d'une publicité ou d'un film qui permette de donner un sens à la blague, à la publicité ou au film.

Pour profiter à la fois des données passées et futures pour comprendre le présent, une autre architecture neuronale a été proposée. Il s'agit des réseaux de neurones récurrents, dits *bidirectionnels*. Pour pouvoir inclure le futur, ces réseaux nécessitent de surcroît un délai de réponse. Il y a fort à parier que des structures similaires sont à l'œuvre dans le cerveau humain. Ceci expliquerait notre faculté à comprendre les blagues avec un délai qui suscite parfois la moquerie des autres.

Enfin, une avancée récente en intelligence artificielle est l'introduction d'un dispositif neuronal qui nous force à oublier. Il s'agit de l'architecture LSTM, pour *long-short-term memory*. En plus des boucles neuronales usuelles des réseaux de neurones récurrents, les LSTM possèdent une boucle supplémentaire qui, si elle est activée, forcera tout signal des boucles neuronales à disparaître. Voilà qui expliquerait pourquoi, coupés dans une discussion, il peut nous être parfois très difficile de retrouver le sujet de la discussion.

De nos jours, les LSTM et leurs variantes semblent être l'état de l'art de l'intelligence artificielle pour le traitement de données à nature très séquentielle, à l'image de la reconnaissance vocale ou de l'analyse du langage naturel. Mais je reconnaiss volontiers mon manque de compréhension de ce succès et mon manque d'expertise quant à prédire l'avenir de l'état de l'art dans ces domaines.

Que faut-il apprendre et enseigner ?

À l'instar de la mémoire humaine, la mémoire des réseaux de neurones artificiels, qu'elle soit encodée dans la connectivité du réseau neuronal ou dans les signaux qui se propagent dans des boucles de ce réseau, n'a aucune garantie de fiabilité. Les ordinateurs qui nous entourent nous dominent largement pour au moins deux tâches algorithmiques : la vitesse de calculs et le stockage fiable de l'information. Il y a fort à parier que l'intelligence artificielle de demain saura tirer parti de ces capacités littéralement surhumaines, sans doute en combinant certains aspects des réseaux de neurones avec d'autres algorithmes beaucoup plus fiables et optimisés pour des tâches plus restreintes.

Les nouvelles technologies ont sans doute déjà modifié nos cortex cérébraux. Notre addiction à nos téléphones, à Google et à Wikipedia semble affecter la gestion de notre mémoire. Ce n'est pas nécessairement une mauvaise chose. Les générations précédentes n'hésitaient pas à glorifier ceux capables de réciter des vers de Corneille ou de Baudelaire, de donner la date du couronnement de Napoléon ou d'énoncer par cœur les équations de Maxwell. Cependant, de nombreux enseignants se plaignent de l'excès de choses à savoir pour réussir à l'école, au détriment d'une réelle compréhension des concepts sous-jacents. Certains arguent que, dans le monde moderne, le savoir-faire est beaucoup plus important que le savoir. Ainsi, selon cette logique, il serait bien mieux de savoir comment trouver ou retrouver l'information que de simplement la savoir.

Personnellement, je pense qu'il y a à la fois beaucoup trop de savoir et de savoir-faire à apprendre à l'école. Pire, ce savoir est souvent enseigné comme étant une vérité absolue, et le savoir-faire comme une solution *sine qua non*. Or, pour la *pure bayésienne* comme pour le *bayésien pragmatique*, « tous les modèles sont faux ». Mais ce n'est pas là le plus problématique à mon sens.

L'excès de savoir et savoir-faire se fait, il me semble, au détriment de la compréhension des notions et des modèles les plus utiles et crédibles pour comprendre le monde qui nous entoure. Cet enseignement ignore trop souvent les raisons de la crédibilité de nos modèles et les limites de leur applicabilité. Je serais d'avis d'enseigner beaucoup moins, et de se restreindre à ce qui est important, contre-intuitif et instructif. Je crois typiquement que les biais cognitifs, les processus clés de la théorie de l'évolution, l'informatique théorique et l'utilitarisme moral devraient être enseignés, au détriment par exemple de la trigonométrie ou de la mécanique quantique.

Par ailleurs, la formule de Bayes semble nous inviter à davantage apprendre par l'exemple, par opposition à la mémorisation de théories — on verra en effet dans les chapitres suivants que nos cerveaux semblent relativement bayésiens et qu'ils sont très prompts à généraliser à partir d'exemples. En particulier, il semble que la pertinence des théories n'advient qu'une fois que suffisamment de données sont connues et « aisément accessibles » pour que les termes d'expérience de pensée de la formule de Bayes soient facilement estimables. Du coup, il semble souhaitable de n'apprendre ces théories qu'après avoir étudié plusieurs exemples où l'*utilité* de ces théories sautera aux yeux. Voilà qui nous invite par exemple à d'abord introduire les mathématiques sous forme de jeux, d'énigmes ou de paradoxes logiques qui attirent l'attention des élèves, avant de leur expliquer qu'il s'agit là de cas d'application de théories plus générales — c'est ce que j'ai essayé de faire tant bien que mal à travers ce livre.

Mais, à mon humble avis, le plus important à enseigner reste l'épistémologie, et les statistiques sans lesquelles l'épistémologie est inapplicable. Bien entendu, tout bayésien extrémiste que je suis, je pense surtout que la formule de Bayes et ses nombreuses conséquences contre-intuitives devraient former l'un des piliers de l'Éducation.

Il est temps, je pense, d'arrêter de cumuler des connaissances dogmatiquement reconnues vraies, et de commencer à enseigner ce qu'est la connaissance, comment l'acquérir et comment distinguer les théories crédibles de celles qui ne méritent pas nos crédences. Malheureusement, de nos jours, même les grands scientifiques ont une compréhension très incomplète de l'épistémologie ; beaucoup ignorent l'existence même du bayésianisme.

Références en français

- ▶ *Les faux souvenirs* | e-penser | B. Benamran (2016)
- ▶ *L'Histoire du Stockage Numérique - 1ère Partie* | Matière Grise (2017)
- ▶ *L'Histoire du Stockage Numérique - 2ème Partie* | Matière Grise (2017)
- ▶ *Le bitcoin et la blockchain* | Science Étonnante | G. Mitteau et D. Louapre (2016)
- ▶ *Le BITCOIN : Révolution économique ?* Micode et Stupid Economics (2017)

- ▶ *Une justice SANS libre-arbitre ?* Démocratie 24 | Science4All | L.N. Hoang (2017)
- ▶ *La mémoire ne suffit pas* | IA 8 | Science4All | L.N. Hoang (2018)
- ▶ *Le Bitcoin, comment ça marche ?* Crypto | String Theory | L.N. Hoang (2018)
- ▶ *Quel problème résout la Blockchain ?* Crypto | String Theory | L.N. Hoang (2018)
- ▶ *La Blockchain* | Crypto | String Theory | L.N. Hoang (2018)

Références en anglais

- ➲ *The Memory Illusion: Remembering, Forgetting, and the Science of False Memory* | Cornerstone Digital | J. Shaw (2016)
- ➲ *Deep Learning* | MIT Press | I. Goodfellow, Y. Bengio and A. Courville (2016)

- ➲ *Two views of brain function* | Trends in Cognitive Sciences | M. Raichle (2010)

- ▶ *Memory Hackers* | NOVA PBS (2016)
- ▶ *Mathematical Way to Choose a Toilet* | Numberphile | R. Symonds (2014)
- ▶ *When To Quit (According to Math)* | Up and Atom | J. Tan-Holmes (2017)
- ▶ *Selective attention test* | D. Simons (2010)
- ▶ *Bitcoin* | ZettaBytes | R. Guerraoui and J. Hamza (2017)
- ▶ *The Blockchain* | ZettaBytes | R. Guerraoui et J. Hamza (2017)
- ▶ *Arm Up for the Big Data Deluge* | ZettaBytes | A. Ailamaki (2017)
- ▶ *All of Humanity's Data in a Backpack* | ZettaBytes | C. Dessimoz (2017)
- ▶ *Why Blockchain is a Revolution* | ZettaBytes | E.G. Sirer (2018)

- ➲ *How much bandwidth does each human sense consume relatively speaking?*
Quora | R. Rapplean (2012)
- ➲ *The Secretary / Toilet Problem and Online Optimization* | Science4All | L.N. Hoang (2015)

L'art de faire des mathématiques consiste à trouver ce cas particulier qui contient tous les germes de la généralité.

David Hilbert (1862-1943)

Rien dans la vie n'est aussi important que vous le pensez au moment où vous y pensez.

Daniel Kahneman (1934-)

17

La nuit porte conseil

D'où viennent les idées ?

« La pensée n'est qu'un éclair au milieu d'une longue nuit. Mais c'est cet éclair qui est tout », écrit le mathématicien Henri Poincaré. Il illustra cet éclair par le récit de l'une de ses grandes découvertes : « [...] au moment où je mettais le pied sur le marchepied, l'idée me vint, sans que rien dans mes pensées antérieures parût m'y avoir préparé, que les transformations dont j'avais fait usage pour définir les fonctions fuchsiennes étaient identiques à celles de la géométrie non euclidienne. »

Poincaré renchérit avec le récit d'une autre de ses découvertes : « Un jour, en me promenant sur la falaise, l'idée me vint, toujours avec le même caractère de brièveté, de soudaineté et de certitude immédiate, que les transformations arithmétiques des formes quadratiques ternaires indéfinies étaient identiques à celles de la Géométrie non euclidienne. »

Ces récits de Poincaré font écho avec le vécu de nombreux mathématiciens. Dans son livre *Théorème vivant*, Cédric Villani raconte qu'après avoir durement travaillé jusqu'à 3 heures du matin pour compléter un trou béant dans une démonstration de 150 pages et s'être couché désespéré, il se réveilla avec la solution du problème ! Le cerveau de ces mathématiciens semble capable de tourner malgré eux.

À titre personnel, j'ai aussi vécu ce genre d'expériences à maintes reprises — même si mes meilleures idées n'arrivent pas à la cheville de celles de Villani ou

Poincaré ! J'irai même jusqu'à postuler avec grande crédence que la capacité du mathématicien à faire travailler son subconscient en continu est la raison première de son aisance avec ses objets mathématiques préférés. Depuis deux ans, la formule de Bayes ne m'a jamais semblé bien loin. Et c'est souvent sans me prévenir qu'elle s'est mise à me chuchoter ses secrets.

Je ne peux que vous suggérer de suivre les pas de Poincaré. Si vous voulez vraiment progresser en mathématiques, je vous conseille de vous passionner pour cette discipline au point que votre inconscient ne la laisse plus tomber, même au moment où vous irez vous coucher. Dans le langage de la psychologie de Kahneman, que l'on a utilisé au chapitre 14, c'est comme si, ce faisant, un système 3 naissait en nous et éduquait le système 1, sans que le système 2 en soit conscient.

Voilà qui soulève toutefois une question intrigante pour le psychologue. Qu'est-ce que ce système 3 ? Que se passe-t-il dans le cerveau d'un mathématicien pour que celui-ci continue à progresser sans qu'il en soit conscient ? Les rêves aident-ils ? Et de façon plus générale, à quoi rêver peut-il bien servir ? Il s'agit là de questions difficiles. Je ne prétends absolument pas vous en fournir des réponses complètes. Néanmoins, j'aimerais vous présenter l'hypothèse avancée par Francis Crick, prix Nobel de médecine, et Graeme Mitchison. Pourquoi celle-ci ? Parce qu'elle repose sur un argument bayésien élégant....

Mais avant d'en arriver là, je vais m'arrêter sur le processus créatif dont les machines sont désormais capables.

L'art créatif des intelligences artificielles

Depuis peu, les machines savent composer de la musique et peindre des tableaux. Comme on en a brièvement parlé dans le chapitre précédent, la clé de ce processus créatif est encore une fois la formule de Bayes.

En effet, dans de nombreux modèles de *deep learning*, il est possible d'exciter certains neurones dits *profonds*, pour créer une combinaison de concepts abstraits dans le réseau de neurones. Certains de ces réseaux de neurones seront alors capables d'imaginer des données brutes qui auraient pu conduire à l'excitation des neurones excités. Autrement dit, alors que les réseaux de neurones cherchent habituellement à déduire des données les concepts abstraits qui les résument, on peut leur demander de deviner les données probables, sachant les concepts abstraits.

En choisissant ensuite des concepts abstraits habituellement déconnectés, on peut alors amener le réseau de neurones à créer des données à la fois inhabituelles et relativement crédibles. Tel est le processus artistique des machines. Il y a fort à parier que ce processus partage au moins quelques similarités avec le processus créatif de nos cerveaux.

C'est ainsi qu'en 2015, Google publie sur son blog de recherche des images générées par son intelligence artificielle *DeepDream*¹. Ces images ont des airs psychédéliques. On y voit des nuages se transformer en poissons, des arbres en temples, et des feuilles d'arbre en oiseaux. Mieux encore, vous pouvez désormais demander à une autre intelligence artificielle, appelée *DeepArt*, de réinterpréter vos photographies à la manière d'un peintre célèbre, qu'il soit Van Gogh, Picasso ou Kandinsky. Mon image de profil Twitter est ainsi le fruit de quelques secondes de travail gratuit de *DeepArt*.

L'une des étapes importantes des procédés créatifs de ces intelligences artificielles est la capacité du réseau de neurones à trouver des données crédibles, étant donné des résumés abstraits de ces données. On dit que le réseau doit être capable d'échantillonner, conformément à la loi de probabilité $\mathbb{P}[\text{data}|\text{resume}]$ des données conditionnellement aux concepts abstraits excités. Ainsi, selon le bayésianisme, *la créativité se résume à l'échantillonnage de croyances contextuelles*.

Mais surtout, échantillonner, c'est fournir des exemples représentatifs, plutôt que de raisonner sur des distributions de probabilité très complexes. Voilà qui peut être très utile pour illustrer un raisonnement et le rendre plus intelligible. Après tout, l'apprentissage naturel de nous autres humains, y compris en mathématiques, semble bien souvent davantage reposer sur des exemples représentatifs que sur une théorie formelle. Notre appareillage cérébral semble bien plus optimisé pour inférer des règles grossières à partir d'exemples, que pour faire coïncider son réseau neuronal avec une théorie formelle. Voilà qui explique pourquoi échantillonner est indispensable pour les humains. De façon curieuse, ceci semble indispensable pour les machines aussi.

Mais avant d'en arriver là, il nous faut au préalable des modèles adéquats pour représenter de telles relations entre des données et des résumés abstraits des données. Plusieurs architectures de *machine learning* ont été proposées pour décrire des lois de probabilité et permettre leur échantillonnage. Ces architectures tombent dans deux catégories (que des modèles sophistiqués peuvent combiner à escient) : il s'agit des réseaux bayésiens² et des champs de Markov. Arrêtons-nous sur les réseaux bayésiens pour l'instant.

L'allocation de Dirichlet latente (LDA)

Outre les *filtres de Kalman* et les *Hidden Markov Models*, l'un des succès majeurs des réseaux bayésiens est l'allocation de Dirichlet latente (ou LDA en anglais), inventée au début des années 2000. LDA a pour objectif de classer des textes par catégorie. Ainsi, de manière parfaitement automatisée, un ordinateur pourrait

¹  Deep Dream - a code example for visualizing Neural Networks | Google Research Blog
| A. Mordvinstor, C. Olah and M. Tyka (2015)

² aussi appelés *forward models*.

utiliser LDA pour ranger vos emails dans des dossiers intitulés « personnel », « travail », « vacances » et « spams ». Mieux encore, LDA est même capable de détecter des combinaisons de catégories, et ainsi dire d'un document qu'il est moitié-travail moitié-vacances, voire qu'il est $\frac{2}{3}$ -personnel et $\frac{1}{3}$ -travail.

Pour ce faire, LDA exploite la notion fondamentale des réseaux bayésiens, à savoir la notion de causalité. Cette notion de la causalité a le bon goût d'être conforme à notre intuition, ce qui fait des réseaux bayésiens des modèles probabilistes relativement simples à interpréter. C'est en vertu de cette conformité des réseaux bayésiens avec notre intuition, que, comme on en a parlé au chapitre 3, Neil et Berger proposent de les utiliser dans le domaine judiciaire.

Revenons-en à LDA. LDA postule que tout mot d'un document est obtenu par le procédé causal suivant. Premièrement, une certaine combinaison de catégories est tirée au hasard pour le document³. Puis, pour chaque mot du document, une catégorie de la combinaison de catégories est tirée au hasard. Enfin, le mot est tiré au hasard en fonction de la catégorie tirée.



Figure 17.1. LDA est un exemple typique de réseau bayésien. Elle possède une structure causale qui vise à déduire les données observées de concepts abstraits. Le schéma ci-dessus est une représentation simplifiée de LDA.

LDA est simpliste. Elle est aussi bien entendu très erronée. Un document écrit par LDA serait un sac de *buzzwords* sans queue ni tête. Ce n'est clairement pas ainsi que j'ai écrit ce livre ! Mais « tous les modèles sont faux, certains sont utiles ». À défaut d'être *vraie*, LDA est *utile* ! Cette technique est utilisée à travers le web et en bio-informatique pour classer des données dans des catégories. Et si cette technique est si efficace, c'est parce qu'elle peut sans cesse s'améliorer à l'aide d'inférences bayésiennes. Ainsi, à chaque fois qu'un nouveau document est présenté, LDA est capable d'analyser ce document pour s'auto-améliorer.

Mieux encore ! LDA n'a pas besoin qu'on lui dise la catégorie à laquelle on pense que le document en question appartient — même si elle gagnerait légèrement à le savoir. On dit de LDA qu'elle est un algorithme d'apprentissage *non-supervisé* : donnez-lui plein de documents sans lui préciser leur catégorie, elle s'améliorera malgré tout !

Mieux encore ! LDA n'a pas non plus besoin qu'on lui liste *a priori* les catégories que l'on veut prendre en compte, ni même le nombre de ces catégories. En particulier, il y a une variante de LDA, appelée LDA hiérarchique, où le nombre de catégories peut être automatiquement augmenté s'il semble émerger des documents d'un nouveau type. On dit de LDA hiérarchique qu'elle est une méthode non-paramétrique, car sa complexité peut croître indéfiniment.

³Ce tirage se fait justement selon une *loi de Dirichlet*.

Le restaurant chinois au secours de LDA

Pour réussir ce tour de force, LDA hiérarchique, à l'instar de LDA, a une touche profondément bayésienne. Ceci veut dire que, en particulier, il est indispensable de disposer d'un *a priori* concernant la manière dont le nombre de catégories va devoir augmenter en fonction de la quantité de documents. L'*a priori* en question est ce que l'on appelle le *processus du restaurant chinois*.

Imaginons-nous dans un restaurant chinois. À chaque instant, un nouveau client débarque. Le nouveau client va alors assigner un numéro à chacun des $n - 1$ clients déjà présents, et le numéro n à une table non-occupée. Il tire ensuite un nombre au hasard entre 1 et n , et va s'asseoir à la table du client dont il a tiré le numéro (ou s'il est tombé sur n , il va s'asseoir tout seul). Et bien, LDA hiérarchique suppose que chaque nouveau document est, *a priori*, une sorte de nouveau client, et que chaque catégorie est une table du restaurant⁴.

Ce procédé est sans doute pertinent pour nous en pratique. Si l'on cherche à organiser nos documents par catégorie, il est raisonnable de ne considérer inventer une nouvelle catégorie pour le n -ième document qu'avec probabilité $1/n$. Ce faisant, on garantit que le nombre de catégories ne deviendra jamais trop volumineux. En effet, je laisse les matheux parmi vous vérifier que le nombre de catégories sera logarithmique en le nombre de données — en particulier, LDA hiérarchique est donc parfaitement adapté au *Big Data* !

Ce qui est stupéfiant avec LDA hiérarchique, c'est que l'ordinateur qui l'applique invente par là ses propres nouveaux concepts ! D'ailleurs, en pratique, il arrive souvent que les catégories inventées par LDA hiérarchique, bien que pertinentes, ne puissent pas aisément être interprétées par les humains. L'ordinateur aura alors inventé un concept totalement absent de notre vocabulaire.

Selon le *naturalisme poétique* de Sean Carroll qu'on a décrit au chapitre 13, on est alors forcé que ce concept *existe*, car il est bel et bien utile à l'ordinateur. De façon plus pragmatique, ce que LDA montre surtout, c'est que l'utilité de concepts abstraits ne requiert pas d'existence physique de ces concepts. Pour expliquer au mieux des jeux de données en pratique, inventer ces concepts abstraits est une étape inéluctable. On reviendra plus longuement sur cette remarque fondamentale dans le prochain chapitre.

Pour l'heure, notez à quel point il est aisément d'échantillonner des données à partir d'un réseau bayésien et de concepts profonds. En effet, par définition du réseau bayésien, la manière dont des données brutes fictives sont générées correspond à un processus causal très précis (même s'il inclut de la randomisation). Dans le cas de LDA, vous pouvez même aisément exiger de LDA qu'elle génère un document moitié personnel, moitié travail. Certes, LDA n'est pas suffisamment sophistiquée pour produire un texte sensé, mais les mots qui seront générés auront de bonnes chances d'effectivement mêler travail et vie personnelle.

⁴Le fait que ce restaurant soit chinois est un mystère. Mais de façon amusante, il en existe une variante appelée *processus du buffet indien*, qui permet la combinaison de catégories.

Les simulations de Monte-Carlo

Si l'échantillonnage ne semble pas très fructueux dans le cas de LDA, son application dans des cadres différents peut être spectaculaire. Prenez une boîte d'aiguilles et une grande feuille de papier. Tracez des droites horizontales espacées d'une distance égale à 4 fois la longueur d'une aiguille. Lancez un très grand nombre d'aiguilles. Comptez la proportion des aiguilles qui intersectent une ligne horizontale. Cette proportion sera environ l'inverse de la constante fondamentale τ de la géométrie, égale au rapport de la circonférence d'un cercle par son rayon.

C'est ce que l'on appelle l'expérience des aiguilles de Buffon⁵. Elle permet de sonder la nature d'une constante mathématique à l'aide d'expériences — ce qui est très différent des expériences scientifiques qui cherchent à découvrir des propriétés de notre univers ! De façon tout aussi étrange, les aléas des expériences sont alors essentiels à l'expérience.

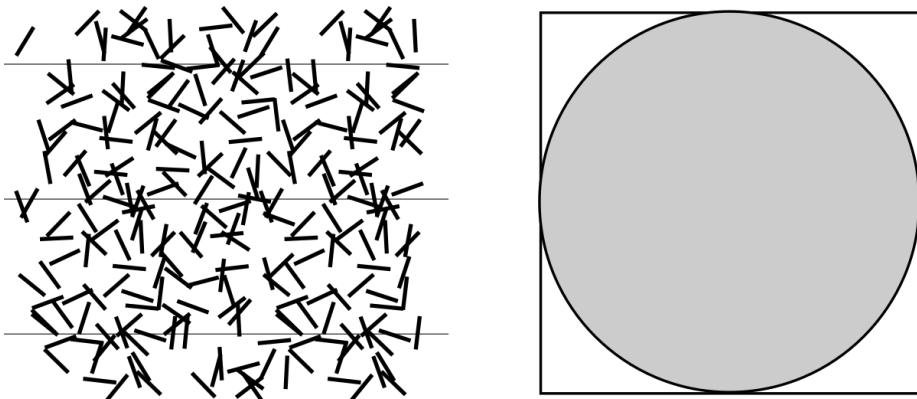


Figure 17.2. Ces deux images illustrent deux méthodes d'approximation de τ par la méthode de Monte Carlo. À gauche, il s'agit de l'aiguille de Buffon. La fraction d'aiguilles qui coupent une ligne horizontale sera environ $1/\tau$. À droite, la fraction de points du carré à l'intérieur du disque sera environ $\tau/8$.

D'autres façons similaires d'estimer τ ont ainsi été proposées. Dessinez un disque de rayon $1/2$, et encadrez ce cercle dans un carré circonscrit de côté 1. L'aire du disque sera alors de $\tau/8$, tandis que celle du carré sera de 1. Jetez ensuite des points (uniformément) aléatoirement sur le carré. On s'attend alors à ce que la proportion de points dans le disque soit environ le rapport de l'aire du disque sur l'aire du carré, soit $\tau/8$. Voilà qui nous donne une autre manière d'estimer expérimentalement τ : il suffit de jeter (uniformément) aléatoirement un grand nombre de points sur le carré, et de multiplier par 8 la proportion de points tombés dans le disque.

⁵ *Pi and Buffon's Matches* | Numberphile | T. Padilla (2012)

Cette expérience a été réalisée par Dianna Cowern et Derek Muller, mieux connus sous les noms de Physics Girl et Veritasium⁶. Cowern et Muller ont lancé un grand nombre de fléchettes sur une cible carrée. Cependant, après un premier jour d'expériences, ils constatèrent que leurs lancers n'étaient pas uniformément distribués. Ils touchaient davantage le centre de la cible que les coins, ce qui les aura conduits à une surestimation du résultat. Le lendemain, ils modifièrent l'expérience en dessinant sur l'arrière de la cible plusieurs disques circonscrits par des carrés. Ils obtinrent enfin l'excellente estimation $\tau \approx 6,28$.

Les expériences de Buffon, Cowern et Muller font en fait partie d'une famille plus générale d'expériences dont la justesse du résultat dépend de la qualité de l'aléatoire. Cette famille est celle des méthodes de Monte Carlo. Ces méthodes ont été formalisées dans les années 1940 par Stanislaw Ulam et John von Neumann, qui cherchaient à calculer la probabilité de gagner à un certain jeu de cartes. Alors qu'Ulam et von Neumann passèrent un temps important à effectuer de difficiles calculs combinatoires, Ulam se demanda s'il ne serait pas plus simple de répéter le jeu de cartes un grand nombre de fois, et d'estimer la probabilité théorique de victoire à l'aide de la fréquence empirique. Von Neumann comprit immédiatement le génie de l'approche d'Ulam et prit le soin de la programmer sur l'ordinateur ENIAC qu'il venait de créer. Les travaux d'Ulam et von Neumann eurent des applications immédiates pour le projet Manhattan, ce projet qui conduira à l'invention de l'arme nucléaire.

Depuis, les applications des simulations de Monte Carlo ont fleuri et envahi de très nombreux domaines. Que ce soit en physique quantique, en aérodynamique, en thermodynamique, en physique statistique, en astrophysique, en analyse d'appareils de mesure, en ingénierie électronique, en géostatistique, en énergie, en écologie, en robotique, en télécommunication, en étude du risque, en traitement du signal, en climatologie, en phylogénie (l'étude de l'arbre du vivant), en biologie moléculaire, en infographie (notamment pour calculer les trajectoires des rayons de lumière), ou encore en finance (notamment en gestion de portfolios). Ces simulations sont particulièrement utiles pour étudier la sensibilité des systèmes à des variations dans les conditions initiales.

Le cas le plus parlant de cette étude est peut-être celui de la météorologie. À cause du fameux effet papillon, la précision des mesures météorologiques est vouée à toujours être insuffisante pour effectuer des prédictions déterministes. Une minuscule erreur de mesure, ou l'absence de mesure d'un petit phénomène, peut conduire à une prédition très erronée. Les météorologues ont pris ceci en compte, et plutôt que de prétendre effectuer des prédictions déterministes, leurs prédictions sont désormais probabilistes. Pour les générer, ils vont typiquement simuler plusieurs météos futures possibles, en variant légèrement les conditions initiales conformément à l'imprécision des mesures. Autrement dit, ils effectuent des simulations de Monte Carlo pour un ensemble de conditions initiales crédibles. Leur pari est ensuite que la fréquence empirique des simulations correspondra à une prédition bayésienne pertinente.

⁶  *Calculating pi with darts* | Physics Girl | D. Muller et D. Cowern (2015)

La descente de gradient stochastique (SGD)

À l’ère du Big Data, une autre application encore des simulations de Monte Carlo peut consister à prélever un échantillon représentatif d’un ensemble de données. Cette approche simpliste s’est d’ailleurs retrouvée au cœur de l’un des algorithmes les plus importants du *machine learning* d’aujourd’hui, à savoir la descente de gradient stochastique (ou SGD en anglais).

Au lieu de chercher à faire coller une théorie à tout un jeu de données, SGD va tirer au hasard quelques données, et faire un pas vers l’explication de ces quelques données. En termes de réseaux de neurones, ceci correspond à ajuster légèrement les connexions synaptiques pour que le calcul du réseau de neurones corresponde mieux à la dernière donnée tirée au hasard. SGD itérera ensuite cette approche un grand nombre de fois, jusqu’à ce que ses explications des données échantillonnées aléatoirement soient suffisamment justes.

On pourrait croire que SGD y perd à ne traiter les données que de manière séquentielle et dans un ordre aléatoire. Ce n’est pas vraiment le cas. D’un point de vue théorique, on peut ainsi prouver que les performances de SGD ne sont pas significativement moins bonnes qu’une descente de gradient exact. Côté application, les gains en temps de calculs de SGD ont fait de SGD la solution préférentielle pour permettre aujourd’hui au *deep learning* de Google et Facebook de devenir l’état de l’art.

Mais il y a beaucoup, beaucoup mieux. En 2017, Mandt, Hoffman et Blei réussirent à réinterpréter SGD comme étant une inférence bayésienne approchée⁷. En particulier, chaque tirage aléatoire de SGD permet alors de faire fluctuer les paramètres du modèle. En ajustant les paramètres de cette fluctuation, les trois chercheurs ont même su montrer que ces fluctuations permettaient d’explorer adéquatement un ensemble de modèles crédibles, plutôt que de se restreindre uniquement au MAP (le modèle le plus crédible). Étrangement, les fluctuations aléatoires de SGD pourraient être non pas une faiblesse, mais un atout !

À l’instar de *dropout* dont on a parlé au chapitre 12, l’aspect stochastique de SGD, et en particulier sa non-convergence vers un MAP, pourraient être une solution incontournable pour mieux approximer la formule de Bayes. En particulier, ce faisant, on obtient une forme de moyennisation, dont on a vu l’utilité pour éviter le problème de l’*overfitting*. Cette découverte a ainsi grandement modifié mes crédences en l’utilisation d’une forme de SGD par nos cerveaux — même si je reconnaissais là encore volontier toute l’étendue de mon ignorance !

D’un point de vue technique, l’échantillonnage de SGD est toutefois simpliste. Il suffit de tirer au sort une donnée dans une liste connue au préalable. Mais dans des cas plus sophistiqués, l’échantillonnage peut devenir tout un champ de recherche en lui-même.

⁷  *Stochastic Gradient Descent as Approximate Bayesian Inference* | S. Mandt, M. Hoffman et D. Blei (2017)

Les nombres pseudo-aléatoires

Comment tireriez-vous un point au hasard dans un carré ? On l'a vu, Cowern et Muller ont eu bien du mal à distribuer uniformément leurs fléchettes sur leurs cibles carrées.

Très rapidement, John von Neumann se rendit compte de la difficulté de générer du hasard. Avant lui, certains statisticiens avaient choisi des nombres trouvés dans des complexes tableaux de chiffres, comme les tables logarithmiques. En 1939, la *RAND Corporation* publia un livre contenant 100 000 chiffres obtenus via un dispositif électronique mesurant des numéros obtenus à la roulette. Mais ceci ne suffisait pas aux simulations de Monte Carlo de von Neumann.

Pour mieux automatiser ses simulations de Monte Carlo, von Neumann chercha alors une manière de générer des nombres aléatoires par une machine. Sauf que, comme von Neumann le fit remarquer lui-même, « quiconque qui considère des méthodes arithmétiques pour produire des nombres aléatoires est, bien sûr, en plein péché. » Mais von Neumann comprit aussi qu'un *vrai* aléatoire n'était pas nécessaire à ses simulations. Il suffisait que ces nombres aient des propriétés « suffisamment aléatoires ». Ainsi naquirent les nombres *pseudo-aléatoires*⁸.

À l'aide de ces nombres pseudo-aléatoires, von Neumann put déterminer des algorithmes déterministes qui génèrent des séries de nombres pseudo-aléatoirement indépendants et uniformément distribués entre 0 et 1. Ces nombres forment le socle de tout échantillonnage des lois de probabilité. Par exemple, tirez deux nombres pseudo-aléatoires (et pseudo-indépendants) entre 0 et 1. Vous obtiendrez alors les coordonnées d'un point pseudo-aléatoire uniformément distribué dans le carré. On y est arrivé !

Mais faisons maintenant plus difficile. Comment obtenir un point uniformément distribué dans le disque inscrit dans le carré ? Et quid de distributions plus générales ?

L'échantillonnage préférentiel

Pour tirer un point pseudo-aléatoire dans un disque, il existe en fait une méthode étonnamment simple. Tirez un point pseudo-aléatoire uniformément distribué dans le carré. Si ce point est à l'extérieur du disque, rejetez-le et recommencez. Sinon, acceptez ce point. On peut alors démontrer que les points acceptés seront uniformément distribués dans le disque !

L'exemple du disque est en fait un cas particulier d'une approche plus générale appelée l'échantillonnage préférentiel, ou *importance sampling*. Cet échantillonnage préférentiel permet un échantillonnage pondéré d'une distribution cible, à

⁸  How to Generate Pseudorandom Numbers | Infinite Series (2017)

l'aide d'une distribution de référence que l'on sait déjà échantillonner. C'est ce qu'on a fait pour échantillonner le disque à partir d'un échantillonnage du carré.

Plus généralement, pour échantillonner la distribution cible, on échantillonne selon la distribution de référence. Puis, on donne au point tiré une importance proportionnelle à la probabilité de ce point tiré selon la distribution cible. Dans le cas du disque, cette importance était ou bien 0 (si le point était à l'extérieur du disque), ou bien 1 (si le point était à l'intérieur). Et oui ! Accorder une importance nulle à un point, c'est finalement la même chose que le rejeter.

L'échantillonnage préférentiel est particulièrement utile lorsque l'on ne connaît pas la loi de probabilité précise d'une variable, mais que l'on est capable de calculer les probabilités relatives de deux valeurs d'une variable. C'est typiquement le cas du disque, où l'on ne connaît pas la probabilité (ou plutôt, la densité de probabilité) d'un point à l'intérieur du disque, mais on sait qu'il sera tout aussi probable que n'importe quel autre point à l'intérieur du disque. C'est aussi le cas pour de nombreux modèles à variables cachées.

L'échantillonnage préférentiel au secours de LDA

Revenons-en à LDA. Imaginons avoir lu un ensemble mots de mots dans un texte. Quels autres mots x aura-t-on de bonnes chances de trouver dans ce texte ? Pour le savoir, LDA va d'abord chercher à déterminer à quelle catégorie ce texte appartient, pour ensuite déduire quels mots sont probables dans un texte de cette catégorie. Pour ce faire, il lui faut donc d'abord déterminer les catégories Cat probables, étant donné les mots. Comme LDA considère que les catégories causent les mots (souvenez-vous, c'est un modèle causal !), il va donc nous falloir inférer des causes étant donné des conséquences. Il nous faut donc utiliser la formule de Bayes. LDA va ainsi devoir calculer :

$$\mathbb{P}[\text{Cat}|\text{mots}] = \frac{\mathbb{P}[\text{mots}|\text{Cat}]\mathbb{P}[\text{Cat}]}{\mathbb{P}[\text{mots}]}.$$

La grande difficulté de cette équation, c'est le dénominateur. Ce dénominateur, aussi appelé marginal ou fonction de partition, nécessite de combiner toutes les combinaisons de catégories imaginables qui ont pu produire l'ensemble mots de mots du texte. Or ces combinaisons de catégories, il y en a en fait une infinité ! Souvenez-vous, un texte peut être 1/3-personnel, 2/3-travail. Mais ces fractions peuvent en fait être n'importe quelle combinaison de nombres positifs dont la somme est égale à⁹ 1. À moins d'hypothèses additionnelles, il est illusoire d'espérer calculer exactement la probabilité qu'un texte appartienne à telle ou telle catégorie selon LDA, étant donné des mots du texte !

⁹Le calcul exact nécessite de calculer une intégrale sur l'ensemble des combinaisons de catégories possibles !

Cependant, et à l'instar du cas du disque, on peut calculer les probabilités relatives de deux catégories différentes $\text{Cat}C$ et $\text{Cat}D$, sans connaître le dénominateur. En effet, on a

$$\frac{\mathbb{P}[\text{Cat}C|\text{mots}]}{\mathbb{P}[\text{Cat}D|\text{mots}]} = \frac{\mathbb{P}[\text{mots}|\text{Cat}C]\mathbb{P}[\text{Cat}C]}{\mathbb{P}[\text{mots}|\text{Cat}D]\mathbb{P}[\text{Cat}D]}.$$

On peut alors appliquer l'échantillonnage préférentiel pour construire un échantillon pondéré représentatif des catégories auxquelles le texte a de bonnes chances d'appartenir, puis d'en déduire d'autres mots probables.

Si compléter un texte à l'aide de LDA ne semble pas d'un grand intérêt, certaines variantes de ce problème correspondent à des milliards de chiffres d'affaire. À l'heure du *Big Data*, l'un des problèmes informatiques les plus lucratifs est celui de la recommandation. Des milliards ont été investis pour répondre au mieux à la question suivante. Étant donné votre historique Facebook, iTunes ou Amazon, quels posts, musiques ou produits serait-il pertinent de vous proposer ?

Cette question a notamment connu son heure de gloire quand, le 2 octobre 2006, l'entreprise de flux continu de films et séries télévisées Netflix a lancé le challenge Netflix¹⁰. Destiné aux *data scientists*, cette épreuve consistait à prédire les notes que des utilisateurs donneraient à certains films, étant donné toutes les notes qui ont été données. Plus précisément, Netflix possédait une base de données de 100 millions de notes données par 480 mille utilisateurs à 18 mille films. Une certaine fraction de ces notes fut rendue publique. Il s'agissait de deviner le reste.

S'il ne s'agit pas là du même problème que celui que LDA résout, j'imagine que vous ne serez pas indifférent à la similarité. Or, améliorer les prédictions de quelques pourcents pourrait augmenter la rétention des utilisateurs de plusieurs pourcents, ce qui augmenterait le chiffre d'affaires de plusieurs pourcents, ce qui représente potentiellement des millions, voire des milliards de dollars !

Certes, LDA n'a pas été un outil central dans la résolution du challenge Netflix. Mais parmi les meilleurs outils pour résoudre ce genre de problème, on trouve d'autres sortes de réseaux bayésiens. De nos jours, le modèle le plus performant semble être celui des *Generative Adversarial Networks* (GANs). Le principe des GANs est la construction de réseaux bayésiens profonds avec peu, voire aucune perturbation stochastique intermédiaire¹¹. La principale source de probabilité est alors l'incertitude sur les variables cachées très profondes. Mais à l'instar des nombres pseudo-aléatoires élémentaires de von Neumann, cette incertitude correspond généralement à une loi de probabilité très simple à échantillonner¹².

¹⁰  *The Netflix Prize* | ZettaBytes | A.M. Kermarrec (2017)

¹¹ L'utilisation de *dropout* correspond justement à rajouter des perturbations stochastiques intermédiaires.

¹² Cependant, l'inférence bayésienne est alors difficile. L'astuce consiste alors à construire d'autres réseaux de neurones dont la tâche est d'aider à calculer cette inférence bayésienne !

Mais pour en revenir au challenge Netflix, il se trouve que c'est l'autre architecture de modèle probabiliste à variables cachées qui aura joué un rôle déterminant. Cet autre modèle nous vient de la physique.

Le modèle d'Ising*

Dans les années 1920, Wilhelm Lenz et son étudiant Ernst Ising cherchèrent à comprendre les transitions de phase. À bien des égards, il s'agit là d'un problème de recherche encore largement ouvert aujourd'hui. Plutôt que de l'attaquer dans toute sa généralité, Lenz et Ising se restreignirent au cas de la transition de phase ferromagnétique : alors qu'à basse température, le fer est magnétique, à haute température, ces propriétés magnétiques disparaissent.

Pour comprendre l'origine de ce phénomène, Lenz et Ising cherchèrent une explication d'origine microscopique. Le moment magnétique du fer est la somme des moments magnétiques de ses atomes, appelés *spins*, dont la valeur peut être $+1$ ou -1 . Lenz et Ising supposèrent que les spins des atomes voisins avaient tendance à s'aligner. Ils décrivirent cela en postulant que l'énergie d'une paire de spins alignés est -1 , alors que celle de spins opposés est $+1$. L'énergie totale E du fer est alors la somme des énergies de ces interactions locales entre spins voisins.

La question que se posèrent Lenz et Ising fut de savoir si, à une température T donnée, les *spins* auraient tendance à s'aligner ou non. Autrement dit, l'ensemble des configurations où les *spins* s'alignent est-il, à température T , plus probable que l'ensemble des configurations où les *spins* ne s'alignent pas ?

Fort heureusement pour Lenz et Ising, un problème similaire avait déjà été résolu un demi-siècle plus tôt par Ludwig Boltzmann. Boltzmann découvrit qu'à l'équilibre thermodynamique et à une température T , la probabilité d'une configuration i à énergie E_i est proportionnelle à $\exp(-kE_i/T)$, où k est ce que l'on appelle la constante de Boltzmann. Plus précisément, la loi de Boltzmann dit que la probabilité de la configuration i est

$$\mathbb{P}[i|T] = \frac{\exp(-kE_i/T)}{\exp(-kE_i/T) + \sum_{j \neq i} \exp(-kE_j/T)}.$$

Le dénominateur de cette équation est la fameuse fonction de partition, dont le calcul est impossible car le nombre de configurations est typiquement exponentiel en le nombre d'atomes. En particulier, si l'on en croit la loi de Boltzmann, une configuration i est exponentiellement rare en son énergie E_i . De façon cruciale, cela est d'autant plus le cas à faible température T .

C'est précisément cela qui explique le rôle de la température dans la transition de phase ferromagnétique. À faible température, les configurations de hautes énergies, celles où les spins ne s'alignent pas, sont exponentiellement rares très

vite ; les configurations crédibles sont donc celles où les spins s'alignent, ce qui rend le morceau de fer magnétique. À l'inverse, à haute température et si l'on suppose que kE_i est beaucoup plus petit que T , alors les quantités $\exp(-kE_i/T)$ sont toutes proches de 1. Cependant, les configurations où les *spins* ne sont pas alignés sont exponentiellement plus nombreuses que celles où les *spins* s'alignent. En effet, si vous tirez au hasard (uniformément et indépendamment) les valeurs des spins, il n'y a quasiment aucune chance qu'ils soient alignés. Du coup, à haute température, l'ensemble des configurations où les *spins* ne s'alignent pas est exponentiellement plus probable que l'ensemble des configurations où ils s'alignent, d'où la disparition du magnétisme du fer.

Le modèle d'Ising est fascinant pour plusieurs raisons. Premièrement, il s'agit de l'un des plus simples modèles expliquant les transitions de phase. En second lieu, la compréhension du modèle d'Ising passe par la distribution de Boltzmann, celle qui lie l'énergie E d'une configuration à sa probabilité. Enfin, et surtout, le modèle d'Ising est un merveilleux exemple de *champs aléatoires de Markov*, qui se trouvent être au cœur de nombreux modèles modernes de *machine learning*.

La machine de Boltzmann

Les champs de Markov sont décrits par des variables reliées entre elles par des arêtes non-orientées. Ils sont ainsi très similaires aux réseaux bayésiens. Cependant, par opposition aux réseaux bayésiens, les arêtes des champs de Markov ne représentent pas des liens de cause à effet. Elles représentent davantage des espèces de corrélations entre variables, par opposition à des variables qui ne sont pas reliées et qui sont intuitivement presque indépendantes. C'est grossièrement cela que l'on appelle la propriété de Markov des champs aléatoires¹³.

Il y a un cas en particulier de champ de Markov qui se prête particulièrement bien au *machine learning*, à tel point qu'il fut l'un des ingrédients indispensables à la résolution du challenge Netflix. Ce cas particulier est appelé *machine de Boltzmann restreinte*. À l'instar d'un réseau bayésien à variables cachées, la machine de Boltzmann consiste à corrélérer les variables observables avec des variables cachées. Autrement dit, la machine de Boltzmann restreinte est un champ de Markov dont toute arête lie une variable observable à une variable cachée¹⁴. Qui plus est, à l'instar du modèle d'Ising, la machine de Boltzmann associe à toute arête une énergie de corrélation¹⁵.

¹³Le théorème de Hammersley-Clifford démontre que la propriété de Markov est équivalente à dire que la probabilité conjointe se factorise en produit de fonctions des cliques, i.e. s'écrit $\mathbb{P}[X = x] = \prod_{\text{Clique } c} f_c(x_c)$, où x_c est le vecteur des x_i pour $i \in c$.

¹⁴Autrement dit, il s'agit d'un graphe biparti, avec les variables observables d'un côté et les variables cachées de l'autre.

¹⁵Il est alors possible de généraliser cette configuration en considérant une couche de données observées, reliée à une première couche de variables cachées, elle-même reliée également à une seconde couche de variables cachées, et ainsi de suite. On obtiendrait là une machine de Boltzmann profonde.

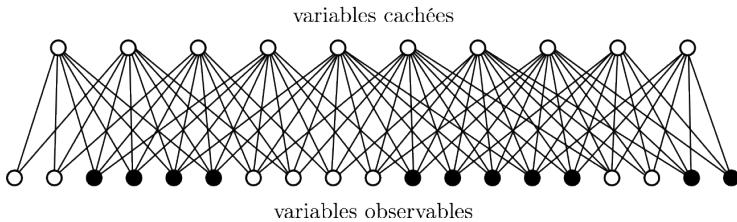


Figure 17.3. Une machine de Boltzmann relie les variables observables via des variables cachées. Quand certaines variables observables sont observées (ici en noires), on peut alors utiliser les liens cachés via des variables cachées pour deviner les valeurs crédibles des variables observables non-observées.

L’application de la machine de Boltzmann restreinte au challenge Netflix se fait alors en deux phases. Dans un premier temps, on utilise les données pour apprendre les paramètres de la machine. Ces paramètres sont les équations des énergies de corrélation des arêtes¹⁶. Une fois les paramètres déterminés, la machine de Boltzmann utilise ensuite la distribution de Boltzmann pour effectuer des prédictions probabilistes qui adressent le challenge Netflix. Étant donné les données observables connues, la machine de Boltzmann infère des variables cachées crédibles, desquelles elle déduit les valeurs crédibles des variables observables non-observées.

La *pure bayésienne* pourrait effectuer ce raisonnement sans difficulté, et en déduire alors les crédences adéquates sur les valeurs possibles des variables observables non-observées. Mais ce raisonnement échappe au *bayésien pragmatique*, parce qu’il nécessite le calcul exact de l’effrayante fonction de partition. C’est pour ça que, bien souvent, le *bayésien pragmatique* devra alors se contenter d’un échantillonnage. Il semble que lui aussi soit contraint de raisonner avec des exemples (représentatifs).

MCMC et Google PageRank

Malheureusement, l’échantillonnage préférentiel est souvent incapable de fournir un bon échantillonnage des machines de Boltzmann en temps raisonnable. De façon générale, si la distribution de référence de l’échantillonnage préférentiel est trop distincte de la distribution cible (et c’est souvent le cas en grandes dimensions¹⁷), alors l’échantillonnage préférentiel peut nécessiter un trop grand nombre d’itérations pour être représentatif de la distribution cible. En fait, déterminer une bonne façon d’effectuer un échantillonnage rapide et représentatif est un champ de recherche difficile.

¹⁶On suppose généralement que l’énergie d’une arête est bilinéaire en les variables observables et cachées liées par l’arête. Reste alors à déterminer le coefficient associé à cette forme bilinéaire.

¹⁷ *Hypersphères* | IA 19 | Science4All | J. Cottanceau et L.N. Hoang (2018)

Bizarrement, pour résoudre ce problème, il est souvent utile de remplacer les répétitions d'expériences indépendantes de Monte Carlo par des méthodes dites de *Markov-Chain Monte Carlo* (MCMC). Ces méthodes consistent à effectuer un parcours aléatoire dans l'ensemble des possibles. Les étapes de ce parcours aléatoire ne seront absolument pas représentatives de la distribution cible.

Cependant, pourvu que chaque transition de ce parcours soit adéquatement choisie, il demeure possible de garantir qu'à l'infini, la fréquence de visite de tout point¹⁸ dans l'ensemble des possibles converge vers la probabilité de ce point selon la distribution cible. Dit plus simplement, à l'infini, l'échantillonnage sera parfaitement représentatif de la distribution cible.

De prime abord, on pourrait croire que MCMC est une très mauvaise idée. Détrompez-vous. De façon étonnante, l'algorithme qui a fait la fortune de Google est un algorithme MCMC ! Cet algorithme, appelé PageRank, est le cœur des premières versions du moteur de recherche. L'astuce de PageRank fut de calculer l'importance de chaque page web, en fonction de l'importance que les autres pages web lui accordent, et de l'importance de ces autres pages web. Ainsi, une page Wikipedia peut être considérée importante, car de nombreuses autres pages web, y compris des pages web importantes, ont des liens qui redirigent vers cette page Wikipedia.

Cependant, parce que l'importance de chaque page web dépend de l'importance des autres pages web, calculer l'importance d'une page web revient à résoudre une équation horriblement compliquée : cette équation a autant d'inconnues qu'il y a de pages web ! L'astuce qu'ont trouvée Larry Page et Sergey Brin, les fondateurs de Google, fut de résoudre le problème via MCMC. Leur algorithme PageRank va imaginer un surfeur imaginaire qui se promène sur le web en cliquant aléatoirement sur un des liens de la page qu'il est en train de visiter. Notre surfeur va ainsi sauter de page en page. Intuitivement, on s'attend à ce que les pages qu'il visitera alors un grand nombre de fois soient celles qui ont de l'importance, puisque beaucoup de chemins mènent à ces pages.

On peut en fait prouver que, si le web est fortement connexe, alors la fréquence empirique à laquelle le surfer arrive à une page donnée convergera exactement vers l'importance que l'on a cherché à calculer précédemment¹⁹. En particulier, au bout d'une simulation suffisamment longue, cette fréquence empirique sera une bonne approximation de l'importance de la page²⁰.

Ce principe remarquable est ce qui a permis à Page et Brin de créer l'une des entreprises les plus puissantes de la planète ! Rien que ça !

¹⁸Dans le cas de distributions continues, il faudrait plutôt parler de la fréquence de visite de tout ensemble ouvert.

¹⁹Avec probabilité 1 (presque sûrement).

²⁰  *Can a Chess Piece Explain Markov Chains?* Infinite Series | K. Houston-Edwards (2017)

L'échantillonnage de Metropolis-Hasting

Si l'approche de PageRank a été excellente pour organiser le web, elle ne semble toutefois pas adaptée pour échantillonner la distribution d'une machine de Boltzmann restreinte. Il existe une autre approche pour ce faire, appelée échantillonnage de Metropolis-Hasting. Comme PageRank, Metropolis-Hasting va nous emmener en promenade. Cependant, cette promenade aura maintenant lieu dans l'espace des valeurs imaginables des variables observables et cachées.

À chaque instant de cette promenade aléatoire, on considère un pas aléatoire à effectuer, que l'on peut refuser s'il nous amène vers un état de trop petite probabilité. Plus précisément, appelons i la position actuelle, et supposons que le pas aléatoire nous amène à une position j . Pour savoir s'il faut accepter ou refuser ce pas aléatoire, Metropolis-Hasting nous dit de calculer au préalable un taux d'acceptation A . On définit A comme suit :

$$A = \frac{\mathbb{P}[j]}{\mathbb{P}[i]} \frac{\mathbb{P}[\text{Pas}(i \rightarrow j)|i]}{\mathbb{P}[\text{Pas}(j \rightarrow i)|j]}.$$

Intuitivement, le taux d'acceptation d'un pas de i vers j est grand, si j est un état plus probable que i et si ce pas est probablement réversible, dans le sens où la probabilité que le pas inverse soit proposé par la promenade aléatoire n'est pas beaucoup plus faible que la probabilité du pas aléatoire. Si le taux d'acceptation A est supérieur à 1, alors l'échantillonnage de Metropolis-Hasting nous dit d'effectuer ce pas. Sinon, il nous faut tirer une pièce aléatoire, dont la probabilité de tomber sur « oui » est le taux d'acceptation A .

De façon cruciale, lorsque chaque pas a la même probabilité que le pas inverse, A peut être calculé sans faire appel à la fonction de partition. Dans le cas de la distribution de Boltzmann, on a $\mathbb{P}[j]/\mathbb{P}[i] = \exp(kE_j/T)/\exp(kE_i/T) = \exp(k(E_i - E_j)/T)$. Néanmoins, malgré la non-utilisation de la fonction de partition, sous certaines hypothèses raisonnables²¹, après un temps suffisamment long et à l'instar de PageRank, cette promenade aléatoire va fournir un échantillon représentatif de la distribution cible.

Mieux encore, il est alors possible de conditionner l'échantillon aux variables observables observées, en forçant la promenade aléatoire à ne jamais modifier leurs valeurs. On obtiendra alors un échantillonnage représentatif des variables non-observées, sachant les variables observées.

L'échantillonnage de Metropolis-Hasting se décline en de nombreuses variantes utiles au *bayésien pragmatique*, pour toutes les distributions dont les fonctions de partition sont trop longues à estimer correctement. D'un côté, il y a des variantes dites *adaptatives*, qui permettent d'optimiser les propriétés des pas aléatoires pendant l'échantillonnage. De l'autre, il existe des approximations

²¹Notamment le fait que la promenade a une probabilité non nulle de nous amener un peu partout.

de Metropolis-Hastings, qui le rendent applicable même lorsque les rapports $\mathbb{P}[j]/\mathbb{P}[i]$ des probabilités de deux positions i et j ne sont pas directement calculables. C'est typiquement le cas des modèles génératifs de données complexes, comme les simulations de l'univers dont on a parlé à la fin du chapitre 15. Dès lors, on peut chercher à remplacer ces probabilités par des mesures de performances $\text{perf}(j)$ et $\text{perf}(i)$ des positions i et j , en espérant que ces quantités seront suffisamment corrélées avec les probabilités de ces positions.

L'échantillonnage de Gibbs

Cependant, c'est une autre méthode encore d'échantillonnage par MCMC qui est la plus souvent utilisée dans le cas de la machine de Boltzmann. Cette autre méthode, appelée échantillonnage de Gibbs, repose sur une propriété remarquable de la machine de Boltzmann restreinte. Souvenez-vous, dans cette machine, les variables observables ne sont reliées qu'à des variables cachées, et vice-versa. En particulier, deux variables observables ne sont jamais reliées, tout comme deux variables cachées ne le sont jamais non plus²².

L'idée de l'échantillonnage de Gibbs, c'est d'alterner l'échantillonnage des variables observables et celui des variables cachées. De façon cruciale, étant donné les valeurs des variables observables, l'énergie totale de la machine de Boltzmann restreinte est alors une fonction linéaire des variables cachées²³. En particulier, la contribution d'une variable cachée à l'énergie totale est un terme qui ne dépend pas des valeurs des autres variables cachées. Du coup, étant donné les variables observées, chaque variable cachée peut être échantillonnée indépendamment²⁴ ! Et ça, c'est facile à échantillonner.

Étant donné certaines variables observables observées, l'échantillonnage de Gibbs consiste alors à d'abord attribuer des valeurs arbitraires aux variables observables non-observées. Ce sera le point de départ arbitraire de la promenade aléatoire. Puis, l'échantillonnage de Gibbs échantillonnera toutes les variables cachées conditionnellement aux variables observables. Ensuite, il échantillonnera toutes les variables observables non-observées conditionnellement aux variables cachées. On aura alors fait un pas aléatoire dans l'espace des variables observables.

L'échantillonnage de Gibbs consiste ensuite à répéter de tels pas aléatoires. Au bout d'un temps suffisamment grand, ce calcul permettra d'obtenir un échantillon représentatif des valeurs crédibles des variables observables non-observées, étant donné les variables observables observées.

²²Ceci se généralise d'ailleurs très bien aux machines de Boltzmann profondes.

²³Ceci nécessite de supposer que l'énergie totale $E(v, h)$ est bilinéaire en les variables observables v et les variables cachées h .

²⁴Autrement dit, conditionnellement aux variables observables, les variables cachées sont indépendantes. Ceci est en fait une conséquence directe du fait que l'ensemble des variables cachées forment un ensemble stable du graphe du champ de Markov.

Dans tous les cas de MCMC que l'on a vus, que ce soient PageRank, Metropolis-Hastings ou Gibbs, il est crucial que l'échantillonnage dure suffisamment longtemps. En effet, sinon, les données échantillonées seront fortement dépendantes des points de départ des promenades aléatoires, lesquelles n'ont aucune raison d'être représentatives de la distribution que l'on veut échantillonner.

Pire encore, il arrive souvent que la quasi-totalité des données imaginables soient en fait d'une probabilité négligeable. Autrement dit, seule une poignée de données parmi un ensemble immense possède une crédence non-négligeable. Voilà qui est d'ailleurs typiquement le cas de l'ensemble des théories crédibles, qui forment souvent une poignée d'îlots dans un océan de très grande dimension rempli de théories farfelues. Dès lors, tant que MCMC ne sera pas tombé sur l'une des rares données ou théories crédibles, alors il sera impossible pour MCMC de se rendre compte que les données et théories explorées jusque-là ne sont pas crédibles. Souvenez-vous. MCMC ne connaît que les crédences des données explorées relativement aux autres données explorées. Du coup, si aucune donnée vraiment crédible n'a été explorée, toutes les données explorées paraîtront alors crédibles !

En bref, les échantillons intermédiaires de MCMC ne sont pas représentatifs de la distribution que MCMC cherche à échantillonner. Pire encore, il est impossible d'anticiper ou de déterminer la représentativité d'un échantillon obtenu par MCMC. Certes, à l'infini, MCMC deviendra valide. Mais MCMC nécessitera beaucoup de temps pour y parvenir. Et il n'offrira jamais aucune garantie.

Néanmoins, parce qu'il permet des échantillonnages qu'aucune autre méthode ne semble capable de fournir avec la même efficacité, MCMC est devenu un outil indispensable au *bayésien pragmatique*.

MCMC et les biais cognitifs

Vu la complexité de MCMC, j'aurais toutefois pu vous épargner sa description détaillée. Si j'ai malgré tout tenu à passer du temps à expliquer son principe et son utilité, ce n'est pas tant pour que vous l'utilisiez en pratique. C'est surtout pour vous expliquer pourquoi l'évolution darwinienne a sans doute été contrainte de faire de nos cerveaux des calculateurs de MCMC ; et pourquoi il est crucial que vous vous en rendiez compte.

Le psychologue Daniel Kahneman raconte que, pendant la période des attaques terroristes répétées dans les bus en Israël, en voiture, il se dépêchait de s'éloigner de tout bus qu'il voyait. Il en eut honte. En bon statisticien, il savait pertinemment que la probabilité de mourir à cause d'un accident de la route suite à un coup de volant demeurait bien plus importante que celle de voir le bus être attaqué. Le terrorisme cause un nombre de morts négligeables devant le nombre de morts sur les routes. Mais même un cerveau bien informé et bien éduqué comme celui de Kahneman n'arrivait pas à ne pas surévaluer le danger du terrorisme.

Ce biais cognitif est ce que Kahneman appelle le biais de disponibilité. On a tendance à donner trop d'importance aux premières idées qui nous viennent à l'esprit. Si je vous demande de visualiser Linda, cette militante anti-nucléaire de 31 ans, il est probable que vous ayiez très vite une image en tête, et que vous fassiez de l'*overfitting* sur cette image. Ce biais semble révélateur de l'utilisation de MCMC par nos cerveaux. À moins de réfléchir suffisamment longtemps, MCMC est fortement biaisée par son point de départ — et l'abattage médiatique fait que le point de départ des pensées catastrophistes est souvent le terrorisme.

Si ce biais peut paraître relativement évident, c'est en partie parce qu'il n'est pas trop dur d'être conscient de ce à quoi on pense lorsque l'on est victime du biais de disponibilité. Cependant, nombre de signaux qui traversent nos cerveaux sont inconscients ; et ceci n'empêche pas MCMC d'en faire le point de départ de sa promenade aléatoire dans le monde des idées. Dès lors, c'est à notre insu que ce que l'on pense est incroyablement affecté par le contexte dans lequel on se trouve. C'est l'effet d'amorçage, ou *priming effect*.

L'une des expériences les plus bluffantes à ce sujet est celle réalisée par Gary Wells et Richard Petty²⁵ en 1980. Wells et Petty invitérent 72 étudiants à tester un casque audio dans divers contextes d'utilisation. 24 étudiants durent écouter un éditorial sur l'augmentation des frais de scolarité de 587 \$ à 750 \$, avec ce casque et sans bouger la tête. 24 l'écouterent en hochant la tête (comme pour dire « oui »), et les 24 derniers l'écouterent en secouant la tête (comme pour dire « non »). À l'issue de cela, en dernière question d'un questionnaire qui portait sur la qualité des écouteurs, les chercheurs demandèrent aux étudiants leurs avis quant aux frais de scolarité. Les résultats sont stupéfiants. En moyenne, ceux qui secouèrent la tête répondirent 467 \$, ceux qui ne bougèrent pas la tête dirent 582 \$, et ceux qui hochèrent la tête avancèrent 646 \$. Incroyable ! Notre jugement est déterminé par nos actions, et nous en sommes complètement inconscients.

Un autre exemple stupéfiant de notre dépendance en le point initial de notre raisonnement MCMC est l'étonnant effet d'ancrage. Kahneman et Tversky ont réalisé une expérience troublante pour le révéler. Kahneman et Tversky tirèrent d'abord un nombre au hasard devant les yeux d'un sujet, qui était soit 10, soit 65. Imaginons que l'on soit tombé sur le nombre 10. Kahneman et Tversky demandèrent alors au sujet s'il y avait plus ou moins de 10 % de pays africains dans le monde. Si le nombre tiré était 65, ils posèrent la même question avec le chiffre de 65 %. Puis ils demandèrent au sujet d'estimer le pourcentage de pays africains. De façon stupéfiante, à cette seconde question, les sujets répondirent en moyenne 25 % lorsque le nombre aléatoirement tiré était 10, et 45 % lorsque celui-ci était 65 ! D'autres variantes montrent que cet effet stupéfiant persiste, même si les nombres proposés par les chercheurs sont absurdement petits ou grands²⁶.

²⁵  *C'est vraiment moi qui décide ?* Science Étonnante | D. Louapre (2010)

²⁶  *L'effet d'ancrage* | Crétin de cerveau | Science Étonnante | D. Louapre (2016)

Un troisième exemple, parmi tant d’autres, qui illustre notre utilisation permanente de MCMC dans la manière dont on pense, y compris sans s’en rendre compte, est l’aversion aux pertes. Cette aversion fut théorisée par la théorie des perspectives par Kahneman et Tversky, qui permet d’expliquer un grand nombre de biais cognitifs à la fois. Le postulat de cette théorie est le fait que nos préférences sont toujours fortement influencées par une référence. Les gains vis-à-vis de cette référence sont bien, mais les pertes vis-à-vis de cette référence sont catastrophiques. Voilà qui explique pourquoi les médailles d’argent des jeux olympiques tirent la gueule, contrairement aux médailles de bronze qui reçoivent leurs médailles avec un grand sourire²⁷.

MCMC n’est pas un problème en soi. Le problème est que la pertinence de MCMC n’apparaît qu’après un très grand nombre de pas aléatoires. Pire, les premières idées suggérées par MCMC peuvent être toutes très peu crédibles, notamment par opposition à une poignée d’idées crédibles dans l’immense espace des idées. Tant qu’aucune de ces idées crédibles n’aura été échantillonnée par MCMC, les conclusions de MCMC seront très erronées.

Sachant que notre cerveau applique sans doute MCMC, il nous faut absolument reconnaître l’étendue de notre ignorance. Et il faut absolument prendre le temps de la réflexion — bien plus que ce n’est le cas en pratique. En particulier, il se pourrait que les bienfaits des longues méditations et du sommeil résident en partie dans le calcul prolongé de MCMC.

La divergence contrastive et les rêves

Ce que l’on a vu jusque-là n’est toutefois qu’une partie de l’explication avancée par Crick et Mitchison. Pour ces deux chercheurs, l’utilité des rêves en particulier résiderait surtout dans un autre algorithme utile à l’apprentissage appelé *divergence contrastive*. Cet algorithme a pour but de calculer le maximum-a-posteriori d’un modèle à variables cachées, notamment lorsque l’on ne connaît que les probabilités des données relativement à d’autres, comme ce fut le cas pour l’inférence bayésienne de LDA ou pour la machine de Boltzmann.

Je vous épargne les étapes de calculs. L’égalité importante est

$$\partial_{\theta} \log \mathbb{P}[\theta|D] = \partial_{\theta} \log \mathbb{P}[\theta] + \partial_{\theta} \log \tilde{p}(D|\theta) - \mathbb{E}_{x|\theta} [\partial_{\theta} \log \tilde{p}(x|\theta)].$$

Autrement dit, pour savoir comment ajuster les paramètres θ pour obtenir une théorie plus crédible, il nous faut comprendre trois termes. Le premier est

²⁷Une autre explication (encore bayésienne !) à cet effet d’ancrage pourrait être le mécanisme d’anticipation et de correction de notre cortex cérébral proposé par Karl Friston. Selon cette théorie et conformément à l’optimisation de la communication proposée par Shannon, notre cerveau effectuerait constamment des prédictions, et ne réagirait que lorsque les observations contredisent ces prédictions, initiant ainsi un processus d’apprentissage.

l'effet des variations de θ sur l'*a priori*. Le deuxième est l'effet des variations de θ sur la probabilité non-normalisée du modèle (qui correspond typiquement à $\exp(-kE_i/T)$ dans le cas de la distribution de Boltzmann). Le troisième, enfin, est l'effet des variations de θ sur les probabilités non-normalisées des alternatives x à D .

Pour de nombreux modèles, notamment les machines de Boltzmann, les deux premiers termes sont très faciles à calculer. Cependant, le dernier terme va typiquement requérir un (long) échantillonnage représentatif. La remarque cruciale, toutefois, c'est que ce dernier terme ne dépend pas des données ! On n'est pas obligé d'observer le monde pour le calculer ! Il suffit de rêver pour le déterminer.

Et bien, pour Crick et Mitchison, c'est peut-être justement à cela que les rêves servent. Ils serviraient à calculer un maximum-a-posteriori, en échantillonnant l'effet d'un changement des paramètres du cerveau sur la probabilité non-normalisée des alternatives aux données observées.

Voilà une raison de plus de prendre le temps de la réflexion.

Références en français

-  *Théorème vivant* | Le Livre de Poche | C. Villani (2013)
-  *C'est vraiment moi qui décide ?* Science Étonnante | D. Louapre (2010)

-  *L'effet d'ancrage* | Crétin de cerveau | Science Étonnante | D. Louapre (2016)
-  *Au cœur de Google : Page Rank* | Wandida | R. Guerraoui (2013)
-  *Hypersphères* | IA 19 | Science4All | J. Cottanceau et L.N. Hoang (2018)

Références en anglais

-  *Deep Learning* | MIT Press | I. Goodfellow, Y. Bengio et A. Courville (2016)
-  *Thinking Fast and Slow* | SpringerFarrar, Straus and Giroux | D. Kahneman (2013)
-  *The function of dream sleep* | Nature | F. Crick and G. Mitchison (1983)
-  *Stochastic Gradient Descent as Approximate Bayesian Inference* | S. Mandt, M. Hoffman and D. Blei (2017)

-  *How a Kalman filter works, in pictures* | Bzarg | T. Babb (2015)
-  *Deep Dream - a code example for visualizing Neural Networks* | Google Research Blog | A. Mordvinster, C. Olah and M. Tyka (2015)

- ▶ *Pi and Buffon's Matches* | Numberphile | T. Padilla (2012)
- ▶ *Calculating pi with darts* | Physics Girl | D. Muller and D. Cowern (2015)
- ▶ *Inside Google: Page Rank* | Wandida | R. Guerraoui (2013)
- ▶ *What's a Random Number?* ZettaBytes | P. Blanchard (2016)
- ▶ *How to Generate Pseudorandom Numbers* | Infinite Series (2017)

La philosophie est écrite dans cet immense livre qui continuellement reste ouvert devant les yeux (ce livre qui est l'Univers), mais on ne peut le comprendre si, d'abord, on ne s'exerce pas à en connaître la langue et les caractères dans lesquels il est écrit. Il est écrit dans une langue mathématique, et les caractères en sont les triangles, les cercles, et d'autres figures géométriques, sans lesquelles il est impossible humainement d'en saisir le moindre mot ; sans ces moyens, on risque de s'égarter dans un labyrinthe obscur.

Galilée (1564-1642)

18

La déraisonnable efficacité de l'abstraction

Le *deep learning*, ça marche !

Le 10 mars 2016, AlphaGo fait les gros titres. À la surprise générale, cette intelligence artificielle de Google vient de battre Lee Sedol, pourtant considéré par beaucoup comme l'un des meilleurs joueurs du monde au jeu de go.

Le go est un jeu à deux joueurs grossièrement similaire aux échecs. Cependant, il a longtemps été aux échecs ce que les échecs sont au morpion. Le go est plus complexe, plus combinatoire et plus imprévisible. Les meilleurs joueurs de go s'appuient même parfois sur des intuitions dont ils ont du mal à expliquer les fondements — certains affirmaient même qu'il s'agissait d'une faculté humaine hors de portée des machines. Et pendant longtemps, les meilleurs algorithmes n'arrivaient qu'à la cheville des joueurs moyens de go.

Cependant, depuis quelques années, un certain modèle de *machine learning* enchaîne les succès époustouflants. Détection d'objets, reconnaissance de visage, reconstruction optique de caractères, traitement du langage naturel, traduction automatisée et systèmes de recommandation sont autant de problèmes auparavant inaccessibles aux machines. Mais tous furent tout à coup résolus par le *deep learning*. Tous les investisseurs de la Silicon Valley n'avaient plus que ces mots au bout de leurs lèvres. Et tous les géants du web se vantèrent des nouvelles fonctionnalités qu'ils étaient désormais capables d'offrir.

Le succès retentissant du *deep learning* contraste fortement avec de nombreuses approches d'ingénierie classiques. D'habitude, on construit une théorie qui garantit le bon fonctionnement d'une technologie. Puis, on constate avec frustration que la technologie échappe à la théorie. Vient alors la sempiternelle question : d'où vient cette différence entre la théorie et la pratique ? *En théorie, c'est la même chose*, dit la boutade. *Mais en pratique...*

Cependant, le cas du *deep learning* semble être l'exact opposé. On dispose là d'une technologie dont aucune théorie ne semble avoir prédit le succès. Les théoriciens sont pris de cours. Les performances du *deep learning* sont époustouflantes, mais personne ne semble savoir pourquoi. Le *deep learning* marche très bien en pratique ; mais est-ce que ça marche *en théorie* ? Et si, pour une fois, cette question intéresse les praticiens, c'est parce qu'eux aussi sentent que tout progrès dans la compréhension théorique du *deep learning* pourrait conduire à des progrès significatifs, voire spectaculaires, de l'intelligence artificielle.

Mais qu'est-ce que le *deep learning* ? De nos jours, la recherche est si foisonnante qu'il est difficile de cerner les frontières de ce domaine. De façon grossière, on peut toutefois considérer que le *deep learning* est l'étude des modèles empilant un grand nombre de couches de variables cachées, à l'instar de LDA et des machines de Boltzmann. Néanmoins, contrairement aux exemples du chapitre précédent, une architecture de *deep learning* ne cherche pas nécessairement à décrire une loi de probabilité. En fait, la plupart sont des réseaux de neurones conçus pour approcher des fonctions déterministes¹. Toutes ces structures ont toutefois un point commun : elles manipulent des variables cachées *profondes*, c'est-à-dire des variables intuitivement très distantes des variables observables. Plus rigoureusement, tout signal qui se propage des variables observables aux variables profondes passe par un grand nombre de variables intermédiaires.

Voilà qui nous invite à considérer que le niveau d'abstraction d'une variable cachée se mesure par sa profondeur. En effet, intuitivement, l'abstraction s'oppose à ce qui est tangible. Il est naturel de considérer que les variables observables sont justement ce qu'il y a de tangible dans un modèle ; et que les variables profondes sont donc abstraites. D'où le titre du chapitre — que l'on reliera ultérieurement à l'abstraction mathématique !

Le problème théorique que pose le succès du *deep learning* est alors l'explication de la déraisonnable efficacité de l'abstraction. On peut grossièrement distinguer trois approches pour ce faire. La première est celle de la synthétisation préalable nécessaire des données brutes. La deuxième est l'expressivité particulière des modèles profonds (en particulier des réseaux de neurones profonds). Enfin, la troisième, la plus prometteuse à mon sens, réside dans les propriétés algorithmiques des données que l'on veut étudier, mesurées en termes de sophistication de Solomonoff et de profondeur logique de Bennett.

Mais commençons par le besoin de synthétisation des données brutes.

¹En particulier, les GANs dont on a déjà parlé consistent à fixer une distribution Z de variables profondes, et à générer des données crédibles $X = f(Z)$, en ajustant la fonction f .

L'apprentissage des *features*

Avant le succès du *deep learning*, les algorithmes de *machine learning* nécessitaient l'intervention de *data scientists* compétents pour « nettoyer » les données brutes et préparer leur analyse. L'une des principales motivations de la recherche en *deep learning* fut de court-circuiter la compétence des *data scientists* en automatisant la préparation des données. C'est ce que l'on appelle le *feature learning* ou *representation learning*. Cette approche permit au *deep learning* de devenir un fourre-tout capable de s'adapter à divers médias, que ce soit l'image, le son, la vidéo, le texte ou d'autres capteurs en temps réel.

Pour comprendre cela, faisons un très rapide point sur les réseaux de neurones artificiels. Ces réseaux sont une collection des neurones (généralement virtuels) interconnectés via des connexions dites *synaptiques*, un peu comme le sont les ordinateurs de l'Internet, ou comme le sont les neurones de nos cerveaux. Certains de ces neurones sont directement reliés à des capteurs de données, qu'il s'agisse de caméras, de microphones ou d'autres appareils de mesures. Ces neurones correspondent aux variables observables².

Un réseau de neurones va ensuite relier ces neurones observables à des neurones cachés. Ce procédé est d'ailleurs très similaire aux facteurs de confusion et de concision dont on a parlé au chapitre 13, qui nous aidaient à choisir entre l'hôpital et la clinique, ou à expliquer la corrélation entre les couleurs des poils d'un mouton. Autrement dit, une mastication de données par un neurone caché peut être interprétée par le calcul d'un concept abstrait et pertinent pour expliquer les corrélations des variables observées. Typiquement, un tel neurone caché pourra déterminer la présence ou l'absence de lignes directrices dans une image captée par les neurones observables. L'empilement de couches de neurones pourra alors permettre d'expliquer les corrélations des couches de neurones intermédiaires. Typiquement, les couches profondes du réseau de neurones infèrent des lignes directrices la présence ou l'absence de moutons à l'image.

Ce processus est d'ailleurs inspiré de notre cortex cérébral. En effet, les neurosciences ont découvert que nos cerveaux analysent ce que nos yeux voient grâce à une abstraction croissante de la vue. Au niveau de l'œil, les capteurs, appelés cônes et bâtonnets, détectent la luminosité et la couleur du rayon de lumière qui les a excités. Un informaticien parlerait de la luminosité et de la couleur de chaque pixel d'une image. C'est ce genre de données brutes que l'on assignera aux variables observables d'un réseau de neurones profond.

Puis, la deuxième couche de calcul mesure typiquement des corrélations entre pixels voisins. Chez l'humain, chaque neurone de cette deuxième couche lie quelques cônes et bâtonnets voisins dans l'œil. À titre d'exemple, il pourrait s'exciter lorsque ces cônes et bâtonnets voisins sont tous les deux excités, ou tous les deux éteints³.

²  Le deep learning | Science Étonnante | D. Louapre (2016)

³ Il s'agit là d'un exemple et je ne cherche pas à coller avec la réalité biologique.

La troisième couche va ensuite combiner ces corrélations pour déterminer des lignes directrices des images que les yeux voient. La quatrième couche va ensuite combiner ces lignes directrices pour déterminer des objets élémentaires de l'image, comme les oreilles, les yeux ou les pattes d'un mouton. Les couches suivantes vont ensuite combiner ces objets élémentaires pour déterminer des structures plus profondes, comme la présence d'un mouton.

De nos jours, l'état de l'art en intelligence artificielle correspond à des réseaux de neurones dits de *convolution*, appelés *convolution neural networks* (CNN) en anglais (car ils sont reliés à une opération mathématique appelée produit de convolution). Ces réseaux sont inspirés de l'architecture grossière du cortex visuel. Devinez quoi. Alors qu'on a avant tout cherché à rendre le réseau de neurones globalement performant, le réseau de neurones, une fois entraîné à l'aide de photographies, calcule des niveaux d'abstraction croissante des images, à l'instar du réseau de neurones humain. Il semble que la performance d'une analyse d'image repose sur cet empilement de niveaux d'abstraction que seuls les réseaux suffisamment profonds permettent.

La représentation vectorielle des mots

Un phénomène similaire a été découvert empiriquement par le traitement du langage naturel. L'une des difficultés est le très grand nombre de mots des langues naturelles. Plutôt que d'avoir un neurone dédié à la compréhension de chaque mot, il est souvent préférable d'interpréter chaque mot en relation avec d'autres mots. En langage mathématique, étrangement, ceci peut joliment se faire en plongeant l'espace des mots dans un espace vectoriel de grande dimension, typiquement de dimension 50 à 100. En langage moins obscur, chaque mot va correspondre à une certaine combinaison d'activations des différents neurones.

De façon remarquable, en 2013, un groupe de chercheurs à Google autour de Tomas Mikolov réussit à faire apprendre à un réseau de neurones une telle représentation neuronale des mots de la langue anglaise. Ils appellèrent cette représentation *word2vec*. Ainsi, cette représentation transforme tout mot anglais en un vecteur dans un espace de grande dimension. Voilà qui permet de compactifier et structurer la représentation de l'ensemble des mots. En effet, d'un point de vue informationnel, la représentation vectorielle des mots est plus compacte. Elle prend moins de bits pour être décrite.

Mais il y a mieux ! La représentation vectorielle des mots permet surtout de révéler de nombreuses relations entre les mots. Ainsi, l'une des grandes découvertes du groupe de recherche de Mikolov fut que l'addition vectorielle entre les mots correspondait à l'intuition. Par exemple, ils découvrirent que si l'on effectue l'addition vectorielle *roi – homme + femme*, on obtient approximativement la représentation vectorielle du mot *reine*⁴ !

⁴  *Machine Reading with Word Vectors* | ZettaBytes | M. Jaggi (2017)

Plus fascinant encore, ce phénomène a aussi été observé chez les *Generative Adversarial Networks* (GANs). Prenez des images d'individus avec des lunettes de soleil et des individus sans lunettes de soleil. À l'aide d'un GAN, calculez la moyenne des vecteurs d'individus à lunettes, et la moyenne des vecteurs d'individus sans lunettes. Prenez ensuite la différence entre ces deux vecteurs. Vous obtenez alors un vecteur *lunettes* qui, intuitivement, correspond à avoir des lunettes de soleil. Prenez maintenant une image d'un individu sans lunettes. Calculez la représentation vectorielle *individu* de cette image, et ajoutez à cette représentation vectorielle le vecteur *lunettes*. Vous obtenez alors un nouveau vecteur *individu+lunettes*. Utilisez maintenant le GAN pour générer une image qui correspond à ce vecteur. De façon stupéfiante, vous obtiendrez alors une image de l'individu initial où celui-ci porte des lunettes⁵ ! Incroyable !

Cette étrange compatibilité entre l'addition vectorielle des concepts abstraits des réseaux de neurones et la combinaison intuitive des concepts selon nos cerveaux est, à ma connaissance, encore très mal comprise. Il s'agit d'un merveilleux mystère de la recherche actuelle en intelligence artificielle — je parierais qu'il sera résolu dans les années à venir, ce que j'attends avec impatience !

Dès lors, pour lire des livres physiques, un réseau de neurones pourrait consister en un réseau de neurones profonds dont les premières couches forment un réseau de neurones de convolution qui reconnaît les caractères d'une image. Puis, au-dessus de ces couches dédiées à la vision, on trouverait une couche de neurones regroupant les caractères en mots. Ensuite, une autre couche transformeraient ces mots en vecteurs, typiquement comme le ferait word2vec. Enfin, d'autres couches interpréteraient les vecteurs qui représentent les mots pour les relier à d'autres concepts issus d'autres analyses d'images ou autres.

En particulier, un réseau de neurones profonds performant doit pouvoir activer les mêmes neurones s'il voit une photographie d'un chat ou s'il lit le mot « chat ». Pour ce faire, la pré-mastication des données brutes semble cruciale. Plus généralement, il semble que la profondeur soit indispensable à la synthétisation de données brutes, que ce soit pour réduire la taille massive des données brutes et permettre des calculs en temps raisonnable, ou pour révéler des explications pertinentes des corrélations dans les données brutes.

L'expressivité exponentielle*

Le 16 juin 2016, une collaboration entre chercheurs de Stanford, de Cornell et de Google Brain mit en ligne sur ArXiV deux articles époustouflants. Ces toutes autres explications du succès de l'abstraction m'ont violemment séduit. Dans ce livre, je ne vais prendre le temps que de décrire le plus séduisant, à mon sens, de ces deux articles. Cet article est intitulé *Exponential expressivity in deep*

⁵  *Unsupervised representation learning with deep convolutional generative adversarial networks* | A. Radford, L. Metz et S. Chintala (2016)

neural networks through transient chaos. Il combine de nombreux concepts mathématiques sophistiqués, profonds et peu connectés, comme la théorie du chaos, la théorie des champs moyens et la courbure géométrique.

Allons-y doucement. L'article commence par considérer qu'un réseau de neurones n'est rien d'autre qu'une fonction mathématique compliquée. Cette fonction prend en entrée une collection de données mesurées et la transforme en une combinaison de concepts profonds. La collection de données mesurées forme mathématiquement ce que l'on appelle un vecteur, que l'on peut voir comme un point dans un espace de très grande dimension. Il en va de même pour la combinaison de concepts profonds, qui est lui aussi un point dans un espace de très grande dimension.

Dès lors, le réseau de neurones peut être vu comme une transformation géométrique du premier espace dans le second. La question que l'article se pose est la suivante : comment un réseau de neurones « typique » déforme-t-il l'espace ? Est-ce qu'un réseau de neurones « pris au hasard » bouge en moyenne les points dans tous les sens ? Ou conservera-t-il les propriétés géométriques des courbes qu'il transforme ?

La réponse intrigante de l'article est une complexification exponentielle des structures géométriques avec la profondeur du réseau de neurones. Plus précisément, l'article s'intéresse à une certaine mesure appelée *courbure globale* des figures géométriques⁶. Le cercle, intuitivement, est la figure fermée la moins courbée puisqu'il se courbe uniquement pour reboucler sur lui-même. D'ailleurs, la courbure globale d'un cercle est toujours égale à $\tau \approx 6,28$, quelle que soit sa taille. À l'inverse, une courbe très alambiquée qui oscillerait dans toutes les dimensions de l'espace aurait une très grande courbure globale.

Ce que montre l'article, c'est que, pour des réseaux de neurones aléatoires suffisamment « agités », la courbure d'une figure géométrique croît polynomialement avec la largeur du réseau, mais exponentiellement avec sa profondeur. Autrement dit, la profondeur permet aux réseaux de neurones de bien plus rapidement complexifier leurs transformations géométriques que la largeur.

En particulier, ce que ceci montre, c'est que la profondeur est cruciale pour la détection et l'analyse de structures fractales, c'est-à-dire dont le comportement n'est pas toujours lisse et régulier. Or, les structures fractales semblent omniprésentes dans le monde qui nous entoure⁷, que ce soit en biologie, en cosmologie ou en finance !

De la même manière, même si ça peut paraître abstrait et farfelu, il y a de bonnes chances que l'ensemble des images contenant un chat soit une structure fractale dans l'immense ensemble de toutes les images. La profondeur serait alors clé pour la détection de chats, tout comme pour de nombreuses tâches plus sophistiquées.

⁶La *courbure globale* intègre la norme des variations du vecteur unitaire.

⁷  *Les fractales* | MicMaths | M. Launay (2015)

L'émergence de la complexité

Cette découverte stupéfiante a résonné avec d'autres lectures auxquelles je fus exposé à peu près au même moment, notamment celle d'un article⁸ co-écrit par Sean Carroll, Scott Aaronson et Lauren Ouellette, sur la complexification temporaire de l'univers. Le principe physique que cet article cherchait à révéler est un phénomène que l'on observe à la fois à l'échelle du cosmos, du vivant et de la tasse de café au lait. À l'origine, ces structures étaient simples. L'univers est un plasma quasi-homogène, la tasse était une superposition d'une couche de café et d'une couche de lait et le vivant n'existe pas. Ces structures étaient aussi à faible entropie, une notion que l'on a introduite au chapitre 15.



Figure 18.1. Fractales dans le café au lait. Par Pexels sur Pixabay.

Or si l'entropie est faible à l'origine, elle ne peut qu'augmenter avec le temps⁹. Ce principe est la seconde loi de la thermodynamique. Or l'augmentation de l'entropie peut être interprétée comme une homogénéisation. À très long terme, la tasse sera un mélange parfait de café et de lait, le vivant aura disparu et l'univers disparaîtra dans un vide intersidéral parfaitement homogène (appelé le *Big Chill*). Cependant, Sean Carroll eut l'idée intuitive selon laquelle, entre les instants initiaux et finaux, toutes ces structures, le cosmos, le vivant et le café au lait, passent nécessairement par des phases de grande complexité, que ce soit la structuration en galaxies et filaments cosmiques que nous observons aujourd'hui, l'extrême complexité des végétaux, des animaux et de nos cerveaux humains, ou l'étrange phase où le lait forme des figures fractales dans le café.

Sean Carroll fut alors épaulé par l'informaticien Scott Aaronson pour formaliser cette notion intuitive. La première étape de cette formalisation fut une digitalisation des phénomènes physiques. Tout pouvant être ramené à une série de 0 et de 1 (par exemple en encodant l'image du café au format PNG ou JPG), Carroll et Aaronson, ensuite rejoints par Ouellette, cherchèrent une formalisation

⁸ Quantifying the Rise and Fall of Complexity in Closed Systems: The Coffee Automaton | S. Aaronson, S. Carroll et L. Ouellette (2014)

⁹ L'argument est plus subtile que cela, mais je ne m'étendrai pas sur ce sujet

de la complexité algorithmique de suites binaires finies x qui soit cohérente avec l'intuition de Carroll.

Les trois chercheurs présentèrent alors 4 définitions déjà introduites dans la littérature scientifique : la sophistication, la *complexité apparente*, la profondeur logique et la complexité du cône de lumière (sur laquelle je ne reviendrai pas ici). De façon intrigante, toutes ces définitions sont en fait subtilement reliées les unes aux autres.

La sophistication de Kolmogorov*

Arrêtons-nous d'abord sur la sophistication. La sophistication est une notion proposée par Kolmogorov, et s'appuie sur la complexité de Solomonoff. On pourrait d'ailleurs croire que cette complexité de Solomonoff serait une bonne candidate pour mesurer la complexité d'une suite binaire finie. Malheureusement, cette complexité est en fait maximale lorsque cette suite est si aléatoire qu'elle ne possède aucune régularité. Or de tels états sans régularité semblent davantage correspondre à une entropie maximale qui, dans le cas du cosmos, de la vie et du café, correspond à un état simple à décrire : il s'agit alors d'un néant homogène.

À l'instar de Boltzmann près d'un siècle avant lui, Kolmogorov eut la brillante idée de séparer autant que possible les structures « macroscopiques » aisément descriptibles, des structures « microscopiques » qui ne peuvent pas être mieux décrites que par un bruit parfaitement aléatoire. C'est cette intuition que la sophistication de Kolmogorov formalise.

Appelons S la structure macroscopique. De manière grossière, dire que les structures microscopiques de la suite binaire x sont suffisamment aléatoires revient à dire que la complexité de Solomonoff de la suite binaire x sachant S est (quasiment) maximale parmi l'ensemble des suites binaires qui ont la structure macroscopique S . Autrement dit, on considère que x est une instance typique de S , et qu'on peut le décrire précisément comme étant S plus un bruit quasi-uniformément aléatoire. Et pour que cette description macroscopique S soit « valide », il faut que la description de x en $S + \text{bruit}$ soit quasiment aussi compacte que la description la plus concise de¹⁰ x . Toute structure macroscopique ayant cette propriété est alors conforme aux prérequis de Kolmogorov.

Kolmogorov définit ensuite la sophistication de la suite binaire x comme étant la plus faible complexité de Solomonoff parmi les structures macroscopiques S

¹⁰Appelons x la suite binaire, et identifions S avec l'ensemble des suites binaires compatibles avec la structure macroscopique S . Dire que x est suffisamment « aléatoire » comme élément de S revient à dire que l'excès de complexité de Solomonoff $K(x) - K(S)$ est quasiment égal à la longueur de l'identification naïve des éléments de x parmi S , qui consiste à numérotter tous ces éléments. De façon formelle, ceci correspond à $K(x) - K(S) \geq \log_2 |S| - c$ avec un « petit » nombre c . Notez que ceci est légèrement différent d'une variante appelée *sophistication naïve*, où $K(x) - K(S)$ est remplacé par la complexité de Solomonoff conditionnelle $K(x|S)$.

conformes à ses prérequis. De manière intuitive, ceci correspond à la meilleure manière de décrire x comme étant une structure macroscopique simple plus un bruit (quasiment) uniformément aléatoire.

Malheureusement, à l'instar de la complexité de Solomonoff, la sophistication d'une suite binaire est en général incalculable. Pour estimer la sophistication des suites binaires de leurs simulations, Aaronson, Carroll et Ouellette se sont alors tournés vers une heuristique de la sophistication, qu'ils appellèrent *complexité apparente*. Le principe de cette heuristique est de se restreindre à des descriptions macroscopiques S qui correspondent à une sorte de lissage. Ce lissage est d'ailleurs inspiré des approches de Boltzmann qui, au lieu de considérer chaque particule individuellement, préférera décrire des propriétés statistiques grossières de l'ensemble des particules se trouvant approximativement au même lieu à un instant donné. C'est cette complexité apparente que les trois chercheurs uti-lisèrent pour analyser des simulations de café au lait¹¹.

La sophistication est un MAP de Solomonoff !*

Si la sophistication de Kolmogorov a une réelle esthétique mathématique et semble bel et bien reliée à l'intuition de notre description effective des objets physiques, on peut toutefois se demander s'il s'agit réellement de la bonne manière de faire. En particulier, n'est-ce pas quelque peu arbitraire que d'exiger de la suite x qu'elle soit quasi-maximalement aléatoire, sachant sa description macroscopique S ? Cela ne ressemblerait-il pas au cas de l'entropie de Boltzmann, qui a fini par être généralisée par un concept plus général et plus fondamental par Shannon ? Ne faudrait-il pas davantage invoquer un formalisme probabiliste pour toucher du doigt une notion plus profonde de sophistication ?

La *pure bayésienne* répond oui ! L'une des plus excitantes réflexions bayésiennes fut ma découverte de l'élégante notion de sophistication de Kolmogorov n'était qu'un maximum-a-posteriori (MAP) d'un sous-ensemble des théories du démon de Solomonoff !

Appelons T la description algorithmique de la structure macroscopique S . T n'est autre qu'une théorie prédictive à la Solomonoff qui tente d'expliquer la suite x . En particulier, la crédence en T sachant x se calcule via la formule de Bayes :

$$\mathbb{P}[T|x] = \frac{\mathbb{P}[x|T]\mathbb{P}[T]}{\mathbb{P}[x]}.$$

Or, on a vu que l'*a priori* $\mathbb{P}[T]$ sur une théorie à la Solomonoff était néces-

¹¹D'une certaine manière, ceci revient à remplacer la sophistication théorique qui a le défaut d'inclure dans son raisonnement des algorithmes aux temps de calcul déraisonnables par une sorte de sophistication « pragmatique » qui n'utilise qu'une poignée d'algorithmes rapides.

sairement exponentiellement faible en la complexité de Solomonoff¹² $K(T)$. Écrivons-le $\mathbb{P}[T] = \exp(-\alpha K(T))$. Pour déterminer le MAP du démon de Solomonoff, il nous suffit alors de maximiser le logarithme du numérateur, ce qui est équivalent à minimiser le négatif du logarithme du numérateur. On a :

$$\text{MAP}(x) = \arg \min_T \{\alpha K(T) - \ln \mathbb{P}[x|T]\}.$$

Devinez quoi ! Si l'on se restreint désormais à l'ensemble des théories T pour lesquelles $\mathbb{P}[\cdot|T]$ est une loi uniforme sur un ensemble de suites binaires, alors la théorie de la dualité¹³ montre qu'il existe justement une valeur de α pour laquelle $\text{MAP}(x)$ n'est autre qu'une description macroscopique optimale au sens de la sophistication de Kolmogorov¹⁴ !

En particulier, à l'instar de la manière dont la généralisation de l'entropie par Shannon nous a amené à étendre notre conception de l'information et de l'incertitude à des distributions non-uniformes, le démon de Solomonoff nous invite à généraliser la description macroscopique des données via la sophistication de Kolmogorov, à des descriptions pour lesquelles les incertitudes microscopiques sont non-uniformes ! En particulier, ce que cherche vraiment à mesurer la sophistication de Kolmogorov semble davantage être la complexité de Solomonoff des théories algorithmiques crédibles, étant donné des données x .

Bien entendu, le démon de Solomonoff nous dit d'aller plus loin que cela encore. Après tout, le MAP n'est qu'une grossière approximation de la formule de Bayes. Ainsi, au lieu de se restreindre à une seule description macroscopique qu'est le MAP, le démon de Solomonoff nous invite à considérer l'ensemble de toutes les descriptions macroscopiques, et à les pondérer conformément à leurs niveaux de crédence adéquats. Autrement dit, il nous invite à introduire la notion de *sophistication de Solomonoff* comme étant l'espérance des complexités de Solomonoff des théories crédibles *a posteriori*. Autrement dit, je propose de définir :

$$\text{sophistication de Solomonoff}(x) = \mathbb{E}_T[K(T)|x].$$

À ma connaissance, cette quantité n'a encore jamais été étudiée.

¹²En fait, ici, je considère plutôt que $K(T)$ est la longueur de la description algorithmique de T , pas celle de sa compression optimale.

¹³En effet, le Lagrangien associé au calcul de la sophistication est $\mathcal{L}_c(S, \mu) = (1 + \mu)K(S) + \mu \log_2 |S| - \mu c - \mu K(x)$. Par dualité forte de la programmation linéaire, on en déduit l'existence de $\mu^* \geq 0$ tel que la sophistication revient à trouver une distribution sur les S qui maximisent $\frac{(1+\mu^*)}{\mu^*} \ln^2 K(S) + \ln |S|$. Or, il s'agit là exactement de l'équation du MAP de Solomonoff avec $\alpha = (1 + \mu^*)(\ln 2)/\mu^*$ et des théories T qui sont des lois uniformes sur S .

¹⁴On voit d'ailleurs que la subjectivité de c dans la sophistication de Kolmogorov correspond à la subjectivité du préjugé de Solomonoff (et en particulier son facteur d'escompte α de la complexité de Solomonoff des théories prédictives).

En particulier, la remarque intuitive de Carroll, Aaronson et Ouellette semble concerner l'étonnamment grande sophistication de Solomonoff de l'état physique actuel de notre univers. Si Leibniz s'étonnait de l'existence de quelque chose plutôt que rien, je ne peux m'empêcher de m'étonner de l'existence d'une grande sophistication de Solomonoff plutôt qu'une petite. Ou pour faire référence à Descartes, outre le fait que je pense, il y a bien autre chose d'indubitable : l'existence d'une grande sophistication de Solomonoff. Tel est le mystère qui me semble être le plus fascinant de l'univers. Et son explication pourrait donc résider dans la tout aussi mystérieuse seconde loi de la thermodynamique...

La profondeur logique de Bennett

Une autre définition attira également mon attention : la profondeur logique de Bennett. De façon grossière, cette profondeur logique mesure le temps de calculs requis pour calculer une structure observée.

Ainsi, pour Bennett, les cafés au lait initial et final sont tout deux « peu profonds », car ils peuvent être calculés très rapidement par un algorithme. Dans le premier cas, l'algorithme dit : blanc en haut, noir en bas. Dans le second cas, le seul algorithme¹⁵ capable de calculer les positions des particules de café et de lait doit posséder en lui-même toutes les informations de toutes ces particules. Mais alors, puisque ces positions sont dans la mémoire de l'algorithme, il lui suffit de lire sa mémoire pour donner ces positions, ce qui ne requiert pas beaucoup de temps¹⁶. D'une certaine façon, ces deux situations n'ont aucune subtilité calculatoire.

En revanche, le café au lait en train d'être mélangé présente des structures complexes qui semblent pouvoir être décrites par un algorithme relativement succinct. Cet algorithme devra toutefois procéder à de longs calculs pour déterminer les positions des particules de café. L'état intermédiaire du mélange entre café et lait aurait donc une grande profondeur logique de Bennett. Aaronson, Carroll et Ouellette suggèrent que cette grande profondeur logique d'un état intermédiaire n'est pas spécifique à leur tasse de café ; l'univers tout entier serait dans un état intermédiaire d'une énorme profondeur logique.

En 2018, Rachid Guerraoui et moi avons utilisé cet argument pour tenter de justifier le succès du *deep learning*¹⁷. On a commencé par remarquer que la quasi-totalité des algorithmes de *machine learning* étaient des algorithmes très

¹⁵Ou plutôt, il s'agit de l'algorithme à plus faible complexité de Solomonoff.

¹⁶En fait, si le monde est déterministe, il suffit de laisser tourner la simulation du monde à partir de l'état initial, ce qui pourrait se faire avec peu de bits si cet état initial est simple à décrire. Cependant, on peut arguer que, sachant l'état final uniquement, cet état initial est impossible à déterminer en temps raisonnable. Du coup, la profondeur logique « pragmatique » reste faible, car le plus court algorithmique rapidement identifiable qui génère l'état final est celui qui connaît cet état final.

¹⁷Deep Learning Works in Practice. But Does it Work in Theory? L.N. Hoang et R. Guerraoui (2018)

rapides. D'ailleurs, l'ensemble des réseaux de neurones peu profonds forme justement l'ensemble des algorithmes rapides (une fois parallélisés). Si l'on croit que l'univers qui nous entoure et les données que l'on en collecte ont une grande profondeur logique (non-parallélisable¹⁸), on en vient à conclure que la faiblesse inévitable de ces algorithmes peu profonds est leur rapidité !

Ajoutons à cela la complexité de Solomonoff des données brutes dont on a déjà parlé au chapitre 14. Nos problèmes de prédiction semblent alors nécessairement requérir à la fois de nombreux paramètres et des temps de calcul suffisamment longs. Or, c'est précisément ce que les réseaux de neurones profonds proposent ! En plus de cela, les réseaux de neurones ont une structure qui leur permet d'apprendre en temps réel, via notamment la descente de gradient stochastique (SGD) dont on a parlé au chapitre 17. Voilà qui explique que les réseaux de neurones profonds atteignent des performances inégalées, ni par les algorithmes développés par des humains, ni par les architectures de *machine learning* alternatives aux calculs trop rapides.

En particulier, si la profondeur logique est nécessaire pour modéliser le monde et résoudre les problèmes de l'intelligence artificielle, alors les étapes intermédiaires de calcul semblent indispensables. Ces étapes de calculs intermédiaires, notamment lorsqu'ils sont récurrents, correspondent alors à des variables cachées. Et plus les calculs sont longs, plus ces étapes intermédiaires ressemblent à des variables profondes — et donc abstraites.

La profondeur logique de notre univers semble donc la clé pour expliquer la déraisonnable efficacité de l'abstraction.

La profondeur des mathématiques

Si le *deep learning* est en train de devenir le sommet de l'abstraction en *machine learning*, il n'est toutefois qu'une petite colline à côté des montagnes que sont les mathématiques. De tous les édifices humains, les mathématiques sont de loin nos créations les plus abstraites et les plus profondes. Des milliers de livres s'ajoutent ainsi pour aller toujours plus dans cette abstraction, à tel point que même les meilleurs d'entre nous luttent sérieusement à se plonger dans la création abstraite d'autres.

Une poignée d'équations peut requérir des années, voire des décennies de méditations, pour qu'une fraction de ses secrets soient révélés. Certains parmi les plus grands mathématiciens ont d'ailleurs passé une grosse portion de leurs carrières à s'attarder à une seule équation. Villani raconte : « l'équation de Boltzmann, la plus belle équation du monde [...] ! Je suis tombé dedans quand j'étais petit,

¹⁸ La parallélisabilité des algorithmes est d'ailleurs reliée au problème ouvert *P* versus *NC* de l'informatique théorique, qui demande si tout algorithme polynomial peut être parallélisé en un algorithme polylogarithmique avec un nombre polynomial de machines à calculer. Comme la majorité des informaticiens, nous conjecturons que la réponse est non.

c'est-à-dire pendant ma thèse. » Dirac aurait dit de l'équation qui porte son nom qu'elle était bien plus intelligente que lui qui, notamment lorsqu'il était encore jeune, n'en mesurait aucunement les implications physiques¹⁹. Quant à moi, j'espère que dans ce livre, j'aurai réussi à partager ma fascination pour la formule de Bayes et ses incroyables conséquences inattendues qui m'auront tenu en haleine depuis deux ans — et risquent fort de continuer à me fasciner pendant de longues années !

En fait, les mathématiques sont bien trop profondes pour être suivies pas à pas par notre cortex cérébral limité. Pour prendre la mesure des objets mathématiques, il nous faut constamment en trouver des interprétations grossières. Il nous faut imaginer une flèche pour raisonner sur un vecteur, il nous faut penser à une toile déformée pour réfléchir à la géométrie non-euclidienne et il nous faut nous arrêter aux propriétés connues des nombres premiers pour démontrer des théorèmes à leurs sujets.

Et bien souvent, quand on est confronté à des empilements d'étapes de calculs au milieu de nos raisonnements mathématiques, dans un premier temps, il peut être souhaitable d'abandonner tout effort de réflexion et de ne suivre que mécaniquement les règles de calcul pour en arriver à bout. *Shut up and compute*, disait David Mermin pour résumer l'interprétation de Copenhague de la mécanique quantique. On pourrait croire qu'il s'agit là de la mauvaise approche des sciences. Ne cherche-t-on pas à *comprendre* le monde autour de nous ? Si tel est l'objectif, il y a en effet de quoi rejeter l'abstraction mathématique excessive.

Mais le rôle de la formule de Bayes n'est pas d'adapter les théories crédibles aux capacités cognitives du cerveau humain. Son objectif est la prédition. Et si l'univers a une grande profondeur logique, les meilleures prédictions auront de bonnes chances de requérir un très grand nombre d'étapes de raisonnement ; et parce que ces étapes de raisonnement correspondent à des calculs profonds, elles échapperont nécessairement à notre intuition.

En particulier, la profondeur des mathématiques est inégalable par des raisonnements intuitifs. Après tout, notre intuition ne semble capable que de calculs rapides. Un raisonnement intuitif n'a donc pas de grande profondeur logique. Voilà, je pense, l'une des principales explications à la déraisonnable efficacité des mathématiques. Autrement dit, cette efficacité ne viendrait pas de la nature mathématique de l'univers (j'ai personnellement du mal à comprendre cette notion). Elle viendrait de la profondeur logique de l'état physique actuel de notre univers, et en particulier de la présence de phénomènes à grandes profondeurs logiques et faibles sophistications de Solomonoff. Et de nos limites cognitives.

¹⁹  *Anti-Matter and Quantum Relativity* | PBS Space Time | M. O'Dowd (2017)

La concision des mathématiques

Une autre explication à la déraisonnable efficacité des mathématiques est leur incroyable concision. Après tout, l'écrasante majorité de ce livre peut se résumer en la formule de Bayes, laquelle tient en une poignée de caractères. Autrement dit, ce livre admet une description remarquablement plus concise que le livre lui-même. Il est rempli de redondance. Sa sophistication de Solomonoff est relativement faible. D'ailleurs, j'ose espérer que quiconque, méditant suffisamment longtemps la Nature d'un bon apprentissage et cherchant à optimiser l'aspect pédagogique de son enseignement, aurait écrit un livre relativement semblable à celui que vous lisez. J'espère ne dire, à ses yeux, que des banalités formidableness compressibles — mais pédagogiquement instructives.

L'une des plus grandes avancées des mathématiques, que l'on peut attribuer notamment à Al-Khwarizmi, fut la synthétisation du langage mathématique. Mais ce n'est pas tout. En plus d'être concis, le langage d'Al-Khwarizmi est extrêmement mécanique à lire. Il n'y a pas trente-six mille interprétations possibles, et il ne faut pas longuement réfléchir au sens de chacun des symboles de ce langage²⁰. En fait, pour déterminer la validité d'une preuve formelle, il suffit de la lire bêtement (mais avec beaucoup d'attention). En termes informatiques, cette lecture requiert un algorithme à faible complexité de Solomonoff, même si son temps de calcul sera potentiellement long.

L'un des exemples les plus frappants de l'effort de concision des mathématiques est le cas des équations de l'électrodynamique. Quand le physicien James Maxwell les introduisit en 1861, ces équations n'avaient rien de concis. Cependant, l'abstraction croissante des mathématiques a su réduire la longue description de ces équations en une poignée de symboles : $\mathcal{L} = -\frac{1}{4\mu_0}F^{\alpha\beta}F_{\alpha\beta} - A_\alpha J^\alpha$, avec $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. Bien entendu, effectuer une prédiction à partir de ces équations nécessitera tout un arsenal algorithmique ; mais cet arsenal, si on le restreint à l'aspect purement calculatoire, n'est pas si long à décrire.

Ceci contraste fortement avec les théories non formelles. Celles-ci reposent fortement sur une interprétation de la langue et le « bon sens » que nous autres humains possédons. Or la description algorithmique de la langue et du bon sens a de bonnes chances de requérir des milliards de lignes de code pour approcher les performances humaines. Comme on l'a vu au chapitre 14, c'est cela qui, selon Turing, explique la nécessité du *machine learning* pour acquérir la maîtrise de la langue et du bon sens. Par conséquent, les théories non formelles pèchent, non pas tant parce qu'elles sont imprécises, mais parce qu'elles requièrent des algorithmes à grande complexité de Solomonoff (nos cerveaux) pour devenir prédictives. Or, si l'on en croit le préjugé de Solomonoff, les théories à grande complexité de Solomonoff sont exponentiellement improbables *a priori*.

Bien entendu, le langage naturel et son interprétation par nos cerveaux humains n'ont rien d'arbitraire. La sélection naturelle a favorisé les langues et les

²⁰  L'émergence de l'intelligence | IA 5 | Science4All | T. Cabaret et L.N. Hoang (2018)

processus cognitifs capables de prédire l'environnement et les relations sociales des tribus primitives. Cependant, cette sélection n'a pas favorisé les langues et processus cognitifs à même de décrire la physique des particules, l'économie des marchés mondialisés ou l'impact des nouvelles technologies. Pour ces problèmes, il n'est pas étonnant que des approches mathématiques, même simplistes, gagnent les faveurs des préjugés de la *pure bayésienne*.

En particulier, l'élégance mathématique semble nécessairement conduire les mathématiciens à bien explorer et comprendre l'ensemble des algorithmes concis, ceux dont les crédences *a priori* sont grandes, selon le modèle de Solomonoff. Ce n'est alors pas étonnant de constater que les meilleures théories prédictives fondées sur le langage mathématique sont davantage crédibles *a posteriori* aussi.

La modularité des mathématiques

J'aimerais conclure ce chapitre avec une troisième et dernière explication à la déraisonnable efficacité des mathématiques, à savoir la modularité des mathématiques. Les théorèmes mathématiques élégants sont souvent à l'intersection de nombreuses sous-disciplines. Ils forment des ponts entre différentes perspectives. Ils sont des sortes de couteaux suisses qui, manipulés astucieusement, permettent de résoudre une large gamme de problèmes. C'est ainsi que, typiquement, les notions de dérivées, d'espaces vectoriels et de graphes sont omniprésentes en géométrie, en optimisation et en probabilité, mais aussi en physique, en informatique, en biologie, en chimie et en économie. Ces notions sont telles les bits, les structures de listes et les algorithmes de tri de l'informatique. Les théorèmes forment le socle des théories prédictives, de la même manière que les algorithmes élémentaires forment le socle de tout code source sophistiqué.

La raison pour laquelle les programmeurs décomposent les algorithmes ainsi est que ce socle est constamment utilisé et réutilisé à différents endroits dans un code plus général. De façon similaire, l'addition et la multiplication sont souvent utilisées encore et encore dans un modèle physique, de la même manière que la notion de dérivées est souvent appliquée à de nombreuses grandeurs physiques distinctes. Il est dès lors beaucoup plus simple et élégant de ne définir la dérivée qu'une seule fois de manière très abstraite et générale, plutôt que de la redéfinir à chaque fois qu'on va la réutiliser. Le langage mathématique permet alors d'étudier un grand nombre de modèles distincts sans avoir à constamment réinventer la roue.

En voici un exemple devenu incontournable depuis quelques décennies. Un très grand nombre de problèmes pratiques, en mathématiques, en *machine learning*, en physique des matériaux, ou encore en économie, peuvent s'écrire sous la forme de la minimisation d'un objectif sujette à diverses contraintes. Ce formalisme est celui de l'optimisation. Il unifie de nombreux domaines. Les outils qui décortiquent et résolvent ce formalisme, comme la descente de gradient, la

recherche locale ou les algorithmes génétiques, sont alors des couteaux suisses très souvent utiles pour adresser n'importe lequel des nombreux problèmes qui peuvent être modélisés par ce formalisme.

Le cas de la physique théorique est encore plus impressionnant. En particulier, la physique quantique des champs, loin d'être une seule théorie rigide, repose davantage sur un principe de *quantification d'une formule lagrangienne*²¹. En effet, depuis l'utilisation du principe de moindre action par Richard Feynman, les physiciens ont pris l'habitude de définir leurs théories quantiques par une seule et unique formule, dite lagrangien, typiquement de la forme $\mathcal{L} = i\hbar\psi\gamma^\mu D_\mu\psi - \frac{1}{2}\text{Tr}(F^{\mu\nu}F_{\mu\nu})$. Quelle que soit la valeur exacte du lagrangien, les physiciens vont ensuite pouvoir utiliser des méthodes systématiques pour transformer ce lagrangien en une équation à dérivées partielles du mouvement (appelée équation d'Euler-Lagrange). Puis, ces équations pourront être *quantifiées*, avant que des prédictions de cette quantification soient alors déduites des équations. Autrement dit, transformer la formule lagrangienne en un ensemble de prédictions n'est qu'un simple (mais long) calcul.

Mieux encore, la théorie de jauge permet même de déduire la formule exacte du lagrangien, à partir uniquement des symétries du lagrangien. C'est ainsi que de façon stupéfiante, la nature des objets physiques et de leurs interactions s'est réduite à l'étude abstraite de groupes de symétries. Ainsi, le théorème de Noether déduit la conservation de l'énergie de la symétrie du lagrangien par translation dans le temps, et la conservation de la quantité de mouvement de sa symétrie par translation dans l'espace²². Mieux encore, une toute nouvelle théorie quantique des champs peut être construite, simplement en postulant par exemple que le lagrangien est invariant par action d'un groupe, disons, $SU(5)$. Voilà qui est absolument remarquable ! La physique théorique a réussi à se détacher de ses objets élémentaires, comme les photons et les électrons, pour ne s'intéresser qu'à des concepts incroyablement abstraits, comme le groupe de symétrie du lagrangien.

En fait, c'est en se restreignant à ce formalisme étrange que deux des plus grandes découvertes théoriques de la physique moderne ont su devancer de loin les résultats expérimentaux. Ainsi, en 1964, Murray Gell-Man et George Zweig postulèrent indépendamment que le lagrangien était invariant par action du groupe $SU(3)$. Ils découvrirent que cette symétrie impliquait la sécabilité des protons et neutrons en particules encore plus élémentaires, appelées quarks. Après des décennies de travaux théoriques, de découvertes expérimentales et de controverses, le modèle de Gell-Man et Zweig finit par être accepté. Il fait désormais partie du modèle standard de la physique des particules. Cependant, à ce moment, Gell-Man avait déjà reçu un prix Nobel pour d'autres travaux. Or, le comité du prix Nobel ne voulut pas récompenser Zweig sans récompenser Gell-Man une deuxième fois, et il ne voulait pas récompenser Gell-Man une deuxième fois. Zweig ne reçut jamais le prix Nobel.

²¹  *La théorie des cordes* | Science Étonnante | D. Louapre (2015)

²²  *The most beautiful idea in physics - Noether's Theorem* | Looking Glass Univers (2015)

Plus surprenant encore, en 1964 toujours, trois groupes de physiciens (François Englert et Robert Brout ; Peter Higgs ; Gerald Guralnik, Carl Hagen et Tom Kibble) découvrirent indépendamment que le formalisme relativiste de la formulation lagrangienne était incompatible avec l'existence de particules massives. Pour sauver le formalisme lagrangien, les six physiciens introduisirent ainsi un nouveau champ quantique, aujourd'hui appelé champ de Higgs, dont la formulation lagrangienne respecte les symétries dites de *jauge*, mais dont l'état physique brise ces symétries. De façon remarquable, l'interaction des particules classiques avec un champ de Higgs à symétrie brisée est alors indiscernable du cas où ces particules ont des masses ! Mieux encore, la quantification du champ de Higgs et son excitation ont conduit ces chercheurs à prédire une nouvelle particule, appelée boson de Higgs²³. Comme vous le savez sans doute, ce boson de Higgs fut découvert expérimentalement par le LHC du CERN en 2012. L'année suivante, Higgs et Englert reçurent le prix Nobel.

L'abstraction avait encore gagné. C'est forcément un coup de chance. Mais, au vu de notre réflexion sur la sophistication de Solomonoff et la profondeur logique de Bennett, il semble que ce ne soit peut-être pas un coup de chance si improbable...

Références en français

- ➲ *Les Métamorphoses du calcul : Une étonnante histoire des mathématiques* | Le Pommier | G. Dowek (2007)
- ➲ *Complexité aléatoire et complexité organisée* | QUAE GIE | J.P. Delahaye (2009)

- ▶ *Les fractales* | MicMaths | M. Launay (2015)
- ▶ *Le deep learning* | Science Étonnante | D. Louapre (2016)
- ▶ *Jeu de go et intelligence artificielle* | À chaud | Science Étonnante | D. Louapre (2016)
- ▶ *La théorie des cordes* | Science Étonnante | D. Louapre (2015)
- ▶ *$E = mc^2$ et le boson de Higgs* | Science Étonnante | D. Louapre (2017)
- ▶ *Jeux et informatique* | Passe-Science | T. Cabaret (2016)
- ▶ *Peut-on coller les côtés opposés d'un carré ?* Relativité 9 | Science4All | L.N. Hoang (2016)
- ▶ *Top 8 des monstres mathématiques* | Infini 11 | Science4All | L.N. Hoang (2017)
- ▶ *L'émergence de l'intelligence* | IA 5 | Science4All | T. Cabaret et L.N. Hoang (2018)

- ▶ *Deep learning* | Podcast Science 228 | N. Tupégabet (2015)

²³ ➡ *$E = mc^2$ et le boson de Higgs* | Science Étonnante | D. Louapre (2017)

Références en anglais

- ✉ *The unreasonable effectiveness of mathematics in the natural sciences* | Communications on pure and Applied Mathematics | E. Wigner (1960)
- ✉ *Logical depth and physical complexity* | The Universal Turing Machine A Half-Century Survey | C. Bennett (1995)
- ✉ *Efficient estimation of word representations in vector space* | T. Mikolov, K. Chen, G. Corrado and J. Dean (2013)
- ✉ *Unsupervised representation learning with deep convolutional generative adversarial networks* | A. Radford, L. Metz and S. Chintala (2016)
- ✉ *Deep learning* | Nature | Y. LeCun, Y. Bengio et G. Hinton (2015)
- ✉ *Exponential expressivity in deep neural networks through transient chaos* | Advances In Neural Information Processing Systems | B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein and S. Ganguli (2016)
- ✉ *Quantifying the Rise and Fall of Complexity in Closed Systems: The Coffee Automaton* | S. Aaronson, S. Carroll and L. Ouellette (2014)
- ✉ *Deep Learning Works in Practice. But Does it Work in Theory?* L.N. Hoang and R. Guerraoui (2018)

- ▶ *Artificial Neural Networks in Machine Learning* | Wandida | E.M. El Mhamdi (2016)
- ▶ *The Rise of Machine Learning* | ZettaBytes | M. Jaggi (2017)
- ▶ *Machine Reading with Word Vectors* | ZettaBytes | M. Jaggi (2017)
- ▶ *4 Big Challenges in Machine Learning* | ZettaBytes | M. Jaggi (2017)
- ▶ *Google Cars versus Tesla* | ZettaBytes | B. Faltings (2017)
- ▶ *Achieving both Reliability and Learning in AI* | ZettaBytes | B. Faltings (2017)
- ▶ *The most beautiful idea in physics - Noether's Theorem* | Looking Glass Univers (2015)
- ▶ *Anti-Matter and Quantum Relativity* | PBS Space Time | M. O'Dowd (2017)

L'inférence bayésienne rend bien compte des processus de perception : étant donné des entrées ambiguës, notre cerveau en reconstruit l'interprétation la plus probable.

Stanislas Dehaene (1965-)

L'apprenant bayésien peut tirer beaucoup plus d'information sur l'extension d'un concept à partir d'un ensemble d'exemples observés [...] et peut utiliser cette information de manière rationnelle pour inférer la probabilité qu'un nouvel objet soit une instance de ce concept.

Joshua Tenenbaum (1972-)

19

Le cerveau bayésien

Le cerveau est incroyable

En septembre 2017, je pensais avoir fini une première version complète de ce livre. Je l'envoyai à Julien Fageot, un ami mathématicien. Julien s'empressa de me conseiller le cours du neuroscientifique Stanislas Dehaene au Collège de France intitulé *Le cerveau statisticien : la révolution bayésienne en sciences cognitives*. « Je pense que le cerveau bayésien mériterait son propre chapitre », ajouta Julien. Voilà qui m'embêtaît. Le livre me semblait déjà bien trop long.

Je me mis toutefois à écouter Stanislas Dehaene. Quel délice succulent ! En deux jours, je dévorai ses deux années de cours sur le sujet — je serais allé plus vite si je n'avais pas un job ! À chaque cours, j'étais tel un gamin découvrant un nouveau bonbon : je m'en léchais les babines ! Plus surprenant encore, malgré toutes mes créances en le bayésianisme, voire mon hooliganisme pro-bayésien, je ne cessais de me répéter : « mais ce n'est pas possible que la formule de Bayes soit aussi centrale à la cognition humaine ! » Tout extrémiste bayésien que je me croyais, il semblait que je ne fusse toujours pas suffisamment bayésien !

Pourtant, à bien y réfléchir, j'aurais dû m'y attendre. Si le bayésianisme est vraiment une forme d'apprentissage optimal, la sélection naturelle aurait nécessairement dû le sélectionner au moment de choisir les espèces intelligentes les plus à même de survivre et de se reproduire. C'est même une prédition du bayésianisme combiné à l'évolution darwinienne : si la formule de Bayes est vraiment la solution à tous nos maux épistémologiques, alors Dame Nature a

nécessairement trouvé des manières pragmatiques de l'approximer par des processus naturels. Et bien, depuis une quinzaine d'années, cette prédiction a été confirmée, encore et encore, par les sciences cognitives ! *Notre cerveau est un formidable calculateur de toutes sortes d'approximations de la formule de Bayes.*

Cette affirmation peut vous sembler troublante. Après tout, j'ai passé une grosse partie de ce livre à critiquer notre incapacité à appliquer la formule de Bayes à des cas simplistes comme le problème de Monty Hall. Je n'ai cessé de souligner l'inlassable excès de confiance qui accompagne notre inaptitude à comprendre le bayésianisme. Daniel Kahneman et Amos Tversky semblent avoir passé leurs carrières à le démontrer. Oui, nous ne comprenons pas la formule de Bayes. Oui, nous sommes incapables de l'appliquer à des problèmes mathématiques. Oui, nous sommes mauvais penseurs bayésiens.

Néanmoins, Dame Nature n'a pas sélectionné notre capacité à réfléchir à des problèmes abstraits ; elle a sélectionné notre capacité à nous adapter à notre environnement. Du coup, les inférences bayésiennes que nos cerveaux effectuent sont le traitement *inconscient* des données auxquelles nos sens nous exposent, surtout lorsque le traitement de ces données a pu être crucial à la survie dans le monde animal ou à notre compréhension de notre environnement social.

« Notre cerveau humain s'appuie sur des compétences anciennes dans l'évolution. Nous héritons de compétences et de représentations intuitives dans des domaines qui étaient et sont toujours très importants pour la survie de notre espèce. Donc tous les enfants naissent avec un concept d'espace, avec un concept de nombres, avec, probablement dans le cas de l'espèce humaine, un circuit spécialisé dans le langage », précise Stanislas Dehaene. « Évidemment l'Éducation cherche à dépasser ces connaissances. L'Éducation nous invite à concevoir des disciplines nouvelles, comme la lecture par exemple, l'écriture, l'arithmétique formelle symbolique, qui n'ont pas été anticipées par l'évolution. Mais nous allons *recycler* [...] des systèmes cérébraux anciens pour ces usages culturels nouveaux. »

C'est ce recyclage qui peut être défaillant et violer les lois des probabilités. Cependant, les processus préexistants, souvent très inconscients, et leurs applications pragmatiques semblent coïncider de manière bluffante avec des calculs bayésiens.

Montagne ou vallée ?

Ouvrez une carte topographique, c'est-à-dire une carte sur laquelle sont représentés les reliefs. Vous pouvez par exemple ouvrir Google Map sur vos téléphones, aller dans les options et activer le mode « relief ». Zoomez sur une chaîne de montagne ou une vallée, comme la vallée de Chamonix. Regardez maintenant la carte à l'envers. Si vous êtes sur votre téléphone, tournez-le sans toutefois tourner l'image dans votre téléphone. Quelque chose d'étrange devrait vous sauter aux yeux...

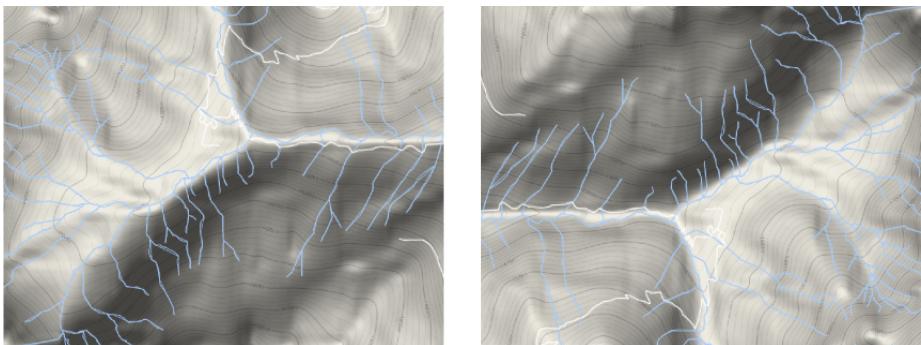


Figure 19.1. Les deux images ci-dessus sont les mêmes, l'une d'elles ayant subi une rotation d'un demi-tour. Ce qui y est représenté est-il une montagne ou une vallée ?

Vous devriez avoir l'impression que les montagnes se sont transformées en vallées, et que les vallées se sont transformées en montagnes ! En particulier, ce qui est, sur la carte, en haut d'une zone ombrée semble être une montagne, tandis que les vallées se trouvent en bas des zones ombrées.

Mais d'où vient cette perception ? Qu'est-ce qui fait que nous puissions distinguer les vallées et les montagnes dans une carte ?

Il se trouve que, comme pour beaucoup d'autres perceptions inconscientes, nos conclusions se déduisent d'un calcul bayésien. En particulier, le préjugé indispensable à notre interprétation des cartes est l'origine de l'éclairage. En effet, les ombres des cartes sont celles d'un éclairage venant du haut de la carte — bien que cet éclairage soit physiquement impossible dans des régions de l'hémisphère nord comme l'Europe ou les États-Unis, où le Soleil éclaire toujours depuis le sud¹ !

S'il est incorrect dans le cas des cartes, ce préjugé est en fait parfaitement justifié dans notre vie quotidienne. L'éclairage vient souvent d'en haut, que ce soit l'éclairage du soleil ou celui de nos lampes électriques. Du coup, quand on observe le visage des autres, on voit généralement le nez émerger au-dessus d'une zone ombrée, à l'inverse d'yeux se trouvant en dessous d'une zone ombrée. D'ailleurs, l'éclairage inverse peut paraître effrayant, ce qui explique pourquoi il est si souvent utilisé dans les films d'horreur.

De façon plus générale, notre cortex visuel possède la faculté remarquable et inconsciente de deviner l'origine de l'éclairage dans une image, pour ensuite mieux interpréter le contenu de l'image. Ce processus est ainsi très similaire aux réseaux bayésiens et aux machines de Boltzmann dont on a parlé dans un chapitre précédent : notre cerveau semble rapidement faire appel à des variables cachées pour comprendre les variables observées.

¹ The "Mountain Or Valley?" Illusion | Minute Physics (2017)

Les illusions optiques

Les préjugés que nous avons sur les éclairages, et qui nous conduisent à mal interpréter des cartes à l'envers, sont d'une efficacité redoutable dans notre analyse de nombreuses images du monde naturel. Cependant, nos préjugés peuvent aussi nous induire en erreur lorsqu'il s'agit d'images construites selon des règles peu usuelles dans le monde naturel.

Prenez deux stylos identiques. Disposez-les l'un à l'horizontal, l'autre juste au-dessus, à la verticale, pour former le symbole \perp que les mathématiciens utilisent pour désigner la perpendicularité. Regardez vos stylos. Vous devriez avoir l'impression que le stylo vertical est plus long que le stylo horizontal — alors que, par construction, ces deux stylos ont en fait la même longueur !

Là encore, cette illusion optique s'explique parfaitement en intégrant un préjugé justifié de notre cerveau bayésien. Habitué à voir des images en perspective, il lui arrive souvent de voir des lignes verticales qui correspondent à des lignes horizontales vues en perspective — typiquement les rails d'une ligne de train. Or la perspective va alors justement écraser ces lignes, si bien que ces lignes apparaissent plus courtes qu'elles ne le sont vraiment. Nos cerveaux bayésiens intègrent alors inconsciemment ce préjugé justifié sur la perspective. Voilà qui revient à inférer le fait que les lignes verticales sont plus longues qu'elles ne le paraissent. C'est sans doute pour cela que nous pensons instinctivement, à tort, que la ligne verticale de notre symbole \perp de perpendicularité est plus longue que la ligne horizontale.

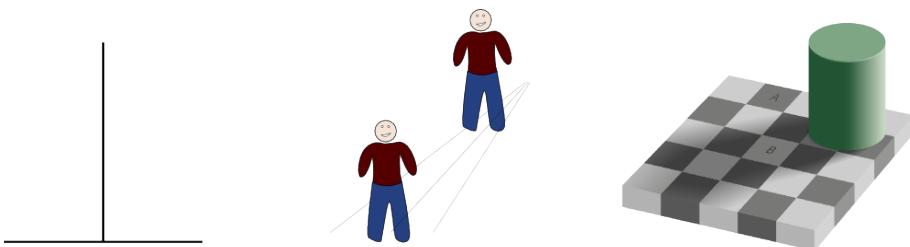


Figure 19.2. Quelques illusions classiques. À gauche, les deux segments ont la même longueur. Au milieu, les deux bonhommes ont la même taille. À droite, les cases A et B ont la même luminosité.

De tels effets de perspective sont utilisés dans d'autres illusions d'optique, où un personnage est typiquement copié-collé en premier et en second plan. Si le personnage au second plan nous paraît alors plus grand, c'est parce que nos cerveaux bayésiens ont inconsciemment appliqué (une approximation de) la formule de Bayes, pour en déduire que le personnage au second plan est très probablement bien plus grand que tel qu'il est représenté. L'illusion d'optique s'explique par une inférence bayésienne.

De la même manière, une autre illusion classique est un damier dont une partie est ombrée. Les cases claires ombrées nous semblent alors plus claires que les cases sombres illuminées. En fait, ceci n'est pas le cas. Néanmoins, nos cerveaux bayésiens effectuent inconsciemment des inférences bayésiennes qui tiennent compte de l'effet de l'éclairage, de façon à obtenir la conclusion plus *utile* selon laquelle les cases sombres illuminées sont plus sombres que les cases claires ombrées.

La perception du mouvement

Prenez un losange très aplati et inclinez-le pour que son axe principal soit légèrement selon la diagonale haut-droite et bas-gauche. Puis, déplacez le losange de gauche à droite. Quand le losange est bien contrasté par rapport au fond (typiquement noir sur blanc), on voit très bien le losange se mouvoir de gauche à droite. Cependant, étrangement, quand on diminue le contraste (gris clair sur blanc), quelque chose d'étrange survient. De nombreux sujets se mettent alors à voir le losange aller dans la direction bas-droite. J'ai moi-même reproduit l'expérience sur Twitter, et selon mon sondage², 39 % des 376 répondants se sont mis à voir le losange descendre !

Comment est-il possible que nos brillants cerveaux bayésiens en viennent à cette conclusion erronée ? Dans un article remarquable³, Weiss, Simoncelli et Adelson montrent que cette conclusion erronée est justement celle d'un cerveau bayésien dont le calcul intègre l'incertitude que cause la faiblesse du contraste.

Mais avant d'en venir aux explications bayésiennes des trois auteurs, arrêtons-nous sur une étonnante faculté de nos cerveaux que l'on a tendance à sous-estimer : quand un losange bien contrasté se déplace, on est capable de déterminer son mouvement. C'est une prouesse remarquable ! Après tout, d'un point de vue sensoriel, tout ce que l'on voit, ce sont des pixels s'allumer et s'éteindre. Comment fait-on pour traduire la dynamique de l'allumage des pixels d'une vidéo en une dynamique de l'objet représenté dans la vidéo ?

On l'a vu, notre cortex cérébral est avant tout capable de détecter les lignes directrices des images. Du coup, quand un losange se déplace, le cerveau voit surtout les bords du losange se déplacer. Sauf que chaque bord du losange est incliné. Du coup, quand le losange se déplace de gauche à droite, le bord du losange, lui, semble en fait se déplacer dans une autre direction. En fait, toute droite qui se déplace semble se déplacer dans une direction orthogonale à la droite. Pour des droites infinies, un tel déplacement est en fait indiscernable de tout autre mouvement de translation de la droite.

²https://twitter.com/science__4__all/status/911943074049396736

³ Motion illusions as optimal percepts | Nature Neuroscience | Y. Weiss, E. Simoncelli et E. Adelson (2002)

Néanmoins, en bon bayésien, notre cerveau sait qu'un déplacement de la droite dans une direction orthogonale à la droite n'est que l'une des explications possibles du mouvement de la droite. S'il s'agit sans doute *a priori* du déplacement le plus probable, d'autres mouvements ont une probabilité non nulle. Selon Weiss, Simoncelli et Adelson, le cerveau va alors combiner les déplacements probables de tous les bords du losange, et supposer que le losange en tant que tel n'a qu'un seul déplacement. Autrement dit, le cerveau va déterminer le déplacement le plus crédible *a posteriori*, étant donné les déplacements probables des différents bords. Voilà qui lui permet de déduire le mouvement du losange du mouvement des bords — quand bien même le mouvement de ces bords est incertain !

Voilà qui explique pourquoi le cerveau effectue les bonnes inférences lorsque le contraste est suffisamment élevé. Que se passe-t-il lorsque le contraste est faible ? Le cerveau voit alors mal le mouvement des lignes directrices. Il a une incertitude quant à la vitesse à laquelle ces lignes se déplacent (qui s'ajoute à l'incertitude sur la direction de déplacement de ces lignes). De façon étonnante, cette incertitude additionnelle conduit à un calcul bayésien différent. Quand celle-ci est suffisamment grande, l'inférence bayésienne conduit alors à préférer l'hypothèse d'un mouvement diagonal vers le bas et la droite — ce mouvement est alors le maximum *a posteriori*.

Incroyable ! La prédiction erronée du cerveau s'explique par un calcul bayésien du cerveau, lorsque celui-ci est soumis à une incertitude additionnelle à laquelle conduit la diminution du contraste ! Cette prédiction est mauvaise. Mais elle est mauvaise pour de bonnes raisons : il s'agissait de la manière optimale de gérer les incertitudes auxquelles le cerveau bayésien était soumis.

Et si le cadre expérimental très artificiel de cette expérience fut choisi pour piéger notre cerveau bayésien, la prédiction bayésienne est sans doute bien souvent beaucoup plus pertinente dans les cas empiriques !

Échantillonnage bayésien

L'un des phénomènes des sciences cognitives qui m'aura le plus impressionné est la capacité de nos cerveaux à effectuer des échantillonnages représentatifs, sans doute via *Markov-Chain Monte-Carlo* (MCMC), cette technique dont on a parlé au chapitre 17. Comme on l'a vu, l'échantillonnage est une approche souvent efficace pour décrire un phénomène probabiliste, notamment lorsque les lois de probabilité sont difficiles à décrire dans le langage mathématique. Or, la quasi-totalité des lois de probabilités sont justement, en pratique, trop difficiles à décrire et à manipuler directement.

Nos cerveaux bayésiens semblent l'avoir compris. Plutôt que de raisonner avec plusieurs théories incompatibles à la fois comme le ferait la *pure bayésienne*, nos cerveaux préfèrent raisonner séquentiellement, d'abord avec une théorie très crédible, puis, si le temps le permet, avec une autre théorie crédible.

C'est typiquement ainsi que l'on va chercher à interpréter des images ambiguës. Vous avez sans doute déjà vu cette image ambiguë qui, sous un angle, semble être l'image d'un canard, alors que, sous un autre angle, elle semble représenter un lapin. Le plus étrange, c'est qu'il semble impossible de voir les deux interprétations crédibles de l'image en même temps. Nos cerveaux bayésiens semblent capables de ne voir qu'une seule interprétation à la fois.

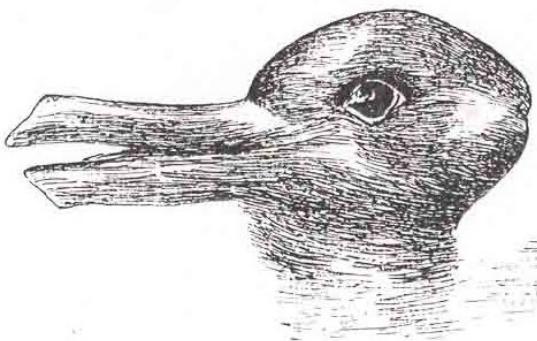


Figure 19.3. Canard ou lapin ?

En 2011, Moreno-Bote, Knill et Pouget⁴ étudièrent ce phénomène. Ils donnèrent à voir à des sujets deux grilles noires sur fond blanc se déplacer l'une par rapport à l'autre. On demande alors aux sujets laquelle des deux grilles est au-dessus de l'autre. Les sujets alternent alors entre les deux interprétations possibles. L'hypothèse du cerveau bayésien postule alors que la proportion de temps pendant lequel le sujet adopte une interprétation est sa crédence bayésienne en cette interprétation.

Pour tester cette hypothèse, le génie des trois chercheurs fut alors d'étudier l'effet de deux variables différentes sur ces proportions de temps consacrés aux deux interprétations, et de vérifier que la combinaison des deux effets était le produit de chaque effet. Les deux effets étudiés étaient le resserrement des droites de l'une des grilles et sa vitesse relative. De façon stupéfiante, la multiplication des effets que prédit l'hypothèse du cerveau bayésien coïncidait formidablement bien avec l'effet combiné des deux effets⁵ ! L'échantillonnage de nos cerveaux bayésiens semble parfaitement obéir aux lois des probabilités ! Mieux encore, Wong et Wang⁶ ont déterminé des mécanismes neuronaux crédibles pour justifier la capacité physique de nos neurones à implémenter un tel échantillonnage bayésien !

⁴ Bayesian sampling in visual perception | Proceedings of the National Academy of Sciences | R. Moreno-Bote, D. Knill, and A. Pouget (2011)

⁵ Ceci repose sur certaines hypothèses d'indépendance des effets.

⁶ A recurrent network mechanism of time integration in perceptual decisions | Journal of Neuroscience | K.F. Wong et X.J. Wang (2006)

Un étrange corollaire de cette conclusion est que nous pouvons améliorer nos prédictions en effectuant plusieurs prédictions correspondant à différentes interprétations. Souvenez-vous. La *pure bayésienne* améliore ses prédictions en prenant la moyenne des prédictions des modèles crédibles. Pour tester ce corollaire de l'hypothèse du cerveau bayésien qui échantillonne via MCMC, Vul et Pashler⁷ ont posé la question suivante à 428 sujets : « quel pourcentage des aéroports mondiaux se trouvent aux États-Unis ? » Ils leur ont demandé deux réponses à cette question. Les secondes réponses étaient souvent moins bonnes que les premières. Mais de façon étonnante, la moyenne des deux réponses étaient néanmoins nettement meilleure que la meilleure des deux réponses !

Mieux encore, pour la moitié des sujets, Vul et Pashler ont attendu 3 semaines avant de leur demander de fournir une seconde réponse, histoire de permettre aux sujets de vraiment changer d'interprétation du problème avant qu'ils ne répondent. Devinez quoi ! L'amélioration de la moyenne des réponses fut alors encore meilleure que dans le cas où les secondes réponses étaient données juste après les premières. « C'est pas mal de se poser soi-même la même question une deuxième fois », conclut Dehaene.

Le scandale de l'induction

En 2011, Josh Tenenbaum et trois collaborateurs⁸ introduisirent un nouveau mot dans la langue anglaise, le concept de « tufa ». Pour expliquer ce que c'est, Tenenbaum propose trois images d'exemples de tufas. Bien entendu, les puristes parmi nous avons envie de dire qu'il s'agit d'une très mauvaise façon de définir un nouveau concept.

Et pourtant. Les quatre chercheurs font alors remarquer qu'avec seulement trois exemples de ce qu'est un tufa, et aucun exemple de ce que n'est pas un tufa, nous sommes néanmoins alors tous à peu près d'accord sur ce qu'est un tufa. En effet, Tenenbaum propose ensuite 39 images, et il nous apparaît alors évident que 6 de ces images sont des images de tufas, et que les autres images ne sont pas des images de tufas ! Incroyable !

Ce phénomène stupéfiant est parfois appelé le *scandale de l'induction*. Il est bien loin des standards de la *p-value* et des méthodes de Fisher. Pourtant, selon Tenenbaum et ses collaborateurs, il se justifie parfaitement par des principes bayésiens. En particulier, il suffit de partir du préjugé selon lequel les trois exemples de tufas sont représentatifs des tufas, et de supposer que ce qui ne ressemble pas suffisamment à ces exemples n'est pas un tufa. Mieux encore, cet apprentissage par l'exemple semble bel et bien être la manière dont on apprend

⁷  *Measuring the crowd within: Probabilistic representations within individuals* | Psychological Science | E. Vul et H. Pashler (2008)

⁸  *How to grow a mind: Statistics, structure, and abstraction* | Science | J. Tenenbaum, C. Kemp, T. Griffiths et N. Goodman (2011)

vraiment le sens des mots. Nous n'avons jamais eu de définition formelle de ce qu'est un chat. Nous avons juste vu des formes similaires, et nos parents nous ont dit qu'on les appelait « chats ».

Une explication simpliste, mais convaincante, du scandale de l'induction consiste à supposer *a priori* que l'ensemble des choses est structuré sous forme d'arbre, à l'image de l'arbre phylogénétique du vivant. En particulier, dès lors, une définition sera admissible si elle est compatible avec la structure de l'arbre, c'est-à-dire si elle correspond à un noeud de l'arbre. C'est d'ailleurs là l'approche de la phylogénie, qui définit une famille d'espèces, comme les mammifères, comme étant les descendants d'un certain ancêtre commun.

L'ensemble des définitions possibles est alors l'ensemble des noeuds de l'arbre. On sait qu'un tufa est nécessairement l'un des noeuds de l'arbre. Il reste à déterminer lequel. Tenenbaum et ses co-auteurs proposent alors d'étudier le cas simpliste où, *a priori*, tout noeud est équiprobable. Le maximum *a posteriori* sera aussi le maximum de vraisemblance, à savoir la définition de tufa pour laquelle la probabilité des trois exemples de Tenenbaum est maximale. Un calcul tout simple montre qu'il s'agit du plus bas noeud de l'arbre qui contient malgré tout tous les exemples fournis.

Tenenbaum et ses co-auteurs postulent que c'est grâce à un tel calcul qu'à l'aide de trois exemples seulement, un consensus s'est formé autour du sens du mot « tufa ». Et, plus généralement, ils postulent que c'est ainsi que les bébés apprennent les mots du langage.

Apprendre à apprendre

Cette explication semble toutefois bien incomplète. En particulier, on peut se demander comment le cerveau a pu déterminer l'arbre qui classe les objets. Mieux encore, d'où vient le choix de la structure d'arbre ? La réponse de Tenenbaum et de ses co-auteurs est fascinante : l'apprentissage de la pertinence de la structure d'arbre, et celui de la structure de l'arbre des objets, semblent tout deux conséquents de calculs bayésiens à plusieurs niveaux. Autrement dit, le cerveau semble effectuer des calculs bayésiens hiérarchiques.

Pour comprendre cela, imaginons qu'un objet A est rond et lourd, un objet B est rond et très lourd, et un objet C est très rond et lourd. Imaginons qu'on vous dise que l'objet A est parfait pour cuisiner des tufas. Pouvez-vous généraliser cela aux objets B et C ? Et si vous ne disposiez que des objets B et C, lequel de ces objets devriez-vous utiliser pour cuisiner vos tufas ?

Pour répondre à cette question, il faut savoir si, pour cuisiner des tufas, la rondeur des objets est une caractéristique plus importante que leur poids, et à quel point des variations de rondeur et de poids affectent la cuisine des tufas. Le calcul bayésien hiérarchique va alors utiliser d'autres exemples similaires, en

s'inspirant par exemple de l'effet de la rondeur et du poids des objets sur la cuisson d'autres aliments.

Cette approche est d'ailleurs celle que l'on a utilisée pour résoudre le paradoxe de Stein. Souvenez-vous, pour juger du niveau d'un pilote à partir de statistiques le concernant, il était alors utile d'invoquer les niveaux d'autres pilotes. De même, ce principe est crucial pour résoudre le problème du mouton d'Écosse. Pour généraliser la couleur noire d'un mouton, il est utile de savoir comment les couleurs d'autres espèces varient géographiquement.

Ce qui est fascinant, c'est que cette approche bayésienne hiérarchique peut être vue comme une façon d'apprendre à apprendre. Après avoir appris les effets de la rondeur et du poids sur la cuisine en général, on est désormais davantage capable de déterminer l'utilité des objets B et C pour la cuisine des tufas, alors même que seule l'information concernant l'objet A nous a été donnée ! L'apprentissage hiérarchique nous permet d'ignorer les variables non pertinentes et de nous focaliser sur celles qui importent.

Bien entendu, je ne vous en ai donné ici qu'un exemple extrêmement simpliste. Mais de façon plus générale, l'approche bayésienne hiérarchique permet de déterminer rapidement la bonne manière de structurer nos modèles du monde, à l'instar du choix de la structure de graphe pour la classification des objets ou de celui des principes causaux pour étudier des phénomènes physiques. Une fois les bonnes structures des modèles découvertes, l'apprentissage est alors accéléré, puisqu'il peut désormais être fait à l'intérieur d'un modèle restreint pertinent.

En fait, on a déjà vu un exemple plus précis d'un tel apprentissage, à savoir celui via l'allocation de Dirichlet latente. Cette structure bayésienne nous permet ainsi d'apprendre au fur et à mesure les catégories pertinentes pour classifier des documents, ce qui rendra la classification de documents futurs beaucoup plus simples. C'est d'ailleurs un modèle similaire qu'étudièrent Tenenbaum et d'autres co-auteurs⁹ en 2011, pour conclure à ce qu'ils appellent « la bénédiction de l'abstraction ».

La bénédiction de l'abstraction

Tenenbaum et ses co-auteurs illustrent cette bénédiction de l'abstraction comme suit. Ils considèrent d'abord une structure hiérarchique. Autrement dit, ils considèrent plusieurs théories très générales, chacune se subdivisant en modèles causaux particuliers, lesquels se différencient ensuite par des cas particuliers. Puis, ils entraînent une intelligence artificielle bayésienne qui applique la formule de Bayes à tous les niveaux de cette hiérarchie.

⁹  *Learning a theory of causality* | Psychological review | N. Goodman, T. Ullman and J. Tenenbaum (2011)

Cette approche s'interprète très bien en termes d'induction de Solomonoff. Souvenez-vous, l'approche de Solomonoff consistait à apprendre des théories capables d'expliquer les données passées et d'effectuer des prédictions pour les données à venir. Il y a fort à parier que les meilleures théories sont celles qui utilisent la formule de Bayes pour effectuer des prédictions à partir du passé. Le calcul bayésien du démon de Solomonoff, qui a lieu au niveau inter-théorie, aura alors de bonnes chances de mettre les crédences sur des théories bayésiennes, lesquelles appliqueront à nouveau la formule de Bayes à un niveau donc inférieur. Mieux encore, ces théories bayésiennes étudieront alors diverses sous-théories, pour finalement préférer des sous-théories bayésiennes. Et ainsi de suite.

Quoi qu'il en soit, les simulations de Tenenbaum et de ses co-auteurs montrent que l'apprentissage bayésien hiérarchique est initialement lent à tous les niveaux. Mais au bout de quelques centaines ou milliers d'échantillons étudiés, la structure hiérarchique finit par mettre ses crédences bayésiennes sur les bonnes théories générales. De façon intrigante, l'apprentissage des bonnes théories générales est bien plus rapide que celui des niveaux inférieurs. Un corollaire de cette remarque est que démarrer directement l'apprentissage avec uniquement les bonnes théories générales n'apporte essentiellement aucun gain de temps !

On y reviendra, mais certains psychologues stupéfaits des prouesses de jeunes enfants ont fini par postuler que de nombreux modèles étaient génétiquement encodés dans le cerveau de ceux-ci. Il semble que l'enfant a, par exemple, une prédisposition à invoquer le principe de causalité. Cependant, les simulations de Tenenbaum et de ses collègues montrent que les principes généraux des modèles sont en fait relativement rapides à apprendre pour une intelligence bayésienne hiérarchique ! Nul besoin de modèles d'apprentissage prémaçhés. Le bayésianisme semble permettre la découverte des paradigmes pertinents pour penser le monde avec une efficacité déconcertante !

Certains théoriciens se plaignent parfois de l'utilisation abusive d'intelligences artificielles en sciences. L'exploration du *Big Data*, disent-ils, ne permettra jamais de découvrir des formules élégantes et synthétiques. Elle ne conduira jamais à des merveilleuses équations comme celles de la relativité générale que découvrit Einstein. Le *machine learning* semble trop mécanique. Il semble manquer du génie dont seuls les grands esprits sont capables.

Cependant, cet argument omet la bénédiction de l'abstraction des méthodes bayésiennes. En particulier, l'approche bayésienne hiérarchique semble parfaitement à même de distinguer parmi un grand ensemble de structures générales de théories celles qui expliquent le mieux les données empiriques. Elles semblent capables de distinguer la forme générique des meilleures théories, et de nous fournir ainsi le paradigme adéquat pour penser les données empiriques.

Et si jamais une équation élégante est la *bonne* façon de modéliser ces données, il y a fort à parier qu'une intelligence artificielle suffisamment proche du démon de Solomonoff saura la déterminer. À l'instar des neuroscientifiques découvrant la pertinence du cadre bayésien pour étudier les processus cognitifs.

Le bébé est un génie

On a vu que Turing avait comparé le bébé à un carnet de notes avec « peu de mécanisme, et beaucoup de pages blanches ». Ce principe aura longtemps prédominé. Ainsi le chef d'État chinois Mao Zedong aurait vanté l'ignorance de son peuple, en soulignant que « sur une feuille blanche de papier sans aucune marque, les mots les plus frais et les plus élégants peuvent être écrits ; les dessins les plus frais et les plus élégants peuvent être dépeints ». Cependant, cette idée est battue en brèche par le psychologue Steven Pinker et par les neurosciences modernes. Et par le bayésianisme.

Avant d'en arriver là, notons que, avant même l'âge de 1 an, le bébé possède déjà un ensemble de *core knowledge*. Il a ainsi une compréhension intuitive des objets, des nombres, de l'espace et de la grammaire. Mieux encore, le bébé semble déjà doué de facultés statistiques. Il va typiquement regarder longuement et avec curiosité des événements que son cerveau statisticien juge improbable.

En 2008, Xu et Garcia¹⁰ le montrèrent à l'aide d'une expérience sur des enfants de 8 mois. L'expérience de Xu et Garcia est inspirée des urnes de Laplace. L'urne contient un grand nombre de balles. Ces balles sont rouges ou blanches. On tire ensuite 5 balles de l'urne. Imaginons que 4 boules rouges soient tirées, contre 1 seule boule blanche. La loi de succession de Laplace, dont on a parlé au chapitre 6, nous invite à penser que la proportion de boules rouges dans l'urne est environ $(4+1)/(5+2) = 5/7$.

Puis, on révèle le contenu de l'urne. Lorsque l'urne contient majoritairement des boules rouges, l'enfant n'est alors pas surpris. Cependant, lorsque l'urne contient en fait majoritairement des boules blanches, l'enfant se met alors à regarder longuement l'urne, comme si celle-ci cachait quelque chose de mystérieux. Incroyable ! L'enfant de 8 mois semble déjà en mesure d'intuiter le calcul bayésien de Laplace, et se met à enquêter, tel un scientifique, les cas où ses prédictions bayésiennes semblent contredites par l'observation !

Le langage

L'un des apprentissages les plus stupéfiant du bébé est celui du langage. On en a tous fait l'expérience. Malgré des décennies d'études, voire d'immersion dans un pays étranger, il nous est très difficile de parler comme les locaux. Les français expatriés dans des pays anglophones conserveront leur accent français. À l'inverse, les bébés ont une faculté bluffante à acquérir le langage de leurs parents. Après quelques années, ils atteignent une maîtrise de la langue que peu d'étrangers atteindront dans leur vie. À l'âge de deux ans, leur lexique s'enrichit à un rythme effréné de 10 à 20 mots par jour ! Comment font-ils ?

¹⁰  *Intuitive statistics by 8-month-old infants* | Proceedings of the National Academy of Sciences | F. Xu et V. Garcia (2008)

Les neurosciences suggèrent que l'apprentissage de la langue par les bébés repose fortement sur l'étude des propriétés statistiques des sons élémentaires de la langue, appelés phonèmes. L'expérience de Saffran, Aslin et Newport¹¹ expose par exemple des bébés à des suites de syllabes. Le rythme auquel ces syllabes sont énoncées est régulier, si bien qu'il est impossible d'inférer quoi que ce soit du rythme auquel ces syllabes sont énoncées. Cependant, une certaine loi statistique se cache derrière l'enchaînement de ces syllabes. Par exemple, un « to » sera toujours suivi d'un « ki », alors que « bu » sera suivi de « gi » seulement une fois sur 3. On parle de *chaîne de Markov*¹².

De façon étonnante, le bébé semble non seulement capable de repérer ces régularités statistiques. Il réussit même, sans doute via une inférence bayésienne, à déterminer les segmentations qui forment probablement des mots. Autrement dit, il semble qu'un calcul bayésien soit crucial à l'apprentissage des mots à partir d'un langage parlé.

Puis, le bébé parvient à combiner des mots pour distinguer des phrases. Il s'agit là d'une prouesse remarquable ! En effet, pour ce faire, le bébé doit parvenir à identifier la structure grammaticale des phrases, c'est-à-dire reconnaître que certains mots sont des verbes, tandis que d'autres sont des noms. Le bébé doit aussi apprendre que, dans certaines langues, l'ordre des mots peut s'inverser quand il s'agit de questions. Cependant, ces efforts, indispensables pour construire des phrases, seront aussi très utiles à l'apprentissage de davantage de mots.

Illustrons cela. Imaginons que deux bols se trouvent sur la table. L'un est bleu, l'autre est de couleur chrome. Un parent demande à un enfant de lui ramener le bol chrome. L'enfant ne connaît pas le sens du mot « chrome ». Cependant, il sait qu'il s'agit d'un adjectif qui décrit le bol. Qui plus est, il devine que le parent ne parle sans doute pas du bol bleu, puisque le parent l'aurait sans doute appelé ainsi. L'enfant conclut alors que le bol chrome n'est sans doute pas le bol bleu, et que le mot « chrome » est la couleur de ce bol chrome.

Incroyable ! L'enfant vient de prédire et apprendre le sens du mot « chrome » sans données. Alors qu'il n'avait jamais entendu ce mot, il a quand même su déterminer son sens. Et son génie fut de s'appuyer sur ses préjugés. Comme le dirait Solomonoff, « il est possible de faire des prédictions sans données, mais il n'est pas possible de faire de prédition sans *a priori* ».

¹¹  *Statistical learning by 8-month-old infants* | Science | J. Saffran, R. Aslin et E. Newport (1996)

¹²  *La machine à inventer des mots (avec Code MU)* | Science étonnante | D. Louapre (2015)

Apprendre à compter

Quand les bébés apprennent l'alphabet ou les nombres, ils apprennent par cœur une suite de mots ou de sons. D'ailleurs, l'apprentissage de l'alphabet est souvent accompagné de celui d'une mélodie qui aide à la mémorisation. De même, l'apprentissage des nombres s'accompagne parfois d'un comptage sur les doigts.

Cependant, connaître et réciter par cœur une suite de mots ne signifie pas être capable de les réutiliser à bon escient dans d'autres contextes. En effet, de nombreux jeunes enfants savent réciter les nombres. Mais si on leur demande de ramener trois objets, ils prennent alors une poignée de ces objets au lieu de n'en prendre que trois.

De façon plus intrigante encore, les enfants vont d'abord uniquement comprendre le sens du mot « un ». Ensuite, ils apprendront le mot « deux », mais n'iront pas plus loin. Pendant plusieurs mois, ils ne connaîtront le sens que de ces deux nombres, quand bien même ils savent réciter des nombres au-delà de « un » et « deux ». Puis ils apprendront « trois ». Et en resteront là.

Puis, tout à coup, vers trois ans et demi, les enfants semblent réussir un saut conceptuel remarquable. Ils arrivent à faire correspondre la suite des nombres au sens des nombres. Ils établissent une relation entre « inclure un objet additionnel » et « passer au mot suivant de la suite des nombres ». Comment cette relation a-t-elle été établie ?

Selon Tenenbaum, et ses collaborateurs Piantadosi et Goodman¹³, l'enfant vient alors de réussir la prouesse fantastique d'avoir appris un algorithme *récursif*. Cet algorithme prend en compte n'importe quel ensemble d'objets. Si l'ensemble est vide, l'algorithme conclut avec le dernier nombre énoncé. Sinon, l'algorithme retire un objet de l'ensemble et énonce le mot qui suit dans la suite des nombres — s'il s'agissait du premier objet retiré, l'enfant énonce alors le mot « un ». Puis l'algorithme compte le reste des objets de l'ensemble, en gardant en tête le dernier mot prononcé.

Ce que je décris là en termes abstraits est en fait un algorithme que vous connaissez (littéralement) sur le bout des doigts, et que vous appliquez à chaque fois que vous comptez un à un les éléments d'un ensemble. De façon remarquable, toutefois, cet algorithme incontournable est en fait étonnamment abstrait. Et le plus étonnant, c'est que le bébé statisticien a su identifier et sélectionner cet algorithme récursif abstrait.

Dehaene suggère que cette capacité à penser et exploiter la récursivité algorithmique pourrait être la différence fondamentale entre le cerveau humain et celui des autres animaux. Quoi qu'il en soit, cette étude bayésienne d'algorithmes récursifs suggère surtout que nos cerveaux bayésiens ne sont finalement peut-être pas si éloignés de celui du démon de Solomonoff.

¹³  *Bootstrapping in a language of thought: a formal model of numerical concept learning* | Cognition | S. Piantadosi, J. Tenenbaum et N. Goodman (2012)

La théorie de l'esprit

L'une des facultés les plus fondamentales que nos cerveaux bayésiens apprennent pendant l'enfance est la théorie de l'esprit. Il s'agit de notre capacité à penser ce que d'autres individus pensent, et à utiliser ce modèle de la pensée d'un autre pour effectuer des prédictions ou apprendre de nouveaux concepts. Ainsi, avant même ses deux ans, le bébé va être capable de suivre le regard d'un autre, d'imiter ses actions, voire de distinguer ses intentions de ses actes manqués. Plus tard, il apprendra que les croyances d'un autre ne sont pas nécessairement les mêmes que celles d'une tierce personne, que les autres peuvent mentir ou parler au second degré avec ironie, sarcasme et humour, et que de nombreux gestes des autres sont involontaires. Peut-être parce que nous sommes des créatures sociales, cette théorie de l'esprit est indispensable à notre apprentissage.

Imaginons une urne transparente contenant beaucoup de jouets bleus et quelques jouets jaunes similaires. Un adulte plonge sa main dans l'urne, et retire trois objets bleus. Il appuie sur chacun de ces trois objets bleus. Chacun fait du bruit. On tire alors un objet jaune de l'urne que l'on donne à un enfant. La question que l'on se pose alors, c'est si l'enfant qui observe cette situation va généraliser la capacité des objets bleus à faire du bruit aux objets jaunes. La réponse est alors oui. L'enfant va alors appuyer 3 fois sur le jouet jaune, qui ne fait pas de bruit, avant d'abandonner ses tentatives.

Jusque-là, rien de très surprenant. Ce qui est amusant est la variante de l'expérience dans laquelle l'urne ne contient que quelques jouets bleus et beaucoup de jouets jaunes. Cette fois encore, l'adulte tire trois jouets bleus et montre qu'en appuyant dessus, il peut leur faire faire du bruit. Puis il tire un jouet jaune. L'enfant va-t-il croire que ce jouet jaune fait du bruit ? Curieux, l'enfant va en effet tester son jouet. Cependant, cette fois-ci, il ne le teste qu'une seule fois !

Il semble ainsi que l'enfant a remarqué que, dans ce second cas, l'échantillonnage de l'adulte était biaisé. L'enfant s'est sans doute dit que l'adulte a fait exprès de ne tirer que des jouets bleus, car ceux-ci sont très différents des jouets jaunes. L'enfant a compris qu'il y avait eu un biais de sélection, et en vertu d'un calcul bayésien, il en a conclu que toute propriété des jouets bleus n'était donc pas clairement transférable aux jouets jaunes !

L'enfant a non seulement su déterminer l'effet du biais de sélection sur la généralisabilité d'une observation — une prouesse dont nous sommes bien incapables dans des cas plus abstraits ! Il a également su modéliser la réflexion de l'adulte pour ce faire ! L'enfant est doué de la théorie de l'esprit, et sait l'appliquer pour éviter des conclusions erronées.

Innée ou acquis ?

Si Dehaene parle de « révolution bayésienne en neurosciences », ce n'est pas uniquement parce que la théorie du cerveau bayésien rend compte d'une quantité monstrueuse de résultats expérimentaux difficilement interprétables autrement. La magie de l'approche bayésienne, c'est aussi de fournir une réponse étonnamment complète à l'un des vieux débats sur l'apprentissage de l'enfant. Ce vieux débat oppose les innéistes, selon qui le cerveau naît avec une connaissance *a priori* de la langue et de la grammaire, aux empiristes selon qui tout est appris. Et ce débat fut personnifié par la mythique opposition entre les psychologues Skinner et Chomsky.

Tout débutea avec la publication de *Verbal Behavior* en 1958 par Burrhus Skinner. Skinner s'appuya notamment sur des expériences où il montra que des pigeons étaient capables d'apprendre le sens de mots comme « donner un coup de bec » ou « tourner sur soi¹⁴ ». En effet, Skinner découvrit qu'en récompensant les pigeons au moment où ils effectuaient l'action associée aux mots imprimés qui leur étaient montrés, les pigeons parvenaient à apprendre le sens des mots. Ou du moins à corrélérer les mots qu'ils voyaient avec l'action à mener pour être récompensés.

Cependant, Noam Chomsky rétorqua que la maîtrise des langages humains était une compétence beaucoup plus complexe et difficile à apprendre que la corrélation entre des mots et des événements. Pour Chomsky, la sophistication de la langue et de sa grammaire ne pouvait être apprise qu'à l'aide d'un cerveau prédisposé à ce faire. Chomsky postula que nos cerveaux étaient génétiquement programmés pour comprendre et manipuler ce qu'il appela la grammaire universelle.

Cependant, selon l'hypothèse du cerveau bayésien, dont on a vu l'incroyable pouvoir prédictif, tout bébé doit avant tout posséder la faculté de générer de complexes modèles de divers phénomènes dans son environnement, et celle d'appliquer la formule de Bayes pour retenir et explorer les modèles les plus utiles. De façon cruciale, les simulations de Tenenbaum et de ses collaborateurs suggèrent que cet appareil est nécessaire et suffisant pour permettre au bébé de modéliser son environnement, de comprendre le langage et d'apprendre à parler — ce que confirme le théorème de complétude de Solomonoff !

D'une certaine manière, la structure innée du cerveau du bébé donne raison à l'innéisme, qui postule la nécessité de prédispositions de notre cerveau. Cependant, cette structure innée semble beaucoup plus abstraite, plus simple et plus merveilleuse que les structures que Chomsky a postulées. On l'a vu. Les grands paradigmes de pensée peuvent en fait être rapidement déterminés par l'approche bayésienne hiérarchique.

¹⁴En anglais, il s'agit des mots « *peck* » et « *turn* », qui ont d'ailleurs aussi le bon goût d'avoir le même nombre de lettres.

À l'inverse, les données empiriques jouent un rôle crucial dans la sélection des modèles utiles via la formule de Bayes. Cependant, là encore, cet apprentissage des données n'est absolument pas réduit au simple calcul de corrélations par apprentissage par renforcement. Cet apprentissage est celui d'un modèle extrêmement complexe et sophistiqué, qui effectue des calculs bayésiens à différents niveaux.

C'est stupéfiant. Je trouve absolument fascinant et dérangeant que de constater que la formule de Bayes, qui me semble si inaccessible tant je me sens incomptéte au moment de manipuler les probabilités, semble en fait calculée par mon cerveau ; le même cerveau que celui qui aspire, en vain, à devenir penseur bayésien conscient et compétent. Nous semblons tous disposer de machines incroyablement sophistiquées et optimisées pour effectuer de complexes calculs bayésiens, de manière massivement parallèle avec une performance énergétique inégalée. Toutefois, étrangement, nous sommes très inconscients de ces calculs. Et nous sommes très incapables de les réutiliser pour penser juste.

Références en français

Le cerveau statisticien : la révolution bayésienne en sciences cognitives | Collège de France | S. Dehaene (2011-2012)

Le bébé statisticien : les théories bayésiennes de l'apprentissage | Collège de France | S. Dehaene (2012-2013)

▶ *La machine à inventer des mots (avec Code MU)* | Science étonnante | R. Koschig et D. Louapre (2015)

Références en anglais

▶ *Statistical learning by 8-month-old infants* | Science | J. Saffran, R. Aslin and E. Newport (1996)

▶ *Motion illusions as optimal percepts* | Nature Neuroscience | Y. Weiss, E. Simoncelli and E. Adelson (2002)

▶ *A recurrent network mechanism of time integration in perceptual decisions* | Journal of Neuroscience | K.F. Wong and X.J. Wang (2006)

▶ *Measuring the crowd within: Probabilistic representations within individuals* | Psychological Science | E. Vul and H. Pashler (2008)

▶ *Intuitive statistics by 8-month-old infants* | Proceedings of the National Academy of Sciences | F. Xu and V. Garcia (2008)

▶ *Infants consider both the sample and the sampling process in inductive generalization* | Proceedings of the National Academy of Sciences | H. Gweon, J. Tenenbaum and L. Schulz (2010)

- ☒ *Bayesian sampling in visual perception* | Proceedings of the National Academy of Sciences | R. Moreno-Bote, D. Knill, and A. Pouget (2011)
 - ☒ *How to grow a mind: Statistics, structure, and abstraction* | Science | J. Tenenbaum, C. Kemp, T. Griffiths and N. Goodman (2011)
 - ☒ *Learning a theory of causality* | Psychological review | N. Goodman, T. Ullman and J. Tenenbaum (2011)
 - ☒ *Bootstrapping in a language of thought: a formal model of numerical concept learning* | Cognition | S. Piantadosi, J. Tenenbaum and N. Goodman (2012)
- ▶ *The "Mountain Or Valley?" Illusion* | Minute Physics (2017)

Il n'y a pas de faits, juste des interprétations.

Friedrich Nietzsche (1844-1900)

L'univers est fait d'histoires, pas d'atomes.

Muriel Rukeyser (1913-1980)

La science remplace du visible compliqué par de l'invisible simple.

Jean Perrin (1870-1942)

20

Tout est fiction

La caverne de Platon

Imaginez-vous enchaîné et contraint à n'observer que la paroi d'une grotte. De temps à autre, vous voyez des ombres apparaître et se mouvoir sur la paroi. Mais, ne pouvant pas tourner la tête, vous ne pouvez pas regarder directement la cause de l'ombre. Tout ce que vous savez du monde qui vous entoure, ce sont les ombres que vous percevez sur la paroi. Votre réalité se limite alors à ces ombres. Vous prenez ces ombres pour la réalité.

Ce que je vous décris là est l'allégorie de la caverne introduite par le philosophe Platon à l'époque de la Grèce antique. Pour Platon, cette allégorie était une métaphore adéquate de l'ignorance de ses concitoyens. Platon va même plus loin encore. Il imagine qu'un de vos camarades, enchaîné lui aussi, est libéré. Ce camarade, appelons-le Pierre-Simon, se retourne, mais est si ébloui par la lumière du soleil qu'il préfère se rasseoir et contempler la paroi et ses ombres. Pour Platon, non seulement les concitoyens sont très ignorants, mais ils se confortent dans leur ignorance.

Platon imagine ensuite que, néanmoins, Pierre-Simon est amené, peut-être de force, à sortir de la grotte. Pierre-Simon est d'abord effrayé. Il est d'abord déboussolé. Cependant, petit à petit, il prend plaisir à découvrir un monde beaucoup plus réel que celui de la paroi et de ses ombres. Tout excité, il finit par retourner dans la grotte pour vous raconter ses innombrables découvertes. Le problème, c'est que vous le prenez désormais pour un fou ! Vous ne le croyez

pas, et rejetez en bloc tout ce qu'il a à vous dire. Pour Platon, ceci illustre la plus grande faiblesse de ses concitoyens : ses concitoyens préfèrent vivre dans leur réalité erronée et ne pas la remettre en cause.

L'allégorie de la caverne de Platon a été reprise dans une scène mémorable du film *Matrix*. Alors que Néo, le personnage principal, commence à douter du réalisme de son quotidien, Morpheus lui propose le fameux dilemme des pilules rouges et bleues. Prenez la pilule bleue, et réveillez-vous ignorant. Vous profiterez de votre quotidien, mais serez incapable de voir qu'il ne s'agit que d'ombres projetées sur le fond d'une caverne. Prenez la pilule rouge, et quittez ce quotidien. Allez découvrir le monde *réel*. Heureusement pour l'intérêt du film, Néo choisit la pilule rouge¹.

Cette pilule rouge permet à Néo de quitter la simulation informatique — la fameuse *Matrix* — dans laquelle il était enfermé. Néo se réveille alors dans un tout autre monde, apocalyptique, dangereux et dominé par les machines. Plus tard, Morpheus accompagne Néo dans une autre simulation informatique. Néo est déboussolé. « Ceci n'est pas réel », s'interroge-t-il. Morpheus répond par une autre question : « qu'est-ce que le réel ? Comment définir le réel ? » Morpheus apporte ensuite une réponse possible. « Si vous parlez de ce que l'on peut toucher, sentir, goûter et voir, alors le réel est simplement des signaux électriques interprétés par le cerveau. »

L'antiréalisme

L'allégorie de la caverne de Platon et la simulation de la *Matrix* sont des exemples fascinants, car elles remettent en cause la nature de la réalité. Mais pour la *pure bayésienne*, elles ne vont pas assez loin. Dans ces deux exemples, il reste admis que ce qui se trouve à l'extérieur du monde dans lequel les individus s'enferment volontiers est le monde *réel*. Il demeure alors sous-entendu que le réel *existe*. Dans ces exemples, ce qui perturbe et intrigue, c'est que le réel n'est pas ce que les habitants de ces mondes croient qu'il est.

Ce discours est récurrent dans la manière dont beaucoup de scientifiques parlent des sciences. On peut ainsi lire par-ci par-là que les sciences *révèlent* une vérité invisible, ou un monde réel caché. Voir qu'il s'agit de la seule voie vers *la* vérité, et la fin des illusions. On pourrait croire que l'eau est infiniment sécable, mais *en vrai*, l'eau n'est qu'un ensemble fini de molécules qui glissent les unes sur les autres. On pourrait croire que le temps et l'espace sont absolus, mais *en vrai*, ils sont relatifs à la trajectoire de l'observateur dans l'espace-temps. On pourrait croire que le corps humain contient uniquement *nos* cellules, mais *en vrai*, la plupart des cellules vivantes qui composent nos corps sont des bactéries en tout genre desquelles dépendent notre santé et notre humeur.

¹  *La machine à expérience de Robert Nozick | Argument frappant | Monsieur Phi | T. Giraud (2018)*

Dans ce chapitre, je vais m'attarder sur ce qui est, à moins d'une erreur dans mon raisonnement, une conséquence inéluctable contre-intuitive du *bayésianisme*. Je l'ai laissée pour la fin, car vous aurez sans doute très envie de la rejeter, même si vous avez bien suivi l'enchaînement des idées de ce livre. Cette conséquence étrange que je prétends être inéluctable est l'affirmation que *tout est fiction*. Certaines fictions sont plus crédibles pour la *pure bayésienne*, d'autres sont plus utiles pour le *bayésien pragmatique*. Cependant, pour tout bon bayésien, il me semble nécessaire que tout ne peut être que fiction ; ou plus précisément, tout n'est que simulations d'un nombre infini d'algorithmes randomisés sur lesquels parier.

En particulier, je prétends qu'un bon bayésien rejette forcément le postulat selon lequel il existerait nécessairement une *réalité* conforme à celle qu'on s'imagine, un univers au-delà de la grotte de Platon, ou un monde physique en dehors de la *Matrix*. Mieux encore, je prétends qu'il n'est ni indispensable, ni même utile, de postuler que les électrons *existent vraiment*, ni même que la physique parle vraiment d'une réalité objective. Certes, ces modèles sont incroyablement utiles et prédictifs et méritent une grande partie de nos croyances. Mais conformément à la citation de Box, je prétends qu'il est davantage utile de garder en tête qu'il ne s'agit là que de modèles, et que « tous les modèles sont faux ». En particulier, en bon bayésien, je défendrai l'*utilité* de cette position fictionnaliste.

La vie existe-t-elle ?

Les débats sur le réalisme virent souvent vers des questions existentialistes comme l'existence de la conscience². Il s'agit là d'un sujet trop controversé pour être adressé directement. Je vous propose donc de commencer par une question d'apparence plus simple. Est-ce que la vie *existe* ?

La vie est l'un des concepts les plus difficiles des sciences — à bien y réfléchir, rares sont les concepts scientifiques qui ne sont pas problématiques. Certains biologistes n'hésitent pas à reconnaître qu'il s'agit tout simplement d'un terme mal défini. D'autres proposent de lister des critères, et de considérer que la vie est l'ensemble des phénomènes naturels qui satisfont ces critères. Parmi les critères habituellement cités, on retrouve les notions de réplicabilité et de variations, sur lesquelles repose notamment la théorie de l'évolution de Darwin. Le problème, c'est que de telles définitions échouent souvent à coïncider avec ce que la plupart d'entre nous aimeraient appeler la vie — en particulier un virus informatique serait alors considéré vivant. Certains biologistes ont du coup choisi de restreindre le vivant à ce qui tourne autour de molécules typiques du vivant, comme l'acide désoxyribonucléique (ADN) et l'acide ribonucléique (ARN). Mais on peut alors se demander si le stockage de données sur des brins d'ADN pourra être considéré vivant.

²  *La conscience* | Science Étonnante | T. Giraud et D. Louapre (2017)

Une alternative consiste à identifier des propriétés physico-chimiques typiques du vivant. C'est l'approche qu'a choisi Karl Friston. Celle-ci repose sur la notion de *couverture de Markov*. Cette couverture, typiquement la membrane d'une bactérie, est une séparation matérielle stable qui distingue le monde extérieur d'une structure intérieure. L'intérieur de la membrane puiserait dans son environnement une certaine forme d'énergie, appelée énergie libre, pour constamment consolider sa structure³. La propriété fondamentale de cette énergie libre est de se trouver très loin de tout équilibre thermodynamique. Pour utiliser à bon escient cette énergie libre et consolider sa structure interne, il est alors crucial pour l'intérieur d'anticiper les perturbations extérieures — et déterminer s'il va y avoir de l'énergie libre à laquelle accéder. Voilà qui exige de l'intérieur de développer un modèle de la réalité extérieure. Mais de façon cruciale, les seules données dont l'intérieur dispose pour ce faire sont uniquement l'information présente sur la couverture de Markov.

En 2013, Friston suggéra que des processus naturels étaient à l'œuvre pour permettre à l'intérieur de calculer une approximation de la formule de Bayes appelée inférence variationnelle bayésienne, que l'on a brièvement mentionnée au chapitre 14. Friston suggère même qu'il s'agit là de la nature fondamentale de la vie : la vie est une structuration d'un environnement restreint et séparé du monde extérieur d'une membrane stable. Ou, en termes thermodynamiques, la vie est un puits d'entropie séparé d'un océan de grande entropie par une couverture de Markov.

Cependant, bien qu'utile, et à l'instar des autres définitions qui s'appuient sur les propriétés du vivant, ou sur les molécules du vivant, il y a fort à parier que la définition de Friston inclut des objets que notre intuition ne considère pas vivants, et vice-versa. À voir la difficulté à définir la vie, on peut se demander si la vie *existe*. Y a-t-il vraiment une réalité au concept de vivant ? Y a-t-il quelque chose de *vrai* dans l'affirmation que le monde physique se divise en parties *vivantes* et parties non-vivantes ? Parle-t-on là de quelque chose de *réel* ?

L'argent existe-t-il ?

Définir le vivant ne semble pas être une question de sécurité nationale pour l'instant — encore que des débats sur l'interruption volontaire de grossesse, la souffrance animale ou les droits des robots soulèvent la question de la nature du vivant. Ce n'est toutefois pas le cas de l'argent. En particulier, en 2008, une monnaie entièrement virtuelle appelée Bitcoin a été introduite et a gagné en valeur. Aujourd'hui, en 2018, l'ensemble de tous les bitcoins en circulation représente une masse monétaire totale d'une valeur de quelques milliards de dollars. Pas mal, pour ce qui n'était qu'un article de recherche publié de manière

³L'énergie libre est aussi étudiée par d'autres chercheurs comme Jeremy England. Voir :  *The Physics of Life (ft. It's Okay to be Smart & PBS Eons!)* | PBS Space Time | M. O'Dowd (2018)

anonyme, il y a moins de 10 ans ! Mais comment cette monnaie, qui n'a pas de support physique, peut-elle avoir une valeur quelconque ? Et si celle-ci n'a pas de support, peut-on encore dire que cette monnaie *existe* ?

Ce qui est étrange, c'est que ces mêmes questions s'appliquent tout autant à l'argent classique dont l'existence semble pourtant reconnue par tous. En effet, dès aujourd'hui, 90 % de l'argent que l'on pense posséder et échanger est électronique. Mais on peut tout autant remettre en cause la réalité de l'argent physique. Après tout, le plus gros de l'argent physique circule aujourd'hui sous forme de bouts de papier. Pourquoi ces bouts de papier auraient-ils une valeur quelconque ? Qu'est-ce qui fait de ces papiers des *vrais* billets d'argent ? Et si des contrefaçons de ces billets sont en libre circulation sans que personne sache les distinguer des « vrais » billets, ces contrefaçons demeureront-elles *fausses*⁴ ? Bref. Qu'est-ce que l'argent ?

Selon Yuval Noah Harari, l'argent, à l'instar de la sacralisation de la vie et des mythes historiques, fait partie de l'ensemble des fictions qui forment la plus grande invention de l'humanité⁵. À titre de comparaison, les fictions en lesquelles croient les chimpanzés sont peu nombreuses. Pour Harari, cette incapacité des espèces autres que *Homo sapiens* à raconter et à croire en des fictions est ce qui les a empêchés de s'organiser en populations de plusieurs centaines d'individus. Par opposition, *Homo sapiens* a su faire coopérer des tribus, puis des civilisations dont les populations se comptaient rapidement en milliers, en dizaines de milliers, voire, aujourd'hui, en plusieurs milliards.

L'une des grandes innovations des sociétés humaines est le troc. Celui-ci permet à deux personnes de toutes les deux gagner en effectuant un échange commercial. Mais le troc a ses limites. Le problème, c'est que de nombreuses tâches requièrent des investissements, et ne seront rentables que sur le long terme. Vient alors la pierre angulaire de l'économie de marché : la création de dettes. Autrement dit, un investisseur peut aider un entrepreneur à se lancer dans un projet, auquel cas l'entrepreneur sera redevable envers l'investisseur. L'investisseur crée par là une dette que l'entrepreneur devra payer. Et la quantité de dettes que l'entrepreneur aura est ce que l'on appelle désormais l'argent.

La création des dettes est sans doute l'une des plus grandes innovations de l'histoire de l'humanité. Elle aura permis l'avènement de l'économie de marché, puis la spécialisation du travail, dont les conséquences sont absolument vertigineuses. Comme l'a fameusement constaté Adam Smith, la fabrication et commercialisation d'une simple veste de laine n'est permise que par l'étonnante et complexe interaction d'un très grand nombre d'individus égoïstes.

Pensez-y. Pour que vous receviez cette veste, ont été nécessaires un berger, un teinturier, un fileur, un tisserand, mais aussi un complexe système de distribution, qui inclut des grossistes, des investisseurs, des navigateurs, des chauffeurs,

⁴  Qui crée l'argent ? Comment ? Heu?reka | G. Mitteau (2015)

⁵  La Grande Histoire des petites histoires | Démocratie 26 | Science4All | L.N. Hoang (2017)

des entreposeurs, des postiers, des vendeurs, mais aussi des constructeurs de navire, des ingénieurs, des techniciens, des responsables de production, sans oublier ceux sans qui les ouvriers n'auraient pas les outils pour fabriquer la veste, y compris les outils les plus simplistes comme des ciseaux. La fabrication de ces ciseaux requiert ainsi des mineurs, des forgerons, des bûcherons pour fournir le bois nécessaires à la consolidation de la mine, des briquetiers et des maçons. Sans compter les siècles de travailleurs du passé sans qui les fondations et l'expertise des travailleurs d'aujourd'hui seraient inexistantes.

Bref. Force est de constater que la liste des compétences et travaux nécessaires à la production et à la distribution de toute veste de laine est interminable ! De surcroît, ce qui est non moins ahurissant, c'est que toute cette machinerie fantastique fonctionne, bien qu'aucun individu ne soit capable de produire une veste de laine tout seul. Abandonné à lui-même, personne ne saurait fabriquer une telle veste seul, à moins d'y consacrer des décennies — et on peut se demander comment une telle personne se nourrirait ! « Nous vivons au milieu d'objets incroyables, des objets incompréhensibles, des objets que personne, littéralement personne, ne sait fabriquer seul », résume Thibaut Giraud sur sa chaîne Monsieur Phi⁶.

La création de dettes est devenue le moteur de l'économie et du progrès technologique. Mais pour le bon fonctionnement de la création de dettes, il fallait un système qui puisse garantir à tout moment et en toutes circonstances le fait que deux individus se souviennent et se mettent d'accord sur les dettes qu'ils se doivent mutuellement. Ce système est le système monétaire. Chaque échange monétaire peut ainsi être vu comme un remboursement ou une création de dettes entre deux individus. *Je te donne un gâteau, tu as donc une dette envers moi, et tu me rembourses la dette par un don monétaire.* Ou de façon équivalente, *tu me donnes de l'argent pour créer une dette, que je rembourse en t'offrant un gâteau.*

Le génie de la monnaie, c'est que cette dette est ensuite transférable. Si Alice me doit de l'argent, Bob peut effacer la dette d'Alice en me donnant cet argent. Les billets de banque, les systèmes bancaires électroniques et le Bitcoin sont alors des technologies pour déterminer à tout moment l'état des dettes à travers le monde. Ils permettent aussi et surtout de ne pas oublier quelles dettes restent à payer. Est-ce à dire que ces dettes *existent* ? Que se passe-t-il si j'ai une dette envers un ami, mais que mon ami et moi oubliions l'existence de cette dette ? Que se passe-t-il si cette information sur la dette est perdue à jamais ? Cette dette qui demeurera toujours impayée sera-t-elle *réelle* ? Comment est-ce possible que l'un des socles de nos sociétés semble aussi peu concret ?

La réponse du bayésien est limpide : *tout est fiction*. La dette, comme l'argent ou la vie, n'est qu'une histoire. Ce qui fait sa force, toutefois, c'est qu'il s'agit d'une histoire qui a gagné les crédences de la quasi-totalité d'entre nous, et qui nous permet de construire et vivre ensemble. L'existence des dettes est un

⁶  Adam Smith - le paradoxe de la veste de laine | Grain de Philo | Monsieur Phi (2016)

modèle faux. Mais il est *utile*. Il est utile non seulement pour la société dans sa globalité, mais aussi pour chacun d'entre nous. Il est utile de penser qu'un billet de 10 euros a vraiment une telle valeur, car cela nous conduira à le garder en poche plutôt qu'à le jeter à la poubelle. On pourra ensuite échanger ce billet contre une bouteille de vin, et profiter de notre croyance, erronée mais utile, en la valeur intrinsèque du billet. Il en va de même avec la monnaie virtuelle Bitcoin. Ce qui importe n'est pas l'existence réelle de la monnaie. C'est l'utilité de croire en son existence. Le modèle de l'existence de l'argent est une fiction utile. Ce n'est pas le seul.

Selon Harari, les fictions qui ont permis à des individus égoïstes de collaborer au service d'un groupe sont les plus grandes innovations de l'humanité. On a tous été bercés par ces mythes. « Liberté, égalité, fraternité », comme le dit la devise de la République française. « Les hommes naissent libres et égaux en droits », rajoute la Déclaration des Droits de l'homme et du citoyen de 1948. Dans le préambule de la Déclaration d'Indépendance des États-Unis, Thomas Jefferson écrit : « nous tenons ces vérités comme évidentes, que tous les hommes sont créés égaux, qu'ils sont dotés par leur créateur de certains Droits inaliénables, parmi lesquels la Vie, la Liberté, et la poursuite du Bonheur. »

Cependant, à l'instar de l'argent et de la vie, et à bien y réfléchir, tous les concepts qui apparaissent dans ces récits séduisants sont en fait loin d'être clairement *vrais ou réels*. Même le concept d'identité personnelle⁷ est remis en cause par certains philosophes. Ainsi David Hume, ce grand-père du bayésianisme dont on a parlé au chapitre 4, affirmait : « les hommes ne sont rien d'autre qu'un faisceau ou une collection de perceptions qui se succèdent les unes les autres avec une inconcevable rapidité. » Ce qui amène Thibaut Giraud à conclure : « Le *moi* est une fiction, mais c'est une fiction utile⁸. »

Arrêtons-nous sur la liberté pour illustrer cette position fictionnaliste.

La téléologie, impasse scientifique ?

Intuitivement, la liberté repose sur l'existence préalable d'un libre arbitre qui nous permet d'effectuer des choix. Cependant, dans un monde déterministe, nos choix sont prédéterminés par les réactions électro-chimiques à l'intérieur de nos cerveaux. Y compris dans un monde quantique avec interprétation de Copenhague, nos choix sont une conséquence des réactions électro-chimiques et d'événements aléatoires. La notion de libre arbitre n'a pas sa place dans les équations de la physique quantique des champs, la meilleure théorie physique dont on dispose aujourd'hui. Accepter la notion de libre arbitre, c'est rejeter la physique moderne.

⁷  *Identité personnelle (1/2) - Téléportation, trous de mémoire & responsabilité* | Grain de Philo | Monsieur Phi | T. Giraud (2017)

⁸  *Je n'existe pas* | Grain de Philo | Monsieur Phi | T. Giraud (2018)

Le libre arbitre peut être vu comme un cas particulier de la téléologie. La téléologie est un ensemble de théories qui expliquent les phénomènes par leurs finalités. Celle-ci fut notamment défendue par Aristote : « il serait absurde de croire que les choses [de la nature] se produisent sans but, parce qu'on ne verrait pas le moteur délibérer son action. » Dans sa version la plus grandiloquente, la téléologie cherche à expliquer l'univers par son but final, lequel pourrait être, par exemple, l'émergence d'une vie intelligente. C'est ce que certains appellent le principe anthropique fort. Cette position est souvent avancée par les déistes qui y voient une forme de *design intelligent*.

De façon étonnante, on retrouve aussi la téléologie au cœur de la physique quantique des champs, où elle est connue sous le doux nom de principe de moindre action. Découvert par Fermat pour la lumière, puis généralisé par Maupertuis à la matière, puis réutilisé par Hilbert pour la relativité générale, puis étendu par Feynman à la physique quantique, le principe de moindre action consiste *grossièrement* à affirmer que la nature cherche constamment à minimiser une certaine quantité appelée action⁹. Dans le cas de la physique quantique des champs en particulier, ce principe téléologique est utilisé quotidiennement par les physiciens théoriciens !

On retrouve même la téléologie derrière d'autres principes physiques, comme la seconde loi de la thermodynamique qui postule que la nature tend vers l'équilibre thermodynamique, le fait que les électrons d'un atome cherchent à d'abord occuper les niveaux de plus faibles énergies ou que la surface d'une bulle de savon minimise l'énergie de tension de surface.

Il y a même tout un champ de la connaissance qui repose presque exclusivement sur un principe téléologique, à savoir la théorie des jeux. Introduite par des mathématiciens comme John von Neumann et John Nash, la théorie des jeux postule que tout individu se comporte stratégiquement, en agissant de sorte à maximiser son utilité à venir. En particulier, dans les jeux séquentiels comme les échecs, la théorie des jeux postule que les joueurs utilisent les principes de la programmation dynamique. Ce principe algorithmique consiste à partir de la finalité, comme gagner aux échecs ou se retrouver dans une configuration avantageuse, pour remonter le temps et déterminer la marche à suivre pour arriver à ses fins.

C'est ce que le chercheur en géopolitique Bruce Bueno de Mesquita appelle la causalité inverse. Ce ne sont pas les marchés de Noël qui causent Noël, mais Noël qui cause les marchés de Noël. Selon Mesquita, ce type de raisonnement est essentiel pour comprendre les sciences sociales. Par exemple, pour un théoricien des jeux, le rôle des lois et de la justice n'est pas de punir les comportements immoraux, mais de décourager les membres d'une société de se comporter de manière immorale. On ne punit pas parce qu'il y a un crime. On punit pour qu'il n'y ait pas de crime¹⁰.

⁹  Moindre action | Passe-Science | T. Cabaret (2016)

¹⁰  Une justice SANS libre-arbitre ? Démocratie 24 | Science4All | L.N. Hoang (2017)

Pourtant, malgré sa grande utilité en physique et son rôle incontournable en sciences sociales, selon Wikipedia¹¹, « le raisonnement téléologique est rejeté par la méthodologie scientifique moderne en raison du principe de causalité qui implique une relation entre une cause et un effet dans laquelle l'effet ne peut précéder la cause. » Et en effet, un tel principe de causalité est incompatible avec la téléologie. Il semble incohérent d'accepter à la fois l'un et l'autre.

Cependant, contrairement à ce qu'affirment Wikipedia et certains scientifiques, de nombreuses théories scientifiques ne sont pas causales. Et quand bien même elles seraient causales, à l'instar de la théorie des jeux, l'effet peut temporellement précéder la cause. D'ailleurs, même les biologistes évolutionnistes, ceux dont on pourrait penser qu'ils sont au front de la lutte anti-téléologique, parlent souvent de l'intention ou des stratégies des gènes des espèces qu'ils étudient, à l'instar du titre du livre *The Selfish Gene* de Richard Dawkins.

En fait, on a déjà caractérisé l'ensemble des modèles causaux : il s'agit des réseaux bayésiens. *A contrario*, à l'instar des champs de Markov, de nombreux modèles scientifiques ne sont pas causaux. En particulier, pour la relativité générale, l'espace-temps existe tel un bloc, et la physique n'est qu'une description des corrélations entre des événements dans l'espace-temps. On n'y trouve pas de principe de causalité — ou du moins, celui-ci n'est pas un concept fondamental. Tout l'espace-temps existe dans son intégralité, pas seconde après seconde. En fait, la notion même de pas de temps qui cadencerait l'évolution de l'univers est rejetée par la relativité générale, pour qui le passage du temps est une fonction du chemin suivi à travers l'espace-temps.

L'univers est-il donc causal ? Ou faut-il rejeter le principe de causalité ? Il existe en fait deux manières de réconcilier le principe de causalité et la téléologie. Pour comprendre la première de ces réconciliations, il est utile de revenir au principe de moindre action. L'analyse variationnelle, et en particulier les équations d'Euler-Lagrange, prouve que, sous certaines hypothèses, ce principe téléologique est en fait mathématiquement équivalent à une équation différentielle causale¹². De la même manière, sous certaines hypothèses, les équations téléologiques de la programmation dynamique sont équivalentes aux équations causales d'Hamilton-Jacobi-Bellman¹³.

Bien souvent, l'approche téléologique est *isomorphe* à une approche causale¹⁴. Face à ce constat, les défenseurs de la causalité s'empresseroont d'exiger le remplacement de toute approche téléologique par son équivalent causal. Après tout, de nombreuses théories physiques fondamentales peuvent généralement se réécrire sous la forme $\dot{y} = f(y)$. Ou dit autrement, le futur proche est une fonction de l'état présent (avec possiblement une perturbation aléatoire).

¹¹Cette citation provient de la page *Téléologie* de Wikipedia (2018). Elle a été toutefois retirée le 27 mars 2018 par l'utilisateur *AhBon?* pour « contre-sens, qui provient de la confusion entre les différents types de cause ».

¹² [Moindre Action | Passe-Science | T. Cabaret \(2015\)](#)

¹³ [The New Big Fish Called Mean-Field Game Theory | Science4All | L.N. Hoang \(2014\)](#)

¹⁴ [Les isomorphismes | Infini 21 | Science4All | L.N. Hoang \(2017\)](#)

Cependant, dans de nombreux cas, il semble nettement plus naturel de préférer l'approche téléologique. Il est difficile d'imaginer que les stratégies des champions d'échec ne sont pas motivées par la finalité. Il est difficile d'imaginer que les bébés ne pleurent pas pour avoir notre attention. Et il est difficile d'imaginer que les scientifiques ne réfléchissent pas pour mieux comprendre le monde qui les entoure. L'équivalence entre ces histoires téléologiques et des approches causales est loin d'être évidente. Mais surtout, il semble douteux que l'approche causale puisse alors être *utile*.

Voilà qui nous amène à la seconde manière de réconcilier téléologie et causalité. Souvenez-vous. Toute prédiction bayésienne s'obtient en combinant les prédictions de modèles distincts. Autrement dit, pour la *pure bayésienne*, la pluralité de modèles utiles incompatibles est non seulement possible ; elle est même souhaitable ! *Une forêt de modèles incompatibles est plus sage que chacun de ses arbres.* Cultivons donc *toute* la forêt.

Qui plus est, la crédence à assigner à un modèle en particulier dépend de la question à laquelle on cherche à répondre. Cela est d'autant plus le cas pour le *bayésien pragmatique* dont les capacités cognitives sont limitées. S'il s'agit de prédire le prochain coup qu'un champion d'échec sera amené à jouer, la théorie quantique des champs ne lui sera daucune utilité.

Toutefois, étrangement, il est souvent utile d'emprunter des idées d'une théorie pour les réutiliser dans d'autres théories. Par exemple, Richard Feynman a emprunté le principe de moindre action de la mécanique classique pour l'appliquer à la mécanique quantique, avec un succès stupéfiant. Il est alors tentant de croire que ces idées communes à de nombreuses théories crédibles, à l'instar des concepts que tout humain comprend, ont leur propre réalité — au moins à isomorphisme près. Et en effet, la clé de la communication entre humains est l'existence d'activations neuronales similaires dans les cerveaux de différents humains, ce qui me permet de postuler que *mon rouge* est relativement *isomorphe* à *votre rouge*. Voilà qui pourrait suggérer que certains objets ont une réalité indépendante de tout modèle¹⁵.

Ce serait toutefois oublier l'exemple de l'argent. Le fait que certaines sous-procédures algorithmiques sont isomorphes dans deux modèles prédictifs distincts ne garantit aucunement que ces procédures seront présentes dans *tous* les modèles prédictifs, ni même que tout modèle prédictif crédible nécessitera l'usage de telles sous-procédures. L'argent n'existe pas en physique quantique des champs.

Les scientifiques parlent parfois du champ d'applicabilité de leurs théories, ou de théories effectives dans certains contextes. En langage bayésien, il faudrait en fait davantage parler du champ de crédibilité des théories. Mieux encore,

¹⁵ Voilà qui permet d'ailleurs de définir une notion d'*utilité* distincte de celle qu'on a introduite pour le *bayésien pragmatique*. Ici, un modèle non-prédictif peut être *utile* s'il est réutilisé par de nombreuses théories prédictives. C'est en fait en ce sens que les lois de Newton ou la théorie de l'évolution sont *utiles*.

le *bayésien pragmatique* assignera à chaque théorie un certain champ d'utilité. Tout modèle prédictif universel sera une combinaison adéquate d'un grand nombre de théories incompatibles ayant chacune son propre champ d'utilité, qui peut d'ailleurs empiéter sur le champ d'utilité d'une autre. En particulier, l'absence d'universalité de cette approche bayésienne n'a en fait rien de problématique. Après tout, « tous les modèles sont faux. »

Ce que la thèse de Church-Turing dit de la réalité

Les plus puristes d'entre vous pourraient être frustrés par les théories effectives. Certains physiciens persistent à (aimer) penser que leur objectif est la quête d'une vérité.

Il suffit pourtant d'accepter une seule hypothèse pour déterminer les lois fondamentales de l'univers ; ou plutôt, pour déterminer *une* loi fondamentale et complète de l'univers. Cette hypothèse est la thèse de Church-Turing (physique), selon laquelle rien dans l'univers n'est capable d'effectuer un calcul que la machine de Turing ne saurait pas effectuer. En effet, de façon intrigante, accepter la thèse de Church-Turing est équivalent à affirmer que l'univers tout entier peut être simulé par n'importe quelle machine de Turing universelle — et la rejeter ne fait que compliquer la tâche de la quête d'une vérité. En particulier, toute machine dite Turing-complète contient en elle toutes les lois de l'univers¹⁶.

Ce qu'il reste alors à déterminer n'est que les données de la machine, qui rendent le comportement de la machine indiscernable de celui de l'univers. Or, il est clair que déterminer les données de cette machine est une tâche illusoire ! On pourrait imaginer que les données qui décrivent tout l'univers pourront être fortement compressées. Mais, même là, il y a fort à parier que la taille compressée de ces données excédera de loin le googol d'octets. Aucun ordinateur à l'intérieur de l'univers ne pourra alors contenir tout le code source de l'algorithme qui fait tourner notre univers !

En plus d'être illusoire, une telle tâche serait aussi sans intérêt. En effet, la simple lecture de ces données nécessiterait un temps de calcul comparable à l'âge de l'univers. Or, pour analyser notre univers, à l'instar de ce qu'il faut faire pour analyser un code, il faudrait également étudier son exécution. En supposant que l'univers a une grande profondeur logique, ceci prendrait un temps inimaginablement long ! En fait, le théorème de Rice prouve même que l'analyse systématique de codes est un problème indécidable¹⁷.

Mais oublions toutes ces contraintes physiques trente secondes. Que dirait la *pure bayésienne* ? Le démon de Solomonoff ne finirait-il pas par déterminer toutes les lois de l'univers, y compris les données qu'il faudrait fournir à une

¹⁶  [La machine de Turing | IA 4 | Science4All | L.N. Hoang \(2017\)](#)

¹⁷  [Basics of Program Verification | ZettaBytes | V. Kuncak \(2017\)](#)

machine de Turing pour que la simulation initiée par la machine soit alors *exactement* l'univers ?

La réponse est non. Souvenez-vous, en bon bayésien, le démon de Solomonoff ne met jamais tous ses œufs dans le même panier. Même si, après l'analyse de googol d'octets d'informations, ses crédences sont quasiment entièrement sur un seul modèle de l'univers, il ne pourra jamais exclure que son maximum-a-posteriori est *le* modèle de l'univers. En fait, si l'univers veut piéger le démon, il pourra toujours y arriver en choisissant un code dont la complexité de Solomonoff excède la quantité de données qu'il révèle au démon.

Ainsi, même avec une puissance de calculs illimitée, aucune vérité ne peut être acceptée. C'est d'autant plus le cas en pratique. En pratique donc, seules les théories effectives importent, y compris dans le domaine de la physique des particules. Nous ne disposons que de théories effectives, et ces théories effectives ne sont donc nécessairement que des fictions. D'où la conclusion stupéfiante de ce chapitre : *tout est fiction*. Et le corollaire, c'est que savoir revient simplement à déterminer les fictions utiles.

L'antiréalisme (instrumentaliste) est-il utile ?

Même s'il s'agit là d'une conséquence du bayésianisme, cette conclusion pourrait néanmoins être rejetée si vous ne la trouviez pas *utile*. Je ne le pense pas. Il me semble, au contraire, qu'il y a au moins quatre arguments en faveur de l'utilité pratique du fictionnalisme.

Mon premier argument est une clarification de l'intérêt des sciences. De nombreux débats sur l'*utilité* des sciences en viennent à questionner la *vérité* de ses découvertes. Malheureusement, certaines défenses de la *vérité* des sciences abusent d'arguments bancals dès qu'il s'agit de défendre les lois de Newton ou les sciences sociales. En particulier, nombreux cherchent à tracer une ligne entre les sciences et les pseudo-sciences, comme s'il existait une frontière naturelle entre les modèles qui méritent toutes nos crédences, et ceux qui n'en méritent aucun. Cette frontière n'est pas la *vérité* ; ni même la quête de la vérité. L'approche bayésienne clarifie, il me semble, grandement ce débat. « Tous les modèles sont faux. » Mais certains sont plus crédibles et utiles que d'autres. La science consiste alors à identifier des modèles crédibles et à cerner leur champs d'utilité. D'ailleurs, le *peer-review* scientifique semble bien plus juger l'*utilité* des contributions scientifiques que leur *vérité* (ou leur validité).

Mon deuxième argument est la lutte contre l'excès de confiance. On l'a vu, cet excès de confiance est l'un des biais cognitifs les plus récurrents et les plus nocifs. Il nous pousse à nous recroqueviller sur ce qui nous paraît *vrai*, et à ne pas remettre en cause ces *vérités*. Ceci me semble être la principale barrière à l'apprentissage de concepts, de phénomènes et d'explications contre-intuitifs. Pour lutter contre cet excès de confiance, il me semble utile de considérer que

nos théories sont en fait davantage des sortes de marteaux que des *vérités*. Elles peuvent être d'une grande utilité. Mais elles sont aussi potentiellement substituables par de meilleurs outils. Embrasser pleinement cette posture philosophique revient alors à rejeter la *vérité* de *toute* théorie. Voilà qui me semble être un pas indispensable dans la lutte contre l'excès de confiance — même s'il est crucial de rajouter qu'à l'instar des certaines pièces d'une boîte à outils, certaines théories demeurent plus *utiles*.

Mon troisième argument est la lutte contre notre sensibilité exacerbée aux connotations¹⁸. En particulier, les mots « réel » ou « vrai » ont une connotation très positive, comme s'il fallait justifier le fait que Néo avait un devoir moral de prendre la pilule rouge. Le travers de cette approche est que de nombreux tenants de théories pseudo-scientifiques ont un attachement irrationnel à la *vérité* de leurs positions. Toute remise en question de ces tenants par des scientifiques prétendant posséder la *vérité* sera alors nécessairement interprétée comme une attaque personnelle, qui risque donc de transformer le plus modéré des tenants en un hooligan déchaîné¹⁹. Remettre en question la *vérité* d'un modèle qui nous est cher peut être très difficile. Il me semble beaucoup plus raisonnable de remettre en question son *utilité*, quitte à ce que le calcul des crédences ne suivent pas tout à fait la formule de Bayes.

Mon quatrième argument est la continuité de l'apprentissage. Trop souvent, on a tendance à imaginer qu'un étudiant passera d'un état ignorant à un état savant après avoir suivi une leçon, lu un cours ou regardé une vidéo. Cet idéal ne semble toutefois pas du tout s'appliquer en pratique. J'ai appris la formule de Bayes pour la première fois il y a plus de dix ans. Cependant, je continue encore à progresser à pas de fourmi dans sa compréhension — et il me reste encore beaucoup de chemin à faire. L'apprentissage est nécessairement progressif. Donnée après donnée, argument après argument, expériences de pensée après expériences de pensée, nos crédences fluctuent graduellement, de façon généralement non monotone. Et ce n'est qu'après un grand nombre de données, de calculs approximativement bayésiens et de pas aléatoires de MCMC, que nos crédences deviennent raisonnablement fiables²⁰. *L'apprentissage est une danse*. Et cette danse me semble bien mieux orchestrée par la quête de théories *utiles* que par l'ambition de la découverte de la *vérité*.

Bref. Le fictionnalisme me semble *utile*. Mais je ne compte pas finir ma défense du fictionnalisme sur cette note purement instrumentaliste. J'aimerais, pour finir, terminer sur deux modèles qui me semblent mériter au moins autant de crédences que le réalisme. Et pour les introduire, il me faut revenir à Karl Friston.

¹⁸  *Les synonymes à connotations opposées* | My4Cents (Mayen) | Science4All | L.N. Hoang (2016)

¹⁹  *Check tes biais cognitifs #01 / Personne ne bouge !* F. Garcia (2017)

²⁰  *Why you shouldn't try to "change your mind"* | J. Galef (2017)

Y a-t-il un monde extérieur au cerveau ?

Nous disposons de la remarquable faculté de concevoir l'existence des chats, du réchauffement climatique et même de l'histoire de l'univers. Pourtant, pour ce faire, nos cerveaux ne s'appuient en fait que sur les données collectées par la vue, l'ouïe, l'odorat, le toucher, la proprioception, l'équilibrionception, la thermoception et de nombreux autres sens dont nous sommes plus ou moins conscients. Or, ce que ces sens perçoivent semble finalement très distant de ce qu'est vraiment un chat, le réchauffement climatique ou encore l'histoire de l'univers.

Dès 1983, le psychologue et informaticien Geoffrey Hinton, l'un des pères fondateurs du *deep learning*, suggéra avec ses co-auteurs que le cerveau se comportait comme une machine à prendre des décisions à partir d'observations faites par les sens²¹. En 1988, Edwin Jaynes suggéra que la manière dont le cerveau y parvient repose sur la formule de Bayes²². Dans les années 1990, Hinton et Friston développèrent alors un modèle où une couverture de Markov séparait le cerveau du monde extérieur²³, et où le cerveau parvenait malgré tout à reconstruire un modèle de tout le monde extérieur, en utilisant notamment l'inférence bayésienne variationnelle que Friston généralisa à la vie en 2013. L'hypothèse de Friston, Hinton et Jaynes, c'est que nos cerveaux reconstituent tout un modèle du monde extérieur à partir uniquement des données sensorielles.

Le plus surprenant dans cette hypothèse sur le fonctionnement de la pensée humaine, c'est que la construction d'un modèle du monde extérieur n'a pour but que l'explication de ce que perçoivent les sens. Le monde extérieur n'a en fait que bien peu d'importance. Ce monde extérieur n'est qu'une construction de l'esprit. Ce qui importe, ce sont les perceptions des sens, et la capacité du cerveau et de son modèle du monde à prédire les futures perceptions des sens — voire à influencer ces perceptions pour le meilleur.

Selon cette logique, la pensée ne peut être que subjective, dans la mesure où elle est une construction enfermée à l'intérieur d'un esprit — ou, dit autrement, d'une couverture de Markov. Ce qui importe alors, c'est ce que l'on trouve au niveau de cette couverture, et la manière dont on peut expliquer ce qu'il s'y passe. L'hypothèse d'un monde extérieur, à l'instar de celle selon laquelle on vivrait dans une simulation virtuelle comme dans le film *Matrix*, n'a rien de dogmatiquement indiscutable. Par opposition aux philosophes réalistes, la *pure bayésienne* n'a pas de crédence aveugle en une réalité extérieure objective.

²¹ *Massively parallel architectures for A.I.: Netl, Thistle, and Boltzmann machines* | Proceedings of the National Conference on Artificial Intelligence | S. Falhman, G. Hinton et T. Sejnowski (1983)

²² *How Does the Brain Do Plausible Reasoning?* Maximum-Entropy and Bayesian Methods in Science and Engineering | E. Jaynes (1988)

²³ *The free-energy principle: A unified brain theory?* Nat Rev Neuroscience | K. Friston (2010)

Y a-t-il un chat dans un code binaire ?

Il est toutefois tentant de penser que tout ce qui se passe à l'intérieur de nos cerveaux a un pendant à l'extérieur de la couverture de Markov. Votre chat semble bel et bien au moins aussi *réel* que l'image que vous vous en faites. Mais comment pouvez-vous vraiment le savoir ? Tout ce que vous apercevez, ce sont des données que vos sens ont captées, et qui sont corrélées avec ce que vous pensez correspondre à l'existence des chats.

Les simulations de Google sont particulièrement intéressantes pour comprendre cela. L'intelligence artificielle de Google a fini par se construire un concept de chat, lequel s'active si et seulement si (avec grande probabilité) cette intelligence artificielle est exposée à des données qui sont cohérentes avec son concept de chat. Pourtant, tout ce à quoi cette intelligence artificielle a vraiment eu accès, ce sont des données brutes, à savoir un énorme fichier composé uniquement de zéros et de uns. En termes bayésiens, l'existence du chat est utile pour modéliser les nombreuses suites binaires auxquelles l'intelligence artificielle de Google a été exposée ; et c'est pour ça que cette intelligence artificielle y a pensé.

De la même manière, l'hypothèse de Friston, Hinton et Jaynes poussée à son extrême postule que chacune de nos vies est ni plus ni moins la lecture d'un très grand nombre de bits auxquels nos cerveaux sont exposés. Dans ce modèle, et à l'instar de l'intelligence artificielle de Google, nous ne sommes que des têtes de lecture d'un énorme fichier informatique, de quelques zettaoctets peut-être, que nous parcourons à une vitesse faramineuse de plusieurs gigaoctets par seconde.

Ce qui est absolument fascinant, c'est que la lecture de ce fichier par une *pure bayésienne* ou par un *bayésien pragmatique* l'amènera à inventer toutes les fictions que tant d'entre nous prenons pour *vraies*. Ce fichier fantastique est tel le meilleur des livres jamais écrit. Alors que les meilleurs romans nous amènent à imaginer des morceaux de monde fictif, ce fichier fantastique permet à la *pure bayésienne* et au *bayésien pragmatique* de vivre nos vies, avec autant de réalisme que la vie que nous croyons vivre.

En particulier, ce qui rend alors ce fichier fantastique est une propriété algorithmique fascinante dont on a déjà parlé dans le chapitre précédent, à savoir son énorme sophistication de Solomonoff et sa gigantesque profondeur logique de Bennett. D'un côté, l'énorme sophistication de Solomonoff nous a conduit à croire en des modèles où des individus autres que nous existent, à développer des théories sophistiquées et à étudier les mathématiques. De l'autre, la gigantesque profondeur logique nous a poussés à croire que la meilleure façon d'expliquer ce que l'on observe est d'imaginer que l'état présent de l'univers est le résultat d'un long calcul, dont l'origine est un état physique beaucoup moins complexe.

Bien entendu, rien ne garantit que cette version extrême de l'hypothèse du cerveau bayésien est *vraie*. En bons bayésiens, il nous faut constater qu'il ne s'agit que d'une fiction. Or « tous les modèles sont faux. » D'ailleurs, cette

hypothèse n'est finalement pas si crédible aux yeux du démon de Solomonoff. En effet, elle postule l'existence d'un énorme fichier informatique de quelques zettaoctets. Elle est donc peu parcimonieuse.

L'antiréalisme du démon de Solomonoff

En fait, aux yeux du démon de Solomonoff, ce qui a davantage d'existence que cet énorme fichier informatique est la manière dont il a été produit. Souvenez-vous. Le démon de Solomonoff croit en des algorithmes randomisés. En particulier, le fichier informatique en lui-même n'est que l'artefact d'algorithmes randomisés plus fondamentaux qui, à l'aide du hasard, ont généré ce fichier informatique. Pour le démon de Solomonoff, n'existent alors vraiment que ces différents algorithmes, ces espèces de générateurs de fictions, et le hasard indescriptible sur lequel ces algorithmes se sont appuyés.

Tous les algorithmes n'ont toutefois pas le même degré d'existence. Le démon de Solomonoff va supposer que certains algorithmes sont plus crédibles que d'autres, et va constamment chercher à ajuster ces crédences, à l'aide de la formule de Bayes. Autrement dit, de manière grossière, le démon de Solomonoff croit uniquement en l'existence d'une superposition d'algorithmes et de hasard. Dès lors, les étapes intermédiaires de calculs de ces algorithmes, les fictions qu'ils racontent, ont elles aussi une certaine existence, même si cette existence est moins fondamentale que les algorithmes eux-mêmes. Tel est, il me semble, la conclusion stupéfiante du *bayésianisme*.

Je n'exclus toutefois pas la présence de failles dans ce raisonnement. Ceci me pousse, en bon bayésien, à ne pas mettre toutes mes crédences en cette étrange approche de la réalité. J'espère toutefois avoir créé une crédence non nulle en vous en la possibilité que tout soit fiction ; à l'exception peut-être des algorithmes qui nous racontent ces fictions.

Références en français

- ▶ *L'Allégorie de la Caverne de Platon* | Le Coup de Phil' | Cyrus North (2013)
- ▶ *Moindre Action* | Passe-Science | T. Cabaret (2015)
- ▶ *Qui crée l'argent ? Comment ?* Heu?reka | G. Mitteau (2015)
- ▶ *Le double visage de l'argent* | Heu?reka | G. Mitteau (2016)
- ▶ *Adam Smith - le paradoxe de la veste de laine* | Grain de Philo | Monsieur Phi | T. Giraud (2016)
- ▶ *Adam Smith - division du travail & main invisible* | Grain de Philo | Monsieur Phi | T. Giraud (2017)
- ▶ *Identité personnelle (1/2) - Téléportation, trous de mémoire & responsabilité* | Grain de Philo | Monsieur Phi | T. Giraud (2017)

- ▶ *Identité personnelle (2/2) - Montez-vous dans le téléporteur ?* | Grain de Philo | Monsieur Phi | T. Giraud (2017)
- ▶ *La machine à expérience de Robert Nozick* | Argument frappant | Monsieur Phi | T. Giraud (2018)
- ▶ *Je n'existe pas* | Grain de Philo | Monsieur Phi | T. Giraud (2018)
- ▶ *La conscience* | Science Étonnante | T. Giraud et D. Louapre (2017)
- ▶ *Check tes biais cognitifs #01 / Personne ne bouge !* F. Garcia (2017)

- ▶ *Les synonymes à connotations opposées* | My4Cents (Mayen) | L.N. Hoang (2016)
- ▶ *Les isomorphismes* | Infini 21 | Science4All | L.N. Hoang (2017)
- ▶ *Une justice SANS libre-arbitre ?* Démocratie 24 | Science4All | L.N. Hoang (2017)
- ▶ *La Grande Histoire des petites histoires* | Démocratie 26 | Science4All | L.N. Hoang (2017)
- ▶ *La machine de Turing* | IA 4 | Science4All | L.N. Hoang (2017)
- ▶ *L'émergence de l'intelligence* | IA 5 | Science4All | T. Cabaret et L.N. Hoang (2018)

Références en anglais

- 📘 *The Selfish Gene* | Oxford University Press | R. Dawkins (1976)
- 📘 *Predictions: How to See and Shape the Future with Game Theory* | Vintage | B. Mesquita (2010)
- 📘 *Sapiens: A Brief History of Humankind* | Harper | Y.N. Harari (2015)
- 📘 *The Big Picture: On the Origin of Life, Meaning and the Universe Itself* | Dutton | S. Carroll (2016)

- .ribbon *Massively parallel architectures for A.I.: Netl, Thistle, and Boltzmann machines* | Proceedings of the National Conference on Artificial Intelligence | S. Falhman, G. Hinton et T. Sejnowski (1983)
- ribbon *How Does the Brain Do Plausible Reasoning?* Maximum-Entropy and Bayesian Methods in Science and Engineering | E. Jaynes (1988)
- ribbon *The free-energy principle: A unified brain theory?* Nat Rev Neuroscience | K. Friston (2010)
- ribbon *Life as we know it* | Journal of the Royal Society Interface | K. Friston (2013)
- ribbon *The New Big Fish Called Mean-Field Game Theory* | Science4All | L.N. Hoang (2014)
- ▶ *Why you shouldn't try to "change your mind"* | J. Galef (2017)
- ▶ *Basics of Program Verification* | ZettaBytes | V. Kuncak (2017)
- ▶ *The Physics of Life (ft. It's Okay to be Smart & PBS Eons!)* | PBS Space Time | M. O'Dowd (2018)

Douter de tout ou tout croire sont deux solutions également commodes, qui l'une et l'autre nous dispensent de réfléchir.

Henri Poincaré (1854-1912)

Le premier principe est que vous ne devez pas vous duper — et vous êtes la personne la plus facile à duper.

Richard Feynman (1918-1988)

21

Aux origines des croyances

Le scandale des séries divergentes

On est le 6 juin 2013. Minuit a sonné il y a six minutes. Je viens de lire l'un des articles les plus troublants qui soient¹. David Louapre, pourtant détenteur d'une thèse en physique théorique, vient d'écrire un billet sur son blog *Science Étonnante* où il semble prouver que la somme des entiers strictement positifs est égale à $-1/12$. Ou dit autrement, $1 + 2 + 3 + 4 + \dots = -1/12$. Je suis à la fois fasciné et profondément confus. Il me faut laisser un commentaire.

J'écris : « Super article ! Je suis troublé par le fait que l'on cherche à construire des sommes, au lieu de chercher à étendre leur définition en utilisant des règles de manipulation des séries telles que celles que vous utilisez. N'est-il pas possible de montrer que, moyennant certaines règles comme la linéarité et l'ajout de zéros dans les séries, il existe une unique sommation naturelle pour toutes les séries que l'on peut obtenir avec ces règles ? Ceci justifierait bien mieux à mes yeux le résultat $1 + 2 + 3 + 4 + \dots = -1/12$ que des sommes de Césaro, qui paraissent encore trop restrictives, ou des techniques de prolongement analytique, dont on peut avoir l'impression qu'[elles] donnent un résultat arbitraire. »

Trois heures plus tard, David Louapre répond : « Tout juste, c'est bien ça qui se passe. On cherche un opérateur S agissant sur l'espace des suites qui soit linéaire, stable par ajout d'un nombre fini de zéro [sic] en début de la suite et coïncidant

¹  $1 + 2 + 3 + 4 + 5 + 6 + 7 + \dots = -1/12$! Science Étonnante | D. Louapre (2013)

avec la définition usuelle pour les séries absolument convergentes. Si on suppose que cet opérateur existe ALORS les manipulations un peu heuristiques que je fais dans le billet sont licites, et on trouve que $-1/12$ est la seule valeur possible pour $1 + 2 + 3 + 4 + \dots$ mais il faut encore prouver que l'opérateur existe (sur certaines suites du moins), d'où l'intérêt des approches type Césaro ou prolongement analytique. »

De passionnantes discussions prolifèrent alors dans les commentaires du billet de David Louapre. La réplique la plus spectaculaire est celle de Rémi Peyre, qui prouve qu'il « n'existe aucune méthode [linéaire, régulière et stable] de sommation de séries divergentes qui permette de donner une valeur finie à [la somme des entiers strictement positifs]. » Quelques semaines plus tard, je publiai à mon tour un billet sur mon blog², avec une preuve que, si les manipulations de David Louapre ne permettent pas de conclure à $1 + 2 + 3 + 4 + \dots = -1/12$, elles permettent toutefois de conclure $1 + 2 + 4 + 8 + 16 + \dots = -1$.

Trois ans plus tard, le 8 septembre 2016, je mis en ligne une vidéo³ qui prouva que toutes les séries définissables par récurrence linéaire non barycentrique à une série convergente près peuvent être sommées de manière unique. C'est ce que j'ai appelé la supersommation linéaire, régulière et stable. Elle permet de prouver de nombreuses égalités stupéfiantes, comme $1 - 1 + 1 - 1 + \dots = 1/2$, $3 + 9 + 27 + 81 + \dots = -3/2$ et $2 + 3 + 5 + 8 + 13 + 21 + \dots = -3$. Qui plus est, en fin de vidéos, j'ai émis la conjecture que toutes ces séries et uniquement ces séries pouvaient être sommées ainsi. Plusieurs de mes (adorables) abonnés se sont précipités sur ce problème et en ont écrit des preuves rigoureuses !

J'adore cette histoire, parce qu'elle illustre parfaitement la curiosité caractéristique d'un (bon) chercheur. L'étrangeté d'un résultat nourrit sa soif de savoir, à l'instar de tant de physiciens déçus par la trop grande prévisibilité du boson de Higgs. Mais surtout, le chercheur se hâte de ne pas conclure. Il va questionner le fondement du résultat, ainsi que les fondements de son intuition. Comme l'affirma Isaac Asimov, « l'expression la plus excitante des sciences, celle qui anticipe les nouvelles découvertes, n'est pas "Eurêka !", mais "c'est marrant". »

Malheureusement, ce ne fut pas la réaction de tous mes auditeurs. « Vous n'avez pas posé votre équation logiquement ! » « Je trouve ça stupide les calculs avec l'infini. » « Bon, ben puisque c'est comme ça je vais prendre la racine d'un nombre négatif et diviser par zéro. » « Ça sert à rien les sommes infinies. » Le billet de David Louapre a eu de nombreuses réactions similaires. « Ce post est d'une stupidité aberrante. » « J'ai ri de ces prétendues "démonstrations" rigoureuses. » « Ce n'est pas de la science étonnante ce sont des pseudo-démonstration [sic] menées n'importe comment ! » David Louapre fut surpris par la violence de ces réactions. « Je ne pensais pas que ce billet provoquerait une telle levée de boucliers », écrivit-il.

²  *The Surprising Flavor of Infinite Series* | Science4All | L.N. Hoang (2013)

³  *La supersommation linéaire, stable et régulière* | Hardcore | Science4All | L.N. Hoang (2016)

Mais c'est faux, non ?

Je vous invite à y réfléchir. Avez-vous vous aussi sauté au plafond au moment de lire $1+2+3+4+\dots = -1/12$? Et si je vous disais que le théorème de Pythagore était faux ? Que π était une imposture ? Que la gravité n'existe pas ? Que le sol accélère vers le haut ? Que les OGM obtenus par CRISPR étaient plus sains pour nous et pour la biodiversité que les méthodes d'hybridation ? Que des physiciens ont réussi des téléportations (des états quantiques) de photons ? Qu'il existe des ensembles ni finis ni infinis ?

Rejeter les hypothèses étranges et contre-intuitives n'est pas un mauvais réflexe — quoiqu'avoir une réaction épidermique au simple fait d'être exposé à ces hypothèses est plus gênant. Si je vous dis que j'ai gravi un sommet de l'Himalaya sans difficulté, je ne vous reprocherai pas de ne pas me croire. J'ai même fait l'apologie des préjugés quelques chapitres plus tôt. Il ne faut pas perdre son temps à méditer trop longuement des réflexions dont on a de bonnes raisons de croire qu'elles ne mèneront nulle part.

De la même manière, des génies et des grandes institutions du passé ont violemment rejeté des idées qui leur semblaient trop contre-intuitives pour être crédibles. On dit ainsi des pythagoriciens qu'ils auraient noyé le pauvre Hippase de Métaponte, parce que celui-ci avait prouvé l'irrationalité du nombre $\sqrt{2}$. En 1632, l'ordre catholique des Jésuites choisit de bannir le calcul infinitésimal des mathématiques⁴. À la fin du XIX^e siècle, les ensembles infinis de Georg Cantor ont suscité la moquerie de ses contemporains, notamment les critiques violentes de Leopold Kronecker⁵, qui utilisa des injures comme « charlatan », « renégat » et « corrupteur de la jeunesse ». Dans les années 1970 encore, la nouvelle géométrie très rugueuse et si inhabituelle des fractales de Mandelbrot allait être vivement critiquée à son tour par de nombreux mathématiciens reconnus, pour qui la vraie géométrie était lisse, continue et différentiable.

Cependant, dans tous ces exemples, les rejets par la communauté mathématique n'ont pas été définitifs. Petit à petit, la communauté mathématique a même fini par changer d'avis, et par encenser ce qu'elle avait pu incendier dans le passé. De nos jours, l'irrationalité des $\sqrt{2}$, les calculs différentiels de Leibniz, les infinis de Cantor et les fractales de Mandelbrot sont tous célébrés comme des joyaux des mathématiques. Même l'équation $1 + 2 + 3 + 4 + \dots = -1/12$ a fini par être défendue par des mathématiciens de tout premier rang, comme Srinivasa Ramanujan, G.H. Hardy ou Terence Tao⁶. Comment est-ce possible qu'en mathématiques, cette science dont on dit parfois qu'elle sait lever le doute et discerner le vrai du faux, de tels changements d'avis aient eu lieu ? Comment les mathématiciens ont-il pu avoir eu tort ? Et qu'est-ce qui leur a permis de changer d'avis ?

⁴  *La quête mathématique de l'infiniment petit | Infini 7 | Science4All | L.N. Hoang (2016)*

⁵  *The Limitless Vertigo of Cantor's Infinite | Science4All | L.N. Hoang (2015)*

⁶  *The Euler-Maclaurin formula, Bernoulli numbers, the zeta function, and real-variable analytic continuation | T. Tao (2010)*

« Une nouvelle vérité scientifique ne triomphe pas en convainquant ses opposants et en les éclairant, mais plutôt parce que ces opposants finissent par mourir, et une nouvelle génération grandit en étant familière avec elle », affirma le physicien Max Planck. S'appuyant sur des décennies de psychologie expérimentale, les psychologues Dominic Johnson et James Fowler rajoutent que « les humains ont de nombreux biais psychologiques, mais l'un des plus récurrents, des plus puissants et des plus répandus, est l'excès de confiance ». Les scientifiques n'échappent pas à cet excès de confiance.

Trop souvent, trop peu font l'effort de comprendre les raisons pour lesquelles d'autres concluent ce qu'ils ont conclu. Plus rares encore sont ceux qui ont médité les raisons pour lesquelles ils pensent ce qu'ils pensent — et l'un des principaux objectifs de ce livre est justement de vous amener à méditer les raisons pour lesquelles vous pensez ce que vous pensez ! J'ai moi-même mis beaucoup de temps avant de méditer les raisons pour lesquelles je pense ce que je pense. Par chance, divers événements dans ma vie ont soulevé cette question. J'espère que l'exemple de mes réflexions personnelles soulèvera cette question chez vous aussi.

Élève officier

Après avoir eu la chance d'intégrer l'École Polytechnique, je fus envoyé à l'École des officiers de l'armée de terre, à Saint-Cyr, par la *magouilleuse* — je vous promets que je voulais y aller et que je ne suis pas tombé dans le piège de la *magouilleuse* ! Ce fut une expérience douloureuse. Se lever tous les matins bien avant le lever du soleil pour récurer les toilettes, marcher au pas pendant des heures en chantant à tue-tête et passer la nuit dans des trous de combat en décembre sous la pluie dans la forêt bretonne. Ce fut dur.

Mais, avec le recul, ce fut une expérience instructive. En quelques semaines seulement, j'avais acquis, sans vraiment complètement m'en rendre compte, toutes les mimiques, les valeurs et les sophismes des militaires. Je me mis à parler sèchement, à sacrifier les symboles patriotiques et à exagérer l'importance du pliage des vêtements au format A4. Je parlai de l'art de commander à longueur de journée et j'affirmai qu'il valait clairement mieux décider plutôt qu'hésiter, comme s'il s'agissait là d'une évidence. Mais le pire dans tout ça, c'est que je ne me rendais pas compte que c'était le contexte qui m'avait poussé vers ces nouvelles convictions. Je ne comprenais pas pourquoi j'en étais venu à croire ce que je croyais. Je n'avais pas idée d'à quel point mon environnement avait déterminé mes pensées.

Fort heureusement, ma courte carrière militaire arriva rapidement à une fin. De retour à l'École Polytechnique, je découvris un tout autre contexte, celui de la vie étudiante. Je fus victime de nouvelles mimiques, d'autres valeurs et de différents sophismes. Mais cette fois-ci, j'eus le bénéfice de la remise en

question. Ces nouvelles valeurs étaient ainsi souvent en contradiction avec mes valeurs militaires, ce qui me permit, peut-être pour la première fois de ma vie, d'enfin me poser la question de l'origine de mes convictions. Cependant, ce fut bref et largement insuffisamment.

Au fil des années, cependant, de temps à autre, je me surpris à découvrir les origines cachées de mes convictions. Je compris que mes aspirations à commander, diriger et prendre des responsabilités étaient nourries par les nombreux discours destinés aux étudiants polytechniciens, qu'ils soient donnés par des conférenciers ou des militaires de haut rang. On avait des cours de *leadership*, on nous encourageait à porter des projets à plusieurs et on y faisait la gloire des entrepreneurs. À tel point qu'en sortant de l'École Polytechnique, j'acceptais de ne faire des mathématiques que si celles-ci étaient appliquées, et si elles me permettraient un jour d'accéder à des postes haut placés en entreprise.

Avec le recul, je ne peux que me considérer incroyablement chanceux d'être tombé ensuite dans l'excellent Groupe d'étude et de recherche en analyse des décisions (GERAD) à l'École Polytechnique de Montréal. Dans ce nouvel environnement qui promouvait la recherche, les mathématiques et le travail de réflexion, j'acquis tout à coup de nouvelles valeurs et convictions.

Vivre à Montréal plutôt qu'en France a fortement modifié ma consommation de l'actualité médiatique. Je n'avais plus de télévision, et je passais de moins en moins de temps à lire les journaux français sur le web. Ce fut dès lors beaucoup plus facile de lister toutes mes sources d'information de la politique française, et de comprendre la manière dont ces sources affectaient mes convictions.

Un jour, fin 2011, alors que la primaire socialiste pour l'élection présidentielle de 2012 approchait, je me surpris à affirmer que Martine Aubry ne semblait pas aimable. Je fus également agréablement surpris d'être surpris. Enfin m'étais-je rendu compte d'un doute dans mes convictions, et en une fraction de secondes, je sus exactement pourquoi j'avais de tels convictions : je regardais encore l'émission télévisée *les Guignols*. Je fus aussi surpris de voir que les gens à qui je parlais se mirent tous à acquiescer à l'unisson. Ces autres personnes en savaient-elles plus que moi ? Ou étaient-elles tout autant victimes, directement ou indirectement, des sous-entendus des émissions télévisées ? Au final, qui connaît vraiment les hommes politiques ?

Mon périple en Asie

Le début de l'année 2012 fut l'un des moments les plus marquants de mes réflexions personnelles. Avant de commencer ma thèse, j'étais parti faire un tour de l'Asie de l'Est, sac sur le dos, pendant un mois et demi, avec un ami. Ce fut un périple fantastique, tant géographiquement et humainement qu'intellectuellement.

Parmi les événements qui me marquèrent le plus, il y eut la lecture d'un article de journal en Chine, écrit en anglais, qui discutait de la condamnation à mort d'une chef d'entreprise pour levée de fonds illégale. Ce qui me stupéfia, ce fut la position très réservée et balancée de l'article. Une moitié de l'article justifia la condamnation, l'autre la condamna. J'en fus déboussolé. La quasi-totalité des documents que j'avais lus jusque-là au sujet de la peine de mort n'avait pas fait preuve d'autant de réserve. Pire, l'Éducation nationale m'avait baigné dans un militantisme contre la peine de mort. Elle faisait la gloire de l'œuvre de Victor Hugo, dont l'ouvrage *Le dernier jour d'un condamné* est l'un des livres qui m'avaient le plus marqué.

Je ne tiens pas ce discours pour défendre la peine de mort — le bannissement de la formule de Bayes des cours de justice et les travaux de Julia Shaw me semblent être des arguments frappants contre la peine de mort. Si je vous raconte cela, c'est pour partager ma soundaine prise de conscience du fait que mes positions idéologiques avaient été guidées par un militantisme univoque permanent. Si je pensais ce que je pensais, c'est parce que mon environnement social, culturel et éducatif m'avait poussé à le penser. Ironie de l'histoire, c'est la Chine, un pays dont la pensée libre est combattue et contrôlée par un État puissant, qui m'aura enfin permis de m'en rendre compte. L'École française et mes compatriotes français avaient largement décidé de mes propres convictions.

Quelques jours plus tôt, mon ami et moi nous étions trouvés dans une situation embêtante. Nous venions d'arriver à Tunxi pour explorer les montagnes de Huangshan. Cependant, quelques jours après le nouvel an chinois, les trains de Tunxi à Hong-Kong étaient pleins. Le problème, c'est que nous avions déjà réservé des billets d'avion partant de Hong-Kong trois jours plus tard. Qui plus est, la seule phrase que je connaissais en chinois était « je ne parle pas chinois ».... Dans la précipitation, après plusieurs péripéties et à l'aide de mimes désespérés, on finit par acheter des billets d'avion à l'aéroport de Tunxi pour plusieurs centaines d'euros. Dégoutés d'avoir dû payer si cher, on reprit notre route vers les montagnes de Huangshan.

Pour ce faire, il nous fallut prendre un van. Celui-ci exigea 160 yuans, soit environ 20 euros. C'était, pour la Chine à cette époque, un prix étonnamment élevé. Faute d'options alternatives, on accepta. Le van partit. Assis dans le van, vint ensuite le moment de payer. Et c'est en voyant les autres passagers payer que l'on comprit tout à coup l'erreur de communication ! Ce n'était pas 160 yuans qu'il nous fallait payer, mais 16 yuans, c'est-à-dire seulement 2 euros !

Je me souviens distinctement de l'énorme sourire qui nous vint alors au visage, comme si nous venions de signer l'affaire du siècle ! Pourtant, cette joie sincère était aussi d'un ridicule flagrant. On venait de perdre des centaines d'euros, mais c'est l'illusion d'une économie de quelques dizaines d'euros qui avait déterminé notre nouvelle humeur ! Pire encore, quelques jours avant, à Pékin, on avait déjà perdu une cinquantaine d'euros en tombant dans le piège de la fameuse *arnaque du thé* ! J'avais découvert l'effet d'ancrage et la forte dépendance de mon humeur aux normes imposées par nos sociétés.

Tous des monstres en puissance ?

Deux semaines plus tard, au cours de ce même voyage en Asie, on visita le musée de la guerre du Vietnam à Hô-Chi-Minh. On y vit d'atroces photographies et d'ignobles récits de guerre de soldats américains. En quelques années et sur un territoire relativement restreint, plus de bombes furent larguées que pendant toute la seconde guerre mondiale. Les soldats américains n'avaient guère de remords à tuer des civils — certains y prenaient même du plaisir. Enfin, pour empêcher les guerriers nord-vietnamiens de se cacher dans la forêt, les américains larguèrent des quantités massives d'agent orange. Des décennies plus tard, cet agent orange est encore la cause d'horribles tragédies, car il cause d'épouvantables malformations des enfants des guerriers vietnamiens qui y furent exposés. Pourtant, les studios hollywoodiens n'hésitent pas à faire l'éloge des soldats américains qui y furent envoyés.

Quelques jours plus tard, ce fut le tour de la visite d'une école transformée en prison par les Khmers rouges à Phnom Penh. On découvrit alors l'un des pires épisodes de l'histoire de l'humanité. Juste après la guerre du Vietnam, les forces communistes cambodgiennes, emmenées par Pol Pot, reprirent le pouvoir de force. S'installa alors un régime dur et violent, et une chasse contre les positions pro-américaines. Tous les jours, des milliers de Cambodgiens furent torturés par les autorités, jusqu'à ce qu'ils avouent avoir conspiré contre le régime communiste et qu'ils soient tués. Certaines études estiment à des millions le nombre de victimes du régime, ce qui représente le décès d'au moins un cinquième de la population en l'espace de 4 ans — certaines sources vont jusqu'à parler du décès d'un tiers de la population !

La prison que l'on visita présentait des témoignages de certaines victimes, mais aussi de certains militaires en charge de la torture des gens suspectés d'être pro-américains. Je fus interloqué par la faiblesse des remords dans ces témoignages. Ces bourreaux arrivaient-ils à dormir ou à se regarder dans une glace ? Quelles sont les valeurs morales et éthiques des soldats des Khmers rouges ? Y a-t-il une limite à l'atrocité humaine ? Aurais-je tué et torturé comme ils l'ont fait, si j'avais été à leur place ? Et si j'avais été Américain, aurais-je pris plaisir à massacrer des Vietnamiens ?

Il est tentant de penser que l'on se serait comporté différemment et que l'on aurait fait preuve de davantage d'empathie et de justice. Cependant, la psychologie comportementale suggère fortement qu'il ne s'agit là que d'une illusion. Dans une expérience devenue célèbre, dans les années 1960, Stanley Milgram demanda ainsi à des sujets A de punir d'autres sujets B pour de mauvaises réponses à des questions. Sujets A et B communiquaient via téléphone. Les punitions étaient des chocs électriques. De façon troublante, 2 sujets A sur 3 suivaient presque aveuglément les ordres d'un scientifique qui exigeaient d'augmenter l'intensité des chocs électriques jusqu'à dépasser le seuil considéré mortel. Les sujets A n'avaient pas l'excuse de l'ignorance : les intensités mortelles étaient clairement accompagnées de symboles de tête de mort. Qui plus est, les sujets B étaient des

acteurs en communication téléphonique qui criaient et pleuraient en agonisant. Néanmoins, 67 % des sujets administrèrent (ou, plutôt, pensèrent avoir administré) des chocs mortels, juste parce qu'on leur en avait donné l'ordre. Cette expérience a même été récemment reproduite par Darren Brown pour l'émission *The Heist*, avec des résultats similaires.

La majorité d'entre nous obéit rapidement à la pression sociale et à l'autorité. Et s'il y a certes des cas où on commence par résister, avec le temps, on finit généralement par céder. Aussi décevant, frustrant et contre-intuitif que cela puisse sembler, je ne vois pas de raison de penser que je suis une exception. Il me semble malheureusement quasiment certain que, dans ces situations inhabituelles, moi aussi, j'aurais (pensé) administré des chocs électriques mortels, j'aurais torturé les potentiels pro-Américains et j'aurais pris plaisir à massacer des Vietnamiens.

Par chance, le hasard m'a fait vivre dans d'autres conditions.

Les histoires ont plus d'effet que les chiffres

Les statistiques des sciences cognitives ne plaident pas en faveur de notre bonté à toute épreuve. Cependant, malheureusement, trop souvent, on a du mal à retenir de telles statistiques. On a encore plus de mal à appliquer la formule de Bayes pour ajuster nos crédences bayésiennes en fonction des statistiques. Mais surtout, on ne pense pas qu'elles s'appliquent aux gens que l'on connaît — et encore moins qu'elles puissent s'appliquer à nous.

C'est ce que montrent les expériences des psychologues Richard Nisbett et Eugene Borgida. L'un des faits bien établis par les sciences cognitives est la diffusion de la responsabilité. En particulier, une expérience classique montre que seulement 27 % des sujets sont prêts à aller aider une victime de convulsions, notamment lorsqu'ils savent que d'autres sont susceptibles d'aller l'aider. C'est un fait surprenant ! On a tendance à penser qu'une plus grande proportion d'entre nous serait venue en aide à une personne en danger.

Mais ce n'est pas ce phénomène que Nisbett et Borgida étudièrent. Ils voulaient savoir si les étudiants qui connaissaient ce chiffre seraient meilleurs pour deviner si des individus interviewés seraient allés aider la victime. À leur grand étonnement, la réponse est non. Avoir appris un fait statistique surprenant ne suffit pas à améliorer les prédictions que cette statistique devrait pourtant aider à effectuer. Les étudiants avaient pourtant bien retenu ce chiffre — ils étaient capables de le réciter à l'examen — mais ils n'étaient pas capables de l'appliquer en pratique.

De façon tout aussi étrange, c'est en présentant des cas particuliers d'individus qui ne sont pas allés aider la victime que les étudiants réussirent à internaliser les statistiques de la diffusion de la responsabilité. Nisbett et Borgida conclurent

leur recherche par une remarque que tout apprenant, pédagogue ou communiquant se doit de méditer : « La réticence des sujets à déduire le particulier du général n'a d'égal que leur empressement à inférer le général du particulier. »

Ainsi, quand je discute de politique, de psychologie ou de sociologie avec des amis, il arrive souvent que mes amis utilisent leurs expériences personnelles et celles de leurs proches pour confirmer ou infirmer les analyses scientifiques. Pourtant, ces cas particuliers, souvent teintés de fortes émotions et de remémorations sujettes aux faux souvenirs, ne font absolument pas le poids devant l'analyse statistique des politologues, des sociologues et des psychologues. Cette nécessité de se défaire du particulier pour inférer des généralités est l'une des difficultés majeures des sciences sociales.

Cependant, ce qui est troublant, c'est qu'enseigner ces généralités n'a que très peu d'effets sur les crédences des apprenants. Pour vraiment questionner les croyances des étudiants, il faut présenter des cas particuliers. C'est pour cela que ce chapitre est dédié à l'exemple de *ma* quête des origines de mes croyances, et qu'il révèle les biais et limites de *mes* cognitions. J'espère que mon cas particulier vous aidera à mieux généraliser les limites cognitives de ceux qui vous entourent. Et à anticiper vos propres limites cognitives.

Même armés de nombreuses anecdotes, déterminer nos limites cognitives demeure toutefois extrêmement difficile. Malgré toute l'expertise acquise, même le grand Daniel Kahneman ne fut pas exempt de cette incapacité à déduire le particulier du général. Après avoir convaincu les autorités d'enseigner la psychologie de la prise de décision au lycée, Kahneman réunit une équipe pour écrire le curriculum de ce nouveau cours. Kahneman demanda à tous les membres de son équipe de lui envoyer une estimation du temps requis à l'écriture d'un livre complet à ce sujet. Les réponses allèrent d'un an et demi à 2 ans et demi.

Kahneman demanda ensuite à Seymour, un expert en curriculum, le temps qu'avaient mis d'autres équipes se lançant dans des projets similaires. Seymour fut gêné. Il fut gêné d'avoir donné une réponse déconnectée des statistiques qu'il connaissait pourtant très bien. Seymour affirma qu'environ 40 % des projets similaires avaient échoué, et que tous les autres avaient mis au moins 7 ans à être accomplis ! Seymour ajouta même que les autres projets semblaient, en moyenne, au moins aussi bien partis que celui de Kahneman. L'équipe de Kahneman avait gravement sous-estimé l'ampleur de la tâche.

Mais ce n'est pas là le plus surprenant. Maintenant que Kahneman avait ces nouvelles données, il aurait dû mettre à jour ses crédences, et se rendre compte que le projet allait être une énorme perte de temps, avec une grande probabilité d'échec. Ceci aurait dû l'amener à abandonner ce projet. Mais abandonner un projet est horriblement difficile⁷. Même Kahneman, tout expert en prise de décision qu'il est, ne prit pas le temps de considérer le très probable échec du projet. Il chercha tant bien que mal à mener son projet à bout. Le livre fut fini 8 ans plus tard, par un successeur de Kahneman. Il ne fut jamais utilisé.

⁷  *Les coûts irrécupérables | Crétin de cerveau | Science Étonnante | D. Louapre (2016)*

L'un des nos biais récurrents est le biais du *raisonnement motivé*. Bien souvent, nous partons de la conclusion que nous voulons atteindre, et notre raison n'est alors plus qu'un outil pour nous conforter dans cette conclusion que nous avons déjà acceptée. Kahneman a voulu croire en son projet. Et il n'a pas hésité à ignorer les statistiques qui auraient dû l'en faire douter.

Le psychologue Jonathan Haidt résume cela à travers l'expression « l'intuition d'abord, la raison ensuite ». Dans tout débat, nous prenons d'abord parti. Ce n'est qu'ensuite que nous justifions notre décision. Le raisonnement vient *a posteriori*. C'est ce que l'on appelle la *rationalisation*, et selon Haidt, il s'agit de la manière dont nous pensons à longueur de journée. Nous réfléchissons tels des avocats dont le client est notre intuition.

Le problème, c'est que ce faisant, on a tendance à exagérer les arguments en notre faveur, et à balayer d'un revers de la main ceux qui pourraient mettre en danger notre conclusion. On va typiquement être beaucoup plus critique envers les sources d'information qui remettent en cause notre conclusion, en questionnant notamment les compétences et les motivations de la source d'information. C'est le fameux biais de sélection. Or, à l'ère du web, Google nous permettra toujours de trouver un blog ou une vidéo qui confirme notre conclusion, quelle que soit cette conclusion.

Les superstitions

La rationalisation de l'intuition suffit d'ailleurs à expliquer l'émergence de superstitions. Le psychologue Burrhus Frederic Skinner l'a parfaitement illustré avec l'amusante expérience des superstitions de pigeons. Comme on l'a vu, Skinner a enseigné à des pigeons à lire des mots comme « donne un coup de bec » ou « fait un tour sur toi-même », en récompensant les pigeons ayant suivi l'instruction écrite. De façon remarquable, ces pigeons apprirent rapidement à lire.

Mais là où l'expérience de Skinner devint vraiment intrigante, c'est au moment où Skinner laissa ses pigeons dans une cage, sans instruction, mais avec une récompense qui tombait à des instants aléatoires. Les pigeons tentèrent différents mouvements, en espérant que ces mouvements seraient la cause de la récompense. Bien entendu, quand la récompense tombait, les pigeons venaient d'effectuer, ou étaient en train d'effectuer, un mouvement particulier. Et bien, les pigeons sur-interprétèrent la corrélation qu'ils avaient constatée, et se mirent à croire que ce mouvement causait la récompense. Ils répétèrent alors ce mouvement, encore et encore. Mais plus ils répétaient ce mouvement, plus grande était la probabilité que la récompense tombe à un moment où ils effectuaient ce mouvement, renforçant ainsi la croyance erronée des pigeons. Les pigeons de Skinner avaient appris des superstitions !

On pourrait avoir envie de croire que les pigeons sont idiots. Mais je vous rappelle qu'ils surpassent les humains au jeu de Monty Hall ! Il y a donc fort

à parier que le mécanisme à travers lequel les pigeons peuvent apprendre des superstitions s'applique à l'homme aussi. En 1985, Tversky, Gilovich et Vallone ont ainsi montré que le mythe de la « main en feu » au basketball, selon lequel les joueurs peuvent avoir des jours exceptionnels où tout leur réussit, n'est en fait qu'une superstition. L'analyse statistique des trois chercheurs montrait au contraire que les réussites des joueurs suivaient toutes les lois les plus basiques du pur hasard ; en particulier, une suite de variables indépendantes aura des chances étonnamment grandes de donner lieu à de longues séries de valeurs identiques⁸.

Dès lors, la présence de superstitions dans les sociétés humaines ne nécessite pas d'explications surnaturelles ; les lois statistiques, combinées à notre incapacité à appliquer adéquatement une version, même approximative, de la formule de Bayes, suffisent à prédire la présence de nombreuses superstitions. C'est pourquoi les témoignages du surnaturel n'augmentent pas la crédence du bayésien dans le surnaturel ; ces témoignages sont tout aussi prévisibles en présence de surnaturel qu'en son absence.

L'évolution darwinienne des idéologies

Quelques jours après notre visite de Phnom Penh, on partit visiter les temples d'Angkor. Quel chef d'œuvre ! Ces temples sont impressionnantes tant par leur qualité que par leur quantité, et tant par leur ingéniosité que par le cadre sauvage environnant. Le plus grand de ces temples, Angkor Vat, est particulièrement monstrueux. Certains disent que 300 000 ouvriers et 6 000 éléphants ont construit ce monument gargantuesque en l'espace de 37 ans. Incroyable ! Comment est-ce possible ? Comment est-ce possible de coordonner le travail de 300 000 ouvriers pendant 37 ans ? D'autant qu'à l'époque, ils ne pouvaient pas utiliser Internet pour communiquer !

Mais c'est surtout un peu plus tard, à Ayutthaya en Thaïlande, que je connus le plus grand bouillonnement intellectuel du périple asiatique. Les yeux rivés sur des statues du Bouddha, je compris tout à coup que les croyances se propageaient à travers le temps et l'espace telles des espèces animales. Les croyances sont en compétition permanente pour gagner les esprits des hôtes humains. Et à ce jeu, les croyances qui ont perduré, au point d'être encore connues aujourd'hui, ne sont pas nécessairement les croyances les plus crédibles. Après tout, aucun d'entre nous n'est capable d'appliquer la formule de Bayes, même dans des cas simplistes ! Les croyances qui ont survécu à la grande compétition entre croyances sont celles qui ont séduit le plus grand nombre, et qui auront permis à ce plus grand nombre de survivre et de proliférer à son tour. Les croyances qui demeurent aujourd'hui sont celles qui ont réussi une symbiose avec les plus grandes civilisations humaines.

⁸  *Is the “hot hand” real?* Numberphile | L. Goldberg (2018)

Je fus vivement saisi par cette pensée. À défaut de comprendre pourquoi je pense ce que je pense, je commençais à comprendre pourquoi les civilisations pensent ce qu'elles pensent !

Le voyage en Asie se poursuivit dans les montagnes du Laos, les îles paradisiaques autour de Koh Lanta et la modernité d'un Kuala Lumpur musulman. Il toucha malheureusement ensuite à sa fin, et il me fallut retourner à Montréal pour débuter ma thèse. Je choisis alors de louer une chambre dans une colocation. L'un de mes colocataires allait devenir prêtre catholique. Appelons-le Bob. Rencontrer Bob fut une chance inouïe dans ma quête des origines des croyances.

Bob avait fait des études d'ingénieur, et s'intéressait à la logique des prédicats. Il était aussi bien sûr un fervent croyant catholique. Mais surtout, il adorait débattre dans le calme et de manière constructive. Et j'adorais débattre avec lui. J'ai beaucoup appris de lui. En particulier, ayant récemment découvert le réalisme modèle-dépendant de Hawking et Mlodinow, je n'avais alors aucun problème à accepter la « réalité » d'un Dieu, pourvu que l'on prenne le soin d'ajouter que ce Dieu existe dans un certain modèle. Au contact de Bob, je découvris de nombreux bienfaits des religions qui sont souvent omis par les anti-religieux — et Bob était lui aussi très attentif aux déviations des religions. On était bien loin des émissions américaines entre clans opposés qui sont constamment à deux doigts de s'insulter !

Les réflexions que j'eus à Ayutthaya, suivies de celles avec Bob, m'amènèrent à postuler par moi-même ce que les biologistes appellent la sélection de groupes. Ce processus permet d'expliquer l'omniprésence des croyances religieuses, notamment dans les siècles précédents. Le raisonnement risque de vous interpeler, voire de vous choquer. Notez bien qu'il s'agit d'une explication purement descriptive (ou prédictive), et que tout jugement moral que vous pourrez lire entre les lignes n'est qu'une maladresse de ma part.

Le raisonnement est le suivant. Les individus humains seuls n'ont aucune chance de survie. Même les petites tribus ont de bonnes chances de finir conquises par les plus grandes. Les humains qui survivent à travers les âges sont donc nécessairement ceux qui ont vécu dans des grandes civilisations. Or, vivre ensemble dans des grandes civilisations est une tâche ardue. Il faut absolument que ces grandes civilisations soient structurées pour ce faire. Il faut qu'elles possèdent une hiérarchie sociale. Mais il faut aussi, du coup, légitimer cette hiérarchie sociale. C'est là qu'entre la religion. La religion est, selon cet argument, l'outil indispensable à l'ordre social des grandes civilisations, laquelle est une condition *sine qua non* à la survie des humains. Autrement dit, au cours de l'histoire, il y a sans doute eu des humains sans religion. Mais ceux-ci n'ont pas fait long feu.

Vous pourriez vouloir alors rétorquer que depuis quelques siècles, les plus grandes civilisations adhèrent de moins en moins à la religion. Là encore, la sélection de groupe peut expliquer cela. En effet, l'avènement de l'économie de marché a permis la spécialisation, qui a permis, selon Adam Smith, d'aligner l'égoïsme

des individus avec les intérêts des sociétés⁹. Mieux encore, cette spécialisation gagne à ne pas être orchestrée par une autorité centrale, car cette autorité centrale connaîtrait moins bien les compétences des membres de la société que les membres de la société ne les connaissent eux-mêmes. Dès lors, suite à la révolution industrielle, les civilisations qui proliférèrent furent en fait celles capables de bouleverser l'ordre social établi¹⁰.

En particulier, si je pense ce que je pense, c'est aussi parce que je descends moi-même d'une lignée de croyants et de rebelles remettant en cause ces croyances.

Mais cela ne suffit pas à expliquer ce que je pense. L'autre explication est environnementale. En particulier, je fus frappé par le fait que Bob ne lisait pas du tout les mêmes sources d'information que moi. À cette époque, je fréquentais essentiellement les grands journaux nationaux français bien vus par l'élite, notamment le journal *Le Monde*. Mais pour la première fois de ma vie, grâce à Bob, je me mis à lire des sites web aux noms bibliques. J'avais découvert le problème des *filter bubbles* : nous lisons ce que nous acquiesçons¹¹ !

Ce phénomène est en fait aggravé par ce que les psychologues appellent la polarisation de groupe. La polarisation de groupe est un phénomène observé à la fois en laboratoires et dans les jurés des cours de justice. Le phénomène est grossièrement le suivant. Faites délibérer un groupe d'individus qui ont tendance à penser que X est bien. Après délibération, le groupe finira par penser que X est la solution à tous leurs problèmes. Pire encore, la délibération amènera chacun des membres du groupe à une conclusion plus extrême encore que celle qu'avait chacun des membres du groupe avant délibération¹² !

Il s'agit là sans doute d'un trait sélectionné par la sélection de groupe, étant donné qu'il permet de fédérer des individus et de les faire coopérer. Mais, de façon étonnante, il y a une explication plus bayésienne à notre désir de croire ce que le groupe auquel on appartient croit.

Croire les superstitions est utile

Certes, la *pure bayésienne* cherchera à déduire la croyance du groupe d'une théorie plus générale, qui expliquerait par exemple au passage les croyances d'autres groupes. Cependant, pour le *bayésien pragmatique*, un tel modèle serait compliqué et nécessiterait de longs calculs, surtout si le but n'est que de prédire, par exemple, ce que dira un croyant. Pour le *bayésien pragmatique*,

⁹  Adam Smith - division du travail & main invisible | Grain de Philo | Monsieur Phi (2017)

¹⁰  Infinitesimal: How a Dangerous Mathematical Theory Shaped the Modern World | Scientific American / Farrar, Straus and Giroux | A. Alexander (2015)

¹¹  Petit communautarisme deviendra grand | Démocratie 6 | Science4All | L.N. Hoang (2017)

¹²  Êtes-vous un hooligan politique ? Démocratie 10 | Science4All | L.N. Hoang (2017)

il sera plus *utile* de croire ce que les autres individus croient, car le modèle commun au groupe permet de rapidement prédire le comportement du groupe. En particulier, la conclusion étonnante de cette réflexion, c'est qu'un *bayésien pragmatique* d'une communauté déiste pourrait attacher une grande crédence à Dieu ! Bien sûr, en bon bayésien, il gardera en tête que tous les modèles sont faux, y compris ceux qui ont une grande crédence. Reste que croire en Dieu peut être un modèle *utile*.

L'introduction du pragmatisme exacerbe la subjectivité du bayésien. C'est tout à fait normal puisque les données auxquelles une personne est exposée diffèrent grandement de celles auxquelles une autre personne est exposée. Et de manière pragmatique, il nous faut nous adapter à notre environnement.

D'ailleurs, Dame Nature l'a bien compris. Elle a sélectionné les individus dont les traits sont adaptés à leur environnement, ou dont les traits sont capables de s'adapter à leur environnement. C'est ainsi que le traitement de l'information sonore par les bébés s'adapte rapidement au langage que ces bébés entendent. En particulier, les bébés apprennent à confondre des sons pourtant distincts, s'ils ne voient pas de gains prédictifs à les distinguer. Cet apprentissage permet aux bébés d'avoir un traitement du signal plus pragmatique ; même s'il a la fâcheuse conséquence qu'une fois adultes, nous sommes incapables de distinguer certains phonèmes distincts d'une langue étrangère.

Notre adaptation à notre environnement proche explique aussi notre étonnante compétence en ce que les *data scientists* appellent le *one-shot learning*. Cet apprentissage consiste à inférer beaucoup d'informations à partir d'une seule donnée, comme reconnaître un tufa dont on n'a vu qu'une photographie. En fait, au chapitre 19, on a vu que le bébé est capable de faire encore mieux, puisque dans certains contextes, il est même capable d'apprendre sans aucune donnée ! Pour réussir cette prouesse, il lui faut posséder un préjugé très complet et très structuré, pour que l'inférence bayésienne modifie grandement l'état de nos connaissances.

Le succès humain au *one-shot learning* révèle d'ailleurs un aspect souvent rejeté par les défenseurs de l'égalitarisme entre individus, à savoir le fait que nous naissions avec des cerveaux contenant déjà des préjugés. À bien y réfléchir, ceci n'a en fait rien d'étonnant. Notre cerveau, avec ses deux hémisphères, son hypothalamus et son cortex pré-frontal, a une structure très particulière. C'est la structure que la sélection naturelle a retenu. Du coup, contrairement à une idée reçue, il ne s'agit pas d'une ardoise vierge (*blank slate* en anglais) sur laquelle tout peut être écrit. Nous naissions avec des préjugés qui nous préparent à l'environnement qui va nous entourer, avec notamment une prédisposition à traiter les signaux que nos oreilles, yeux et autres nez vont nous envoyer.

Une variante de l'hypothèse inadéquate du *blank slate* est celle selon laquelle nous naissions avec des cerveaux identiques, et que nous sommes donc tous égaux à la naissance dans nos facultés d'apprentissage. Ce n'est pas le cas non plus. Des études sur des jumeaux suggèrent que nos gènes nous prédisposent

à certaines convictions politiques. En effet, des jumeaux identiques séparés à la naissance ont plus souvent le même avis politique que des frères et sœurs adoptés.

Cependant, le fait que nos réflexes pragmatiques et nos prédispositions génétiques nous préparent à un certain contexte signifie aussi que nos cerveaux ne nous préparent pas nécessairement à sortir de ce contexte. Nous sommes peut-être bons pour effectuer des prédictions concernant notre quotidien ; ceci ne signifie pas que les modèles qui ont gagné nos créances auront une quelconque utilité au-delà de ce quotidien. S'il y a bien une chose que ma courte carrière militaire, mon voyage en Asie et ma rencontre de Bob m'auront appris, c'est bien la faiblesse de mes modèles pour expliquer et comprendre des univers qui m'étaient étrangers. Je compris que la créance que j'avais gagnée en mes modèles avait un champ de crédibilité beaucoup plus restreint que ce que je croyais.

Je compris que j'avais vécu en excès de confiance. Et j'en avais compris plusieurs causes. J'avais subi l'influence biaisée de mon éducation et de mes pairs, hérité des gènes et de la culture de mes ancêtres, souffert d'innombrables biais cognitifs et vécu dans ma petite bulle dont les propriétés diffèrent grandement de ce qui se trouve au-delà de cette bulle.

La magie de YouTube

À ce moment, malgré toutes mes péripéties, je ne me rendais pas encore compte de l'ampleur de mon excès de confiance. Mais pour la première fois de ma vie, j'étais curieux de le découvrir. Par chance, c'est à ce moment-là que je découvris YouTube.

Pendant les années qui suivirent, je devins accro des pionniers d'une vulgarisation de très grande qualité sur le web, comme Singingbanana, VSauce, Veritasium, CGP Grey, e-penser, Dirty Biology et Science Étonnante. Grâce à eux et à d'autres, je découvris l'histoire du New Coke, l'expérience de conformité d'Asch, l'expérience d'obéissance de Milgram, l'expérience de la prison de Zimbardo, les effets placebo et nocebo ou encore les expériences sur le libre arbitre de Libet et Haynes. J'ouvris un blog à mon tour, et j'écrivis un article pour décrire ces expériences¹³. Et c'est en écrivant cet article que je me mis à vraiment sentir au plus profond de moi ce que ces expériences disent des hommes en général ; et de moi en particulier.

Sur YouTube, je me mis à *binge-watcher* les documentaires et les conférences grand public disponibles (parfois illégalement). YouTube a changé ma vie, et ma façon de voir le monde. Lors d'une conférence donnée le 7 février 2016 à Lyon Science, David Louapre prétendit que « YouTube est le truc le plus incroyable qui soit arrivé depuis l'invention de l'écriture ». Il explique que « l'évolution nous

¹³  *The Most Troubling Experiments on Human Behavior* | Science4All | L.N. Hoang (2014)

a rendu bons à la communication par langage oral ». D'ordinaire, toutefois, la communication orale entre deux personnes exige d'elles qu'elles soient présentes au même endroit au même moment. YouTube, et la vidéo en ligne de façon plus générale, a permis au langage oral de s'affranchir de sa délimitation spatio-temporelle, de sorte que tout individu aujourd'hui peut facilement parler à des millions d'autres individus dans le monde entier, et à tout moment dans le futur, à la demande de ces autres individus. Trois jours après la conférence de David Louapre, je lançai ma chaîne YouTube (francophone), Science4All. Le début d'une grande aventure.

Je ne cesse d'être impressionné par la qualité que l'on peut trouver sur YouTube. J'ai des dizaines et des dizaines de chaînes préférées, dont je ne manque aucun épisode. En vrac, je peux citer, côté francophone, Monsieur Phi, El Jj, Heu?reka, MicMaths, Scilabus, Hygiène Mentale, Micode, La statistique expliquée à mon chat, Risque Alpha, Passe-Science, Balade Mentale, Podcast Science et tant d'autres auprès desquels je m'excuse. Côté anglophone, citons 3Blue1Brown, PBS Infinite Series, PBS Space Time, Physics Girl, Kurzgesagt, Numberphile, Looking Glass Universe, Up and Atom, Minute Physics et Tipping Point Math, parmi tant d'autres.

Sur YouTube, certains noms revenaient avec insistance, ceux de Kahneman, Tversky, Haidt, Ariely, Shaw, Hawking, Mesquita, Harari, Pinker, Bostrom ou encore Kuhn. C'est aussi ça la merveille de la vulgarisation scientifique. Elle nous donne envie d'en savoir toujours plus. Intrigué, je me mis à lire tous ces grands chercheurs. Je découvris tout un univers qui s'était demandé pourquoi on pense ce que l'on pense, et comment penser mieux que ce que l'on pense. J'ai appris énormément. Mais surtout, je mesurais de mieux en mieux l'étendue de mon ignorance¹⁴.

Le périple continue

Mais toutes ces connaissances diverses et variées manquaient de structure. « Une accumulation de faits n'est pas plus une science qu'un tas de pierres n'est une maison », disait Poincaré. Je partis en quête d'une théorie des théories, quelque chose qui me permettrait d'unifier et de mieux comprendre toutes ces connaissances diverses et variées. Et pendant longtemps, je ne me rendis pas compte que la réponse résidait dans une formule que j'étudiais pourtant régulièrement, et dans un terme, le mot « bayésien », qui se trouve être l'un des premiers mots du titre de ma propre thèse de doctorat¹⁵. Début 2016, je me rendis compte petit à petit de l'importance de cette formule. Après avoir longtemps erré autour du pot, j'entamai enfin ma marche vers le bayésianisme.

¹⁴  *L'étendue de mon ignorance* | Démocratie 32 | Science4All | L.N. Hoang (2017)

¹⁵  *Conception bayésienne de mécanismes et quantification de l'équité appliquées à la construction d'horaires personnalisés* | Thèse de doctorat | L.N. Hoang (2014)

Ce ne fut pas une courte balade. Ce fut davantage un long périple dont j'ai l'impression, deux ans plus tard, de n'avoir effectué que les premiers pas. L'une des plus grosses difficultés, ce fut d'abord de m'extirper de la contrée « scientifique » dans laquelle j'avais baigné depuis si longtemps. Il me fallait me débarrasser des méthodes de *p-value*, de l'exigence de réfutabilité et de l'espoir d'objectivité. Il me fallait d'abord rejeter la « méthode scientifique », celle que, par une combinaison de raisonnement motivé, de dissonance cognitive et de polarisation de groupe, les scientifiques s'étaient presque tous mis à défendre. Il me fallait m'opposer à ceux que j'avais tant admirés.

Mais là n'est pas la plus grande difficulté. Le plus grand obstacle, qui se tient encore devant moi, reste de vraiment comprendre le bayésianisme, de calculer ses conséquences et d'acquérir la capacité à l'appliquer (approximativement). J'ai fait de mon mieux dans ce livre pour vous aider à ce faire. Mais mes propres limites cognitives sont béantes. Les exemples de Sally Clark, de Monty Hall et du mouton noir d'Écosse l'ont montré encore et encore. Je ne suis toujours pas capable d'appliquer la formule de Bayes. Ni même une version simpliste et très approximative de celle-ci.

Il me reste beaucoup de chemin à parcourir. Mais je comprends beaucoup mieux ce que je ne sais pas, et pourquoi je ne le sais pas. Je sais que mon incapacité à estimer la bonne réponse au problème de l'étudiant *troll* a pour conséquence mon incapacité à calculer de manière fiable la crédence adéquate à attribuer aux différents modèles que mon cerveau a en tête. Et je sais que mon cerveau est bien trop limité pour intégrer des modèles à trop grande complexité de Solomonoff ou à trop grande profondeur logique de Bennett.

Voilà qui me force à mieux reconnaître toute l'étendue de mon ignorance. Et, je l'espère, à réduire autant que possible mes futurs excès de confiance.

Références en français

- ➲ *Conception bayésienne de mécanismes et quantification de l'équité appliquées à la construction d'horaires personnalisés* | Thèse de doctorat | L.N. Hoang (2014)
- 🌐 $1+2+3+4+5+6+7+\dots = -1/12$! Science Étonnante | D. Louapre (2013)
- 📺 *Les coûts irrécupérables* | Crétin de cerveau | Science Étonnante | D. Louapre (2016)
- 📺 *Adam Smith - division du travail & main invisible* | Grain de Philo | Monsieur Phi (2017)
- 📺 $1+2+3+4+5+\dots = -1/12$??? Infini 5 | Science4All | L.N. Hoang (2016)
- 📺 *La supersommation linéaire, stable et régulière* | Hardcore | Science4All | L.N. Hoang (2016)
- 📺 *La quête mathématique de l'infiniment petit* | Infini 7 | Science4All | L.N. Hoang (2016)

- ▶ *Petit communautarisme deviendra grand* | Démocratie 6 | Science4All | L.N. Hoang (2017)
- ▶ *Êtes-vous un hooligan politique ?* Démocratie 10 | Science4All | L.N. Hoang (2017)
- ▶ *L'étendue de mon ignorance* | Démocratie 32 | Science4All | L.N. Hoang (2017)

Références en anglais

- ➲ *Predictions: How to See and Shape the Future with Game Theory* | Vintage | B. Mesquita (2010)
- ➲ *Thinking Fast and Slow* | SpringerFarrar, Straus and Giroux | D. Kahneman (2013)
- ➲ *Infinitesimal: How a Dangerous Mathematical Theory Shaped the Modern World* | Scientific American / Farrar, Straus and Giroux | A. Alexander (2015)
- ➲ *The law of group polarization* | Journal of political philosophy | C. Sunstein (2002)

- ➲ *The Euler-Maclaurin formula, Bernoulli numbers, the zeta function, and real-variable analytic continuation* | T. Tao (2010)
- ➲ *The Surprising Flavor of Infinite Series* | Science4All | L.N. Hoang (2013)
- ➲ *The Most Troubling Experiments on Human Behavior* | Science4All | L.N. Hoang (2014)
- ➲ *The Limitless Vertigo of Cantor's Infinite* | Science4All | L.N. Hoang (2015)

- ▶ *Why "scout mindset" is crucial to good judgment* | TEDxPSU | J. Galef (2016)
- ▶ *ASTOUNDING: $1 + 2 + 3 + 4 + 5 + \dots = -1/12$* | Numberphile | E. Copeland et T. Padilla (2014)
- ▶ *Ramanujan: Making sense of $1+2+3+\dots = -1/12$ and Co.* | Mathologer | B. Polster (2016)
- ▶ *Numberphile v. Math: the truth about $1+2+3+\dots=-1/12$* | Mathologer | B. Polster (2018)
- ▶ *Is the “hot hand” real?* Numberphile | L. Goldberg (2018)

Science sans conscience n'est que ruine de l'âme.

François Rabelais (1483 ou 1494-1553)

La compréhension des mathématiques est nécessaire pour une compréhension cohérente de l'éthique.

Socrate (-469- -399)

Résoudre le problème [de programmer la morale des IA] est un défi de recherche digne des plus grands talents mathématiques de la prochaine génération.

Nick Bostrom (1973-)

22

Au-delà du bayésianisme

Le bayésien n'a pas de morale

« Dieu est mort », affirma Friedrich Nietzsche. Nietzsche désapprouvait le christianisme. Néanmoins, sa sentence de Dieu n'avait rien d'une célébration. Il ne s'agissait pas d'un triomphalisme de l'athéisme. Pour Nietzsche, la mort de Dieu était avant tout source d'inquiétude, car Nietzsche voyait bien tous les bienfaits de la croyance en Dieu pour une société.

Pour le *bayésien*, on l'a vu, *tout est fiction*. Le corollaire immédiat est que tout principe moral n'est aussi que fiction. Après tout, vu la difficulté que l'on a eue à définir la vie, peut-on vraiment donner un sens précis au commandement « tu ne tueras point » ? Cette phrase inclut-elle les milliards de bactéries qui habitent nos corps ? Et si elle se restreint à l'homme, est-on sûr de pouvoir distinguer un homme de celui qui n'en est pas ? À partir de quand un fœtus est-il humain ? Et s'il fallait tuer un homme pour en sauver mille autres ? Aurait-il fallu tuer Hitler avant que celui-ci n'accède au pouvoir ?

Si le *bayésien* rejette l'existence d'une morale fondamentale, il n'a toutefois pas son mot à dire sur la *bonne* morale à suivre. Il s'agirait là d'une philosophie morale prescriptive, c'est-à-dire d'une philosophie qui parle de ce qui *devrait* être. Or, le bayésianisme n'est pas une philosophie morale prescriptive. Être bayésien, c'est utiliser une certaine approche pour organiser son savoir. Mais ce n'est *pas* donner des leçons morales sur ce qui est bien et ce qui est mal, sur ce qu'il faudrait faire et sur ce qui devrait être interdit. La *pure bayésienne* et le

bayésien pragmatique n'ont pas de philosophie morale. Ils ne considèrent même pas qu'être bayésien est *bien* ou *souhaitable*. Et ils ne chercheront absolument pas à vous convaincre que vous *devriez* davantage vous reposer sur la formule de Bayes ! Un bayésien n'est qu'une machine à appliquer (des approximations de) la formule de Bayes.

Cela ne veut pas dire que la morale ne fait pas partie du langage du bayésien pour autant. Pour expliquer pourquoi les humains se comportent comme ils se comportent, laissent des pourboires au restaurant, tirent la chasse d'eau en sortant des toilettes et redistribuent les richesses produites par leurs sociétés, le bayésien trouvera *utile* de supposer que chaque humain possède sa propre morale, et que les individus des mêmes groupes sociaux ont des morales souvent similaires.

La morale (sélectionnée par la sélection) naturelle

De prime abord, il y a de quoi s'étonner de l'existence même d'une morale chez l'homme. Chacun d'entre nous n'y gagnerait-il pas individuellement à nier tout sens moral pour profiter de l'altruisme des autres sans donner en retour ? Cette question devient particulièrement intrigante quand, à cela, on ajoute le fait que la sélection naturelle tend à privilégier ceux dont le comportement permet de se reproduire en plus grand nombre. *A priori*, il semble donc que cette sélection naturelle devrait favoriser les égoïstes immoraux au profit de l'altruisme des sages à l'éthique irréprochable.

Il y a cependant plusieurs façons d'expliquer l'émergence et la survie des comportements altruistes dans la nature. Pendant longtemps, l'hypothèse de la sélection de parentèle a régné sans partage. Cette hypothèse fait des gènes l'objet sélectionné par la sélection naturelle, et des individus des outils que les gènes utilisent pour se reproduire au mieux. En particulier, les gènes poussent leurs individus à maximiser le nombre de descendants dont les gènes seront similaires. Ainsi, il peut être bénéfique pour une abeille de se sacrifier pour que sa reine ponde de nombreux descendants, puisque les descendants de la reine auront des gènes nécessairement similaires à ceux de l'abeille en question.

Cependant, cette hypothèse semble avoir des limites, notamment quand on cherche à l'appliquer à l'homme. D'autres hypothèses ont ainsi été proposées. L'une d'elles fait de la notion de choix de partenaires la clé de la morale. En particulier, notamment à l'époque des chasseurs-cueilleurs, un individu immoral aurait sans doute rapidement été exclu des circuits de coopération, ce qui le force à vivre seul dans la nature et ne lui laisserait aucune chance de survie. Des simulations simplistes suggèrent que cette hypothèse pourrait suffire à expliquer l'altruisme¹.

¹  *Le paradoxe de la morale* | Démocratie 25 | Science4All | S. Debove et L.N. Hoang (2017)

Une troisième hypothèse avancée est celle de la sélection de groupe. Là encore, la clé de la morale est sa capacité à faire coopérer différents individus. En particulier, un groupe a plus de chance de survie s'il croît en population, et si, malgré cela, les individus de la population continuent à coopérer. La sélection de groupe suppose que la plupart des groupes ont échoué ce faisant, faute de principes moraux suffisamment adéquats pour la vie en grande société. Les groupes qui ont survécu sont nécessairement ceux dont les individus avaient une morale très développée qui permet au groupe de primer sur l'individu.

La sélection de groupe prédit une facette importante des morales que l'on retrouve à travers le monde : l'exclusion des individus différents et des traîtres. En effet, pour qu'un groupe survive, il faut non seulement que ses individus aient des principes moraux forts et adéquats, mais il faut aussi que le groupe résiste à l'infiltration d'individus égoïstes. Et pour ce faire, le groupe doit avoir une manière de détecter ces individus égoïstes et de les exclure, à l'instar de notre système immunitaire qui combat les cellules cancéreuses. À l'inverse, le groupe doit célébrer ses membres, en les unissant via des symboles qu'ils sacralisent, comme une langue, un drapeau ou un hymne. La sélection de groupe prédit ainsi notre hooliganisme pour les groupes auxquels on s'identifie². Voici, là encore, un trait que l'on retrouve dans de nombreuses sociétés.

D'ailleurs, Nietzsche suggère que l'hooliganisme aristocrate et l'hooliganisme populaire ont conduit leurs tenants à deux antonymes différents du « bon ». Pour les uns, ceux qui ne sont pas bons sont *mauvais*³. Pour les autres, ceux qui ne sont pas bons sont *méchants*⁴. Comme l'explique le philosophe Thibaut Giraud : « Tandis que le *mauvais*, c'est celui qui voudrait être bon mais ne *peut* pas, le *méchant*, c'est à l'inverse celui qui pourrait être bon mais ne le *veut* pas. »

Mais il semble que cet exemple de morales issues de notre identification à un groupe n'est qu'un exemple parmi tant d'autres. On peut ainsi identifier bien d'autres hooliganismes, qu'ils soient libertariens, égalitaires, traditionalistes, progressistes, nationalistes ou mondialistes. Dans tous ces cas, il semble que les individus de ces groupes s'identifient d'abord à leur groupe, puis cèdent à l'irrationalité pour défendre la cause de leurs groupes⁵ et cherchent ensuite à rationaliser leurs prises de positions⁶.

J'irai même jusqu'à reprocher aux scientifiques et aux amateurs des sciences d'avoir développé un hooliganisme scientifique, qui tend à défendre irrationnellement la légitimité des sciences. Cet hooliganisme scientifique explique ainsi

² *La morale des hooligans (la nôtre !!)* | Démocratie 27 | Science4All | L.N. Hoang (2017)

³ *Nietzsche - La morale des winners ! Généalogie de la morale (1/2)* | Grain de Philo | Monsieur Phi | T. Giraud (2017)

⁴ *Nietzsche et les méchants - Généalogie de la morale (2/2)* | Grain de Philo | Monsieur Phi | T. Giraud (2017)

⁵ *Êtes-vous un hooligan politique ?* | Démocratie 10 | Science4All | L.N. Hoang (2017)

⁶ *La rationalisation* | La Tronche en Biais | V. Tapas et T. Durand (2015)

l'exigence d'une objectivité (pourtant illusoire) et l'acceptation de la « méthode scientifique » — alors que la *pure bayésienne* et le *bayésien pragmatique* affirmeraient qu'il ne s'agit ni d'une bonne théorie descriptive du fonctionnement des sciences, ni d'une théorie normative adéquate⁷.

J'espère qu'en bon bayésien, vous ne cherchez pas à déterminer laquelle de ces trois hypothèses expliquant l'origine de la morale intuitive est *vraie*. « Tous les modèles sont faux. » Et différents modèles peuvent être utiles dans différentes circonstances. L'hypothèse de la parentèle est utile pour comprendre l'importance de la famille, l'hypothèse du choix de partenaires est utile pour comprendre notre addiction aux ragots et l'hypothèse du groupe est utile pour comprendre la polarisation de groupe. Dans tous ces cas, ce qui est remarquable, c'est que l'on parvient à déduire les morales des sociétés humaines de principes évolutionnistes. La morale que les individus ont aujourd'hui n'a alors rien de mystérieux ou de mystique ; elle ne semble donc pas mériter de statut plus fondamental que les individus qui la possèdent.

Mais ça, vous le saviez. « Tous les modèles sont faux. » Y compris les philosophies morales.

Inconscients de nos morales

En fait, les expériences de psychologie montrent encore et encore que nos morales intuitives ont un très grand nombre de défauts, à commencer par le fait que nous connaissons très mal nos préférences et les causes de nos préférences. Des expériences montrent ainsi que nous jugeons tellement le vin en fonction de son prix d'achat, que nos papilles gustatives parviennent à distinguer des vins identiques vendus à des prix différents.

Le cas le plus spectaculaire est sans doute l'histoire du New Coke. Dans les années 1980, les États-Unis sont divisés entre deux marques de soda, Coca-Cola et Pepsi. En aveugle, des expériences avaient montré que le Pepsi était préféré au Coca-Cola. Coca-Cola réagit alors en modifiant sa recette et commercialisa ce qu'il appela le New Coke. En aveugle, ce New Coke était préféré au Pepsi et à la recette originale du Coca-Cola.

Cependant, la vraie dégustation du New Coke ne se fait pas en aveugle ! Quelle qu'en soit la raison psychologique, les Américains se soulevèrent alors contre l'innovation de Coca-Cola, et exigèrent le retour à l'ancienne recette. Coca-Cola finit par se conformer à la demande populaire. Le New Coke fut supprimé, et l'ancienne recette fut à nouveau commercialisée. Plus étrange encore, les ventes de Coca-Cola explosèrent et Coca-Cola acquit une notoriété jamais égalée jusque-là⁸.

⁷  *L'hooliganisme politique a gâché ma marche pour les sciences* | My4Cents (Genève) | L.N. Hoang (2017)

⁸  *New Coke - A Complete Disaster? Company Man* (2017)

Derrière le succès de Coca-Cola se cache en fait toute l'industrie de la publicité. Cette industrie exploite, peut-être sans le savoir, un biais psychologique connu sous le nom de l'effet de simple exposition. Cet effet est merveilleusement illustré par une expérience conduite dans deux universités de l'État de Washington. Des chercheurs publièrent dans les journaux des universités des publicités avec des mots inventés comme Kardiga, Saricik ou Nansoma. De façon cruciale, ces mots apparaissaient bien plus souvent dans une université que dans l'autre. Puis les chercheurs demandèrent aux étudiants de noter ces mots sur une échelle de mal à bien. Et les résultats furent univoques : les mots plus fréquents étaient mieux connotés par les étudiants. Nous aimons ce qui nous est familier.

Selon le psychologue Kahneman, cette expérience s'explique par un phénomène qu'il appelle l'aisance cognitive. L'idée est que le cerveau a une aversion pour la réflexion. Par conséquent, il apprécie ce qui lui vient aisément à l'esprit. C'est ainsi que les avocats dont les noms sont faciles à prononcer sont sur-représentés, et que les compagnies dont les abréviations à la bourse sont prononçables ont de meilleures performances que les autres. Cependant, ce biais pour l'aisance cognitive peut se faire au détriment de l'effort intellectuel nécessaire pour résoudre certains problèmes. Une autre expérience étonnante montre que, face à un examen piégeux, les étudiants s'en sortent nettement mieux si la police de caractères utilisée pour écrire l'examen est *moins* lisible. Selon Kahneman, lorsque la police est trop facile à lire, les étudiants se laissent emporter par l'aisance cognitive, et ne prennent pas suffisamment le temps de la réflexion⁹.

Si ces petits biais peuvent nous nuire individuellement, ils ne semblent pas correspondre à des questions morales. Détrompez-vous. Un grand nombre d'études montrent que la manière dont on vote est grandement influencée par les cent premières millisecondes pendant lesquelles on découvre le visage des candidats¹⁰. Cent millisecondes, c'est un clignement d'œil ! En particulier, notre jugement de la compétence du candidat en cent millisecondes à partir uniquement de son visage suffit à déterminer une grande partie de ce que l'on pense du candidat, de son caractère et de sa capacité à diriger un pays.

Bien entendu, ce n'est jamais l'explication que l'on avance au moment où l'on nous demande de justifier nos convictions politiques ! Nous ne connaissons quasiment jamais les vraies causes de nos croyances. Cependant, nous cherchons constamment à les justifier. Comme le dit Jonathan Haidt, « l'intuition d'abord, la raison ensuite ». Selon Haidt, notre raison passe son temps à lister des arguments ad hoc pour justifier la position que notre intuition a déjà choisie ; et à rejeter tous les arguments qui pourraient remettre en question cette position.

L'un des cas les plus spectaculaires de la manière dont notre intuition détermine le but de notre raison est l'expérience menée par Dan Kahan. En 2013, Kahan proposa à ses étudiants un exercice classique de mathématiques où, à partir de données numériques, une règle de trois permettait de conclure quant

⁹  *The Illusion of Truth* | Veritasium | D. Muller (2016)

¹⁰  *Choisir son président en 100 millisecondes* | Homo Fabulus | S. Debove (2017)

à l'efficacité d'une crème. Le succès à cet exercice ne fut pas excellent ; mais il fut raisonnable. Kahan modifia ensuite uniquement l'emballage de l'exercice. Il fallait désormais conclure quant à l'efficacité d'une loi sur le port d'armes. Mais les chiffres étaient les mêmes. La méthode de résolution de l'exercice était donc inchangée ; et sa difficulté mathématique aussi. Néanmoins, les résultats devinrent catastrophiques. Pire encore, quelle que soit la conclusion attendue par l'exercice, les participants en vinrent constamment à une conclusion qui allait dans le sens de leurs convictions. Leurs intuitions avaient pris parti ; leur raison devait s'y conformer¹¹.

Si l'exemple de Kahan est un cas manifeste, le choix des mots dans les discours militants affecte nécessairement l'intuition des gens qui prononcent ou qui écoutent ces discours, pour faire basculer notre intuition et ses valeurs morales vers un clan plutôt qu'un autre. En particulier, subtilement et sans que l'on s'en rende compte, la connotation des mots va souvent grandement affecter la position de notre intuition ; et donc les arguments que l'on va supporter ou défendre. De façon étonnante, nous avons énormément de mal à faire le lien entre des mots dont les connotations sont opposées, mais dont la signification est pourtant la même.

C'est quelque chose dont je me suis rendu compte il y a quelques années, et qui m'a amené à vouloir chercher les synonymes à connotations opposées de mots fortement connotés. Bien souvent, il est difficile de trouver un synonyme parfait ; mais pour déterminer les causes du parti pris par mes intuitions, même les synonymes imparfaits m'ont été grandement utiles, pourvu qu'ils soient à connotations opposées. Je vous invite ainsi à réfléchir aux paires de mots qui suivent : démocratie et populisme, terrorisme et résistance, communauté et secte, tyran et leader, naturel et sauvage, conditionnement et éducation, PIB et flux de dettes, hypocrite et diplomate, prudent et paranoïaque, préjugé et *a priori* bayésien... Je vous invite notamment à compléter et utiliser cette liste à chaque fois que vous serez exposé à des discours militants¹² — surtout s'il s'agit de vos discours.

Mieux encore, à l'instar de mon utilisation du mot « préjugé » dans ce livre, je vous invite à utiliser la connotation défavorable à vos discours quand vous défendez vos positions. Certes, il vous sera plus difficile de persuader autrui ce faisant. Du coup, si vous pensez que le but d'un débat est de gagner le débat ou de gagner du prestige, il s'agit là clairement d'une mauvaise stratégie. Mais si votre objectif est davantage la clarification des idées (y compris les vôtres !) et le calcul bayésien des crédences en diverses théories, alors cette utilisation de la connotation opposée à votre position sera d'une très grande utilité. Cela évitera, entre autres, de convaincre autrui (ou vous-mêmes !) pour des mauvaises raisons. « Le premier principe est que vous ne devez pas vous duper — et vous êtes la personne la plus facile à duper », disait Feynman. Ainsi, si vous défendez

¹¹  Politics and Numbers | Numberphile | J. Grime (2013)

¹²  Les synonymes à connotations opposées | My4Cents (Mayen) | Science4All | L.N. Hoang (2016)

la consommation de produits *naturels*, cherchez davantage à montrer en quoi les produits *sauvages* sont meilleurs.

Il n'y a malheureusement pas que la connotation des mots qui décide de nos convictions morales. Comme on en a déjà parlé au chapitre 17, notre intuition est guidée par tout notre environnement immédiat. C'est l'*effet d'amorçage*. Notre morale dépend de stimuli dont nous sommes souvent inconscients, à l'image de l'expérience de Wells et Petty qui montre qu'un simple mouvement de tête guide inconsciemment notre jugement moral de ce que devraient être les frais de scolarité. De même, des études montrent que le simple fait de placer les bureaux de vote dans des écoles modifiait significativement l'importance que les électeurs attachaient à l'éducation.

Des bâtons et des carottes

Notre inconscience de nos morales intuitives n'est toutefois qu'une partie du problème ! Un autre corollaire de l'explication évolutionniste de la morale est le fait que notre morale intuitive est adaptée à une époque ancienne. Pire encore, elle n'est *bonne* que dans l'optique de la survie de nos gènes. On peut donc largement douter de sa pertinence dans nos sociétés modernes. D'ailleurs, les morales ont grandement évolué au cours des derniers siècles, surtout suite à la Révolution industrielle. Le populisme a renversé le royalisme, l'homophobie a souvent été remplacée par une stigmatisation de l'homophobie et l'égalité entre les sexes est devenue une priorité sociétale — j'essaie là tant bien que mal d'être descriptif et toute impression de jugement moral n'est que maladresse de ma part.

Si la *pure bayésienne* et le *bayésien pragmatique* ont leurs prédictions (sans doute pleines d'incertitudes) quant à la morale des sociétés du futur, je crains que mes capacités cognitives soient malheureusement beaucoup trop limitées pour effectuer une prédition crédible sur la nature des morales de nos descendants. Le bayésien en moi serait néanmoins prêt à parier que nos morales continueront à être bouleversées dans les décennies à venir, sans doute à un rythme inégalé dans l'histoire de l'humanité, à tel point que nos descendants considéreront que nos morales d'aujourd'hui sont complètement arriérées, irrationnelles, voire... immorales.

Une raison pour cela est la manière dont nous acquérons notre morale. Certes, celle-ci est en partie déterminée par notre patrimoine génétique. Mais une grande partie de notre morale est aussi acquise, à l'école ou avec les parents, à l'aide de carottes et de bâtons. Ces carottes et ces bâtons sont d'ailleurs ce qu'utilisent les chercheurs en *machine learning* pour façonner les fonctions objectifs de leurs intelligences artificielles. À force de se faire taper sur les doigts pour avoir mal conjugué le verbe « avoir », les jeunes enfants et les machines finissent par apprendre la « bonne » conjugaison — et à corriger, parfois violemment,

toute personne qui serait en train de se tromper. On parle d'apprentissage par renforcement, ou *reinforcement learning*.

C'est ce type d'apprentissage qui a permis à Google DeepMind de résoudre un grand nombre de jeux d'arcade, via une intelligence artificielle qui n'était pourtant capable que de voir les couleurs des pixels de l'écran et ne cherchait qu'à atteindre le score maximum. Ce score servait alors de carotte et de bâton. De façon étonnante, il aura suffi. En s'appuyant uniquement sur ce score, Google DeepMind aura réussi des performances surhumaines¹³.

Cependant, les choix des carottes et des bâtons peuvent avoir des conséquences inattendues. En 2016, Microsoft lança Tay, une intelligence artificielle aux commandes du compte tweeter @TayTweets. Le problème, c'est que Tay apprit par apprentissage par renforcement, en étudiant notamment les réactions à ses tweets. En moins de 24 heures, les *trolls* de Twitter ont transformé Tay en monstre nazi raciste rejettant l'existence de l'holocauste et supportant l'idée d'un nouveau génocide. Il va sans dire que Tay fut rapidement éteinte par Microsoft.

Cependant, Tay en dit long sur la manière dont nos propres valeurs morales se forment. Notre morale se construit en partie par apprentissage par renforcement, ce qui explique pourquoi les morales des individus géographiquement ou socialement proches sont souvent assez similaires. Si nous pensons ce que nous pensons, c'est en grande partie parce que notre environnement, social et culturel, nous a poussés à penser ce que nous pensons. Y compris quand il s'agit de questions morales.

La morale du plus grand nombre ?

Un sophisme récurrent assigne à la démocratie un objectif bien défini, lequel émerge des volontés individuelles des citoyens. *Il faut suivre la volonté du peuple*, entend-on souvent. Cependant, une société n'est pas un individu avec une volonté unique. Quand bien même chaque membre d'une société aurait un ensemble de préférences cohérentes, le fameux paradoxe de Condorcet et le célèbre théorème d'impossibilité d'Arrow montrent qu'il n'y aucune façon naturelle de déduire de préférences individuelles (ordinaires) une préférence du groupe. À moins de solutions insatisfaisantes comme la dictature¹⁴.

Pendant des millénaires, les décisions collégiales des groupes humains ont ainsi été une décision essentiellement dictatoriale d'un petit nombre, ou un consensus

¹³  *Human-level control through deep reinforcement learning* | Nature | V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg et D. Hassabis (2015)

¹⁴  *3 théorèmes anti-démocratiques (et la lotocratie)* | Démocratie 3 / Science4All / L.N. Hoang (2017)

qui mettait du temps à se dessiner et qui était remporté par les plus charismatiques, au dépens des plus introvertis. Plus tard, l'invention du vote a permis de répartir équitablement les voix des citoyens. Toutefois, nos scrutins d'aujourd'hui ont de très mauvaises propriétés mathématiques qui favorisent le bipartisme, l'hooliganisme politique et le vote utile. Des scrutins alternatifs avec de bien meilleures propriétés mathématiques ont été proposés au cours des dernières décennies. On peut citer le jugement majoritaire et le scrutin de Condorcet randomisé¹⁵.

Mais il serait erroné de penser que le choix d'un scrutin démocratique suffira à résoudre notre problème de morale prescriptive. Comme on l'a vu, nos morales intuitives sont très déficientes et perfectibles. Pire encore, selon l'économiste Bryan Caplan, les électeurs sont pires qu'ignorants au moment du vote ; ils sont irrationnels. Ils votent par hooliganisme politique, refusent de faire le nécessaire pour effectuer un vote informé et ne culpabilisent aucunement. Et à bien y réfléchir, ceci n'a rien d'étonnant. S'informer a un coût, mais voter informé n'a essentiellement aucun gain. Pour Caplan, les électeurs sont rationnellement irrationnels¹⁶.

Bref. Il semblerait que l'objectif moral de nos sociétés ne peut pas se déduire des morales intuitives des citoyens de nos sociétés, et encore moins de ce que ces citoyens pourraient *dire* désirer. Notre morale intuitive collective semble lunatique, injustifiée, manipulable, incohérente et inadaptée à la vie moderne. Selon Daniel Kahneman, « il est utile de voir la cohérence logique des préférences humaines pour ce qu'elle est : un mirage sans espoir ». Il ne s'agit pas là d'un jugement moral ; Kahneman ne cherche là qu'à décrire notre morale naturelle.

En particulier, l'expérience d'Allais montre que nos préférences violent les axiomes de von Neumann - Morgenstern¹⁷. Il en très certainement de même de nos morales. Concrètement, ceci signifie qu'il nous arrive de penser qu'une situation *A* soit moralement préférable à *B*, *B* à *C* et *C* à *A*. Mais alors, l'argument du « pari hollandais » (ou *Dutch book*) montre qu'un *bookmaker* pourrait vous amener à payer de petites sommes pour passer de *A* à *B*, de *B* à *C* et de *C* à *A*. *In fine*, vous aurez alors dépensé de l'argent sans que la situation ait changé. Vous aurez uniquement gaspillé du temps, de l'énergie et de l'argent.

Exprimé de manière abstraite, ce paradoxe d'Allais peut sembler idiot. Cependant, dès lors qu'il s'agit de sujets sensibles qui touchent à la politique, à la discrimination ou à nos valeurs, ou dès que l'incertitude entre en jeu, il nous est très impossible de nous rendre de compte de ce type d'incohérence. La situation est encore pire à l'échelle de la société, notamment sachant le théorème d'impossibilité d'Arrow. Dans tous ces cas, les incohérences de nos morales, individuelles et collectives, conduisent à un énorme gaspillage de temps, d'énergie

¹⁵  *Le scrutin de Condorcet randomisé (mon préféré !!)* | Démocratie 5 | Science4All | L.N. Hoang (2017)

¹⁶  *Rationnellement irrationnels* | Démocratie 11 | Science4All | L.N. Hoang (2017)

¹⁷  *Argent, risques et paradoxes* | Démocratie 12 | Science4All | L.N. Hoang (2017)

et d'argent. Voilà qui devrait nous inciter à davantage clarifier et formaliser nos morales¹⁸.

Est-ce à dire que notre morale intuitive n'est pas souhaitable, et qu'il convient de la remplacer autant que possible par une morale plus appropriée ? Pour la *pure bayésienne*, non. Souvenez-vous. La *pure bayésienne* n'a pas son mot à dire sur la morale prescriptive.

Mais comme je sens que vous mourez d'envie de parler de morale, je vous invite à dépasser les frontières du royaume de la connaissance pour explorer (très) brièvement celui de la morale prescriptive. Et l'on va voir que pour de nombreuses morales prescriptives, si le bayésianisme n'est pas le fondement de ces morales, il n'en demeure pas moins un outil indispensable — tout comme l'ensemble des mathématiques d'ailleurs !

La morale déontologique

Les deux principales approches à la morale prescriptive sont la déontologie et le conséquentialisme. La déontologie consiste à prescrire des droits et des devoirs. Le conséquentialisme ignore les moyens de nos actions morales, et n'en juge que les conséquences (ou ce que l'on croyait que ces conséquences seraient au moment de nos actions).

L'un des grands défenseurs de la morale déontologique est le philosophe Emmanuel Kant. Kant distingue deux types de devoirs moraux, qu'il appelle impératifs. D'un côté, l'impératif hypothétique est un devoir circonstanciel. Il s'agit d'une action à mener pour une fin prédeterminée. De l'autre côté, l'impératif catégorique est un devoir moral absolu, indépendant des circonstances.

Pour Kant, la propriété fondamentale des impératifs catégoriques est leur universalité. « Agis seulement d'après la maxime grâce à laquelle tu peux vouloir en même temps qu'elle devienne une loi universelle », a-t-il écrit. Ainsi, pour Kant, tout devoir moral est tel qu'il est souhaitable que tout le monde y obéisse¹⁹.

C'est essentiellement cette morale déontologique qui apparaît dans les directives religieuses ou les textes de loi et les statuts des associations. « Tu ne tueras pas. » « Nul n'est censé ignorer la loi. » Ce n'est sans doute pas un hasard. Il est plus facile pour un juge de vérifier si le droit a été respecté, et d'en déduire une sentence. Qui plus est, l'approche déontologique est utile pour l'uniformisation des décisions judiciaires, que beaucoup reconnaissent comme un impératif catégorique. « Les hommes naissent libres et égaux en droits »,

¹⁸En particulier, le théorème de von Neumann - Morgenstern montre que toutes les préférences cohérentes sont équivalentes à maximiser l'espérance d'un score. On décrira cela plus en détail à la fin de ce chapitre.

¹⁹  *À chacun sa morale ? Relativisme vs. réalisme* | Grain de philo | Monsieur Phi | T. Giraud (2017)

affirme le premier article de la Déclaration des Droits de l'homme et du citoyen de 1789.

Cependant, la morale déontologique connaît également son lot de critiques. En particulier, résumer la morale en une poignée d'impératifs catégoriques, à l'instar de définir la vie en quelques phrases, semble être une entreprise bien complexe. Voire illusoire. Voici trois arguments contre l'approche déontologique de la morale.

Pour commencer, quels que soient les principes déontologiques, il semble qu'il demeurera toujours des exceptions. Le cas le plus célèbre est celui du mensonge selon Kant. Imaginez-vous à table avec vos deux enfants, Claude et Dominique. Un homme armé d'un pistolet affirme qu'il veut tuer vos enfants pour avoir joué trop bruyamment dans le parc quelques heures plus tôt. Claude et Dominique ayant senti le tueur venir sont partis se cacher dans la cave. Le tueur vous demande si vous savez où se cachent vos enfants. Malgré ces circonstances, Kant affirme que vous avez le devoir moral de dire la vérité.

De façon plus générale, il est difficile d'anticiper le fait qu'un principe déontologique sera *toujours* souhaitable (disons, selon la morale intuitive). À bien y réfléchir, même le commandement « tu ne tueras point » semble avoir des exceptions — les expériences de pensée qui le montrent font généralement intervenir le personnage d'Adolf Hitler ! Pour garantir qu'un principe moral soit *toujours* valide, il semble nécessaire d'anticiper tout ce qui peut se passer — ou du moins tout ce qui a une chance non-infinitésimale de se passer. Mais on peut alors douter du fait que les limites cognitives humaines soient suffisantes pour vraiment garantir qu'un principe moral soit *toujours* valide. De façon générale, la déontologie semble souffrir d'un excès de rigidité²⁰.

Une seconde limite de l'approche déontologique est le fait qu'elle soit vouée à être mal définie. On l'a vu, pour prétendre que « tu ne tueras point » est un devoir déontologique, il nous faut d'abord définir la vie, la mort, le meurtre, l'intention, le libre arbitre et tout plein d'autres concepts. Or tous ces concepts sont en fait nécessairement mal définis, en tout cas si l'on en croit le bayésianisme. Ou plutôt, comme on en a parlé au chapitre 20, la réalité d'un concept est dépendante du modèle considéré. Or, aucun d'entre nous n'a exactement le même modèle de la réalité. Qui plus est, toute formalisation rigoureuse de ces concepts, à l'instar de celui de chat, nécessitera sans doute des gigaoctets de bits d'information, qu'aucun d'entre nous n'aura le temps ou la patience de lire et comprendre²¹.

Pire encore, l'approche déontologique risque fort d'inciter les individus à tordre les définitions des mots présents dans les principes déontologiques. En effet, il est plus facile de chercher à être moral en jouant avec les interprétations des

²⁰  *Fat Tony et Dr John (biais-variance)* | IA 12 | Science4All | G. Mitteau et L.N. Hoang (2018)

²¹  *Why Asimov's Laws of Robotics Don't Work* | Computerphile | R. Miles (2015)

principes déontologiques qu'en changeant de comportement. Le plus problématique, c'est que cette stratégie est généralement très inconsciente. « L'intuition d'abord, la raison ensuite », comme le dit Jonathan Haidt. Sans que l'on s'en rende compte, on est ainsi amené à bidouiller l'interprétation des principes déontologiques pour rester dans notre droit moral sans avoir à modifier nos comportements. Voilà qui explique pourquoi tant de débats sur fond déontologique ne sont qu'une triste guerre de définitions sémantiques sans fin — surtout quand les parties prenantes cherchent à défendre la supériorité morale de *leurs* actions.

Enfin, une troisième limite de l'approche déontologique est son manque de discernement entre des options plus ou moins bonnes. Bien souvent, cette approche liste des actions à faire, ou des actions à ne pas faire. Mais alors, que faire si l'on est contraint de choisir entre des actions à faire ? Ou s'il faut choisir entre la peste et le choléra ?

Dans ses livres de science-fiction, Isaac Asimov propose une formulation déontologique de la morale des robots. Cette formulation consiste à lister l'ordre de priorité entre plusieurs actions. Ainsi, un robot doit avant tout protéger l'intégrité physique des humains. Ce n'est que sous réserve de cette première loi qu'un robot devra ensuite suivre les ordres d'un humain. Cependant, il est très illusoire d'espérer dresser une telle liste de priorité pour tous les problèmes de prises de décisions morales. En effet, la meilleure action à faire dépend non seulement de l'ensemble des actions envisageables, mais aussi du contexte de la prise de décision. Or, le nombre de contextes imaginables est exponentiellement gigantesque. Lister toutes les actions à faire dans tous les cas revient à écrire un algorithme qui dit tout ce qu'il faut faire dans tout contexte. Or, cet algorithme aura très certainement une énorme complexité de Solomonoff. Écrire, lire et appliquer cet algorithme est complètement illusoire. Par des humains, comme par des machines.

A contrario, le conséquentialiste va rejeter tout impératif catégorique. Dès lors, tout ce qui importe est la conséquence. Une action sera moralement bonne, si et seulement si, ses conséquences sont désirables. La fin, pourvue qu'elle soit vraiment désirable, justifie les moyens.

Le savoir est-il une fin raisonnable ?

Reste à déterminer quelle fin est souhaitable. Qu'attendons-nous vraiment de la société ? Quel est le but final de nos civilisations ? Quelle est notre fonction objectif ? Telle est la question fondamentale du conséquentialiste.

Une idée souvent défendue par les scientifiques est le désir de savoir. Serait-il moralement justifiable de faire du savoir l'objectif de nos sociétés ? Est-il vraiment raisonnable d'exiger la connaissance chez le plus grand nombre ? Faut-il forcer tout le monde à prendre une pilule rouge ? Sans doute pas. En fait,

désigner le savoir comme étant l'objectif des sciences a même de nombreuses conséquences étranges.

Le problème est que le monde est gigantesque et complexe. La quantité de chose à savoir du monde transcende largement nos capacités intellectuelles. Ernest Rutherford va même jusqu'à prétendre que « toute science est soit de la physique, soit de la collection de timbres ». Et comme le dirait Poincaré, « une accumulation de faits n'est pas plus une science qu'un tas de pierres n'est une maison. » Je n'ai personnellement aucune patience pour l'apprentissage par cœur de listes de savoirs disparates.

Plutôt que de s'intéresser aux données, on peut alors préférer partir en quête des théories suggérées par ces données, typiquement en appliquant la formule de Bayes ! Mais dès lors, quel serait l'objectif ? Serait-il prédictif ? Peut-on considérer que chercher à avoir raison est une fin raisonnable ?

Absolument pas. Chercher à avoir raison, même dans un sens subtile comme celui de la divergence KL, a ses propres limites.

En effet, pour avoir souvent raison, il suffit de s'enfermer dans un problème de prédiction facile. Il suffit de ne s'intéresser qu'à ce que l'on connaît déjà très bien. C'est ce que l'on a tendance à faire, en nous enfermant dans un quotidien qui nous est familier — il nous arrive même de prétendre que ce quotidien est *la réalité*.

La curiosité est donc une mauvaise stratégie pour qui veut avoir souvent raison. *Si un ingénieur ne sait pas ce qu'il fait, il faut qu'il arrête de le faire.* Par opposition, la carrière d'un chercheur l'expose constamment à l'erreur. Personne ne se trompe aussi fréquemment qu'un mathématicien, qui passe ses journées à raturer ses brouillons. *Si un chercheur sait ce qu'il fait, il faut qu'il arrête de le faire.*

Ce problème est très bien mis en valeur par une expérience reproduite par Derek Muller sur sa chaîne Veritasium²². Derek Muller est ainsi allé demander à des gens dans la rue de deviner une règle secrète qu'il avait en tête. Il leur donna un indice : la suite 2, 4, 8 satisfait cette règle. Les personnes interrogées pouvaient alors proposer d'autres suites de 3 nombres, et Derek Muller leur disait si les suites proposées suivaient la règle secrète. Le comportement des personnes interrogées était constamment le même. Elles demandèrent — sans doute après un calcul bayésien inconscient ! — si la règle secrète était une multiplication par 2. Derek Muller répondit non. Elles proposèrent ensuite 16, 32, 64, puis 3, 6, 12, ou encore 10, 20, 40. À chaque fois, la réponse de Derek Muller était la même : oui, ces suites satisfont la règle. Mais non, la règle n'était pas la multiplication par 2. Ce qui laissait les interviewés sans voix.

Le problème, c'est que les propositions étaient constamment motivées par le désir d'avoir raison. Les interviewés avaient leurs règles en tête, et cherchaient

²²  *Can You Solve This?* Veritasium | D. Muller (2014)

constamment à la confirmer — quand bien même Derek Muller avait rejeté cette règle. Néanmoins, ils étaient capables d'effectuer tout plein de prédictions justes. Si leur objectif était qu'on leur dise oui, ils jouaient une stratégie optimale !

Pour trouver la règle secrète, il faut s'exposer au *non*. Il faut chercher à avoir tort. Il faut vouloir rejeter notre intuition. Invité par Derek Muller à ce faire, les interviewés proposèrent ensuite des suites comme 5, 10, 15, ou 2, 4, 7, ou encore 10, 9, 8. Ce à quoi Derek Muller répondit oui, oui et non. Enfin un non ! Les interviewés trouvèrent alors immédiatement la règle secrète : toute suite croissante était un oui pour Derek Muller²³.

Ainsi, un *bayésien* qui cherche à affiner ses crédences bayésiennes ne cherchera pas à avoir raison. Au contraire, il cherchera à expérimenter les domaines où l'étendue de son ignorance est grande. Il va chercher à s'exposer à l'erreur de ses prédictions. En particulier, *avoir raison n'est pas une fin désirable*.

Pire encore, pour avoir l'illusion d'avoir toujours raison, nos cerveaux et ses nombreux biais cognitifs vont constamment rejeter tous les cas où on a tort, et sacrifier les cas où on a raison. Le plus troublant, c'est que ceci peut se faire de manière consciente ou inconsciente. Tel est le fameux biais de confirmation. Ce biais est très dangereux. Il nous pousse constamment à la sur-interprétation et à l'excès de confiance.

L'utilitarisme

La philosophie morale conséquentialiste prédominante consiste à affirmer que le but à atteindre est le bonheur du plus grand nombre. Il s'agit de l'*utilitarisme*. L'utilitariste pense qu'une action est morale si, *in fine*, elle rend plus de gens plus heureux. Reste à déterminer une mesure plus rigoureuse de ce bonheur, ainsi qu'à déterminer ce que l'on entend par « le plus grand nombre ». On peut aussi se demander si l'on tient aussi en compte l'équité, si le bonheur futur compte autant que le bonheur présent, ou s'il y a d'autres objectifs qui pourraient être désirables comme la biodiversité ou la connaissance. L'utilitarisme est malheureusement mal défini. Il en existe un très grand nombre de variantes.

Je ne chercherai pas à lister toutes les difficultés que pose l'utilitarisme ici. La philosophie morale est un sujet passionnant, subtile et difficile. Mais ce n'est pas le sujet de ce livre. Néanmoins, il y a une dimension fascinante à la morale utilitariste souvent négligée, y compris par ses défenseurs : la morale conséquentialiste en général, et utilitariste en particulier, requiert une philosophie de la connaissance performante. En effet, si l'on cherche à rendre les gens heureux, encore faut-il savoir ce qui rend les gens heureux, et comment faire pour arriver à cette fin. Ou plus généralement, pour déterminer quelles actions entreprendre, il faut d'abord *prédir* les conséquences des différentes actions envisageables.

²³Il y a d'ailleurs un champ de recherche qui cherche à optimiser le choix des expériences à effectuer, dont une branche bayésienne appelée *Bayesian experimental design*.

Autrement dit, un bon utilitariste doit d'abord étudier l'épistémologie ; et si l'on en croit le livre que vous êtes en train de lire, il se doit d'être bayésien.

Il arrive souvent que des militants soient religieusement anti-capitalistes, anti-communistes, anti-royalistes ou anti-religions. Pour l'*utilitariste*, de tels jugements ne sont pas permis — ou du moins n'ont pas une forte crédence — tant que l'*utilitariste* ne comprend pas suffisamment les implications du capitalisme, du communisme, du royalisme et des religions, leurs bienfaits et leurs défauts dans le contexte présent. En particulier, Jonathan Haidt affirme que nous sous-estimons constamment la valeur de ce qu'il appelle le *capital moral*²⁴, à savoir toutes les structures sociétales, souvent cachées, et sans lesquelles nos civilisations s'effondraient. Pour l'*utilitariste*, posséder une maîtrise solide des sciences sociales, à commencer par la psychologie, l'économie et les sciences politiques, est un prérequis indispensable, avant d'émettre un avis tranché sur ce que la société devrait faire. En morale (conséquentialiste) plus encore qu'en sciences, « il faut se hâter de ne pas conclure ».

Malheureusement, la quasi-totalité des citoyens a une compréhension très incomplète des concepts de base des sciences sociales. Pire encore, la quasi-totalité des citoyens fait énormément de contre-sens à ce sujet, et ne fait aucun effort pour que cela change. Voilà qui a amené plusieurs intellectuels, comme Bryan Caplan et Jason Brennan, à se positionner contre la démocratie. Leur argument principal est essentiellement utilitariste : si le but est le bonheur, certaines décisions que la société impose à tous ses membres sont nettement meilleures que d'autres ; et une poignée d'experts sera certainement bien plus à même de prendre des décisions plus conformes à l'objectif utilitariste que l'électeur médian.

L'expertise n'est pas le seul argument anti-démocratique de l'utilitariste. Selon Jason Brennan, « la politique abrutit et corrompt ». Plus on s'informe, plus l'hooligan politique en nous prend les commandes et nous impose un parti pris qu'il nous faudra défendre, comme le montrent de nombreuses expériences²⁵. Voilà qui nous conduit à combattre ardemment les idées qui s'opposent à notre intuition, et à refuser toute concession. La politique nous pousse à nous faire des ennemis, et à détester nos opposants. Pour Brennan, un idéal de société n'est pas une société dont les individus participent tous activement à la vie politique ; la société idéale est celle dont les individus passent leur temps à faire ce qui les passionne²⁶.

La démocratie n'est pas le seul pilier de nos sociétés que l'utilitariste remet en cause. L'utilitariste s'oppose aussi au devoir déontologique de réciprocité, que l'on retrouve pourtant dans toutes les grandes religions. « Ne blesse pas les autres de manière que tu trouverais toi-même blessante. » « Tu aimeras ton

²⁴  *La grande Histoire des petites histoires* | Démocratie 26 | Science4All | S. Mombo et L.N. Hoang (2017)

²⁵  *The Righteous Mind: Why Good People are Divided by Politics and Religion* | Vintage | J. Haidt (2013)

²⁶  *7 arguments CONTRE la démocratie* | Démocratie 30 | Science4All | L.N. Hoang (2017)

prochain comme toi-même. » « Ne fais pas aux autres ce que tu ne voudrais pas qu'ils te fassent. » « Aucun d'entre vous ne croit vraiment tant qu'il n'aime pas pour son frère ce qu'il aime pour lui-même. » « Tu aimeras ton prochain comme toi-même. » Voyez-vous la limite de ce principe ? Il présuppose que tout autre a les mêmes préférences que vous.

Par opposition, l'utilitariste va tenir compte du fait que les préférences varient d'un individu à l'autre. Un individu pour qui le sentiment de liberté n'est pas primordial pourrait préférer qu'on le force à se laisser tenter par de nouvelles expériences, tandis qu'un autre individu pourrait sacrifier sa liberté. L'utilitariste sera amené à traiter ces deux individus de manière très différente ; et cette manière de traiter ces individus pourrait différer de la manière dont l'utilitariste souhaite être traité²⁷.

Le problème qui se pose alors à l'*utilitariste*, c'est qu'elle ne connaît pas bien *a priori* ce que l'autre préfère (ou va préférer). Pour le déterminer, il lui faut une philosophie du savoir. Et si l'on en croit ce livre, elle doit être *bayésienne*.

La conséquentialiste bayésienne

En particulier, l'*utilitariste bayésienne* doit exploiter ses préjugés. Ces préjugés sont indispensables pour se comporter de façon optimale pour le bien-être des autres. Si l'*utilitariste bayésienne* est à un enterrement, elle supposera que les blagues morbides ne seront pas bienvenues. Si elle est à un séminaire de mathématiques, elle supposera que les personnes présentes veulent se creuser les neurones. Si elle est en boîte de nuit, elle supposera que personne ne veut y entendre parler de la formule de Bayes.

Toutefois, en bons bayésiens, il ne faut pas oublier que « tous les modèles sont faux ». Pire, même les préjugés bayésiens peuvent conduire à des prédictions erronées. Or, comme on en a déjà parlé au chapitre 9, agir selon certaines prédictions, erronées ou non, peut causer une grande tristesse chez d'autres. Même si l'*utilitariste bayésienne* pense qu'une blague a de très grandes chances de faire rire, le fait qu'il persiste une probabilité non-négligeable que cette blague cause énormément de tort peut l'amener à se retenir. C'est ce qui explique que les premiers rendez-vous sont aussi délicats. Quand il faut apprendre à connaître l'autre sans le vexer, force est de marcher sur des œufs. Ce comportement est celui de l'*utilitariste bayésienne* lorsqu'elle rencontre une personne nouvelle dont elle connaît nécessairement mal les préférences.

L'*utilitariste bayésienne* doit constamment se demander ce que l'autre pense et préfère. Et elle n'agit pas selon son estimation moyenne. Elle va faire attention à éviter les paroles et les gestes potentiellement vexants, même si ces paroles et gestes ne l'auraient pas vexée, elle, et même si la probabilité que ces paroles

²⁷  *La morale des gens heureux : l'utilitarisme* | Démocratie 28 | Science4All | T. Giraud et L.N. Hoang (2017)

et gestes soient vexants est relativement faible. Mieux encore, elle ne perd pas de vue que l'autre a de bonnes chances de mieux savoir ce qu'il préfère, ce qui explique pourquoi l'*utilitariste bayésienne* va souvent préférer laisser à l'autre la liberté de faire ce qu'il veut faire.

De façon plus générale, tout bayésien cherche constamment à quantifier l'étendue de son ignorance. Tout bayésien cherchera aussi à estimer l'ignorance des autres — qui, comme on l'a vu, n'est généralement pas corrélée avec leur manque de confiance en soi. Dès lors, si la *conséquentialiste bayésienne* estime que d'autres sont plus informés qu'elle, et si elle pense que ces autres personnes plus informées prendront des actions bienveillantes, alors il s'agira d'un devoir moral pour elle que de ne pas donner son avis pour ne pas polluer le débat — elle pourrait toutefois intervenir pour questionner les fondements des avis des autres, tester leur expertise, clarifier leurs positions ou apprendre autant que possible d'eux²⁸. Et si elle ne comprend pas tout le raisonnement d'individus plus compétents, mais si elle croit que ces individus sont informés et bienveillants, alors la *conséquentialiste bayésienne* cherchera à laisser ces personnes décider à sa place, ainsi qu'à la place des moins informés²⁹.

Enfin, il est important d'insister sur le fait que la *conséquentialiste bayésienne* raisonne toujours dans l'incertitude. Plus généralement, son processus de décision est un cas classique de la théorie statistique de la décision. Pour commencer, elle assigne un score moral $\mathcal{M}(x)$ à tout état x du monde. Plus l'état x est désirable, plus $\mathcal{M}(x)$ sera grand. Pour la *conséquentialiste bayésienne*, il s'agira d'un devoir moral d'entreprendre une action a qui maximise le score moral espéré, c'est-à-dire la quantité $\mathbb{E}_x [\mathcal{M}(x)|a]$. De façon remarquable, à isomorphisme près, cette approche de la morale est la seule à satisfaire les axiomes de von Neumann-Morgenstern, et donc à ne jamais être victime du paradoxe d'Allais³⁰.

En particulier, le formalisme de la *conséquentialiste bayésienne* permet de traiter les problèmes à faibles risques de grandes catastrophes. Imaginons deux options ✗ et \checkmark . Et considérons trois conséquences possibles : \odot , \ominus et \square . Supposons que ✗ consiste à ne rien faire, et conduit nécessairement à \ominus . Et postulons que \checkmark est l'action risquée, qui peut conduire à \odot ou \square .

On avance parfois le *principe de précaution* pour préférer ✗ . Cependant, \square est alors potentiellement très improbable — et malheureusement, l'invocation du principe précaution n'est pas toujours une invitation au calcul de la probabilité de \square . Or, même si \square est catastrophique, selon la *conséquentialiste bayésienne*, il peut alors être rationnel (ou non) d'entreprendre \checkmark . De façon plus formelle,

²⁸ En pratique, comme on en parlé au chapitre 5, d'un point de vue pédagogique, il est aussi très utile d'exprimer de vive voix ses préjugés pour mieux en être conscient et pour mieux les corriger. C'est typiquement le cas dans l'apprentissage des mathématiques. Cependant, si le but d'un débat est la prise de décision, il pourrait être inopportun de ralentir le processus de prise de décision pour corriger vos préjugés.

²⁹ *Wikipedia et l'épistocratie | Démocratie 31 | Science4All | L.N. Hoang (2017)*

³⁰ *Argent, risques et paradoxes | Démocratie 12 | Science4All | L.N. Hoang (2017)*

la *conséquentialiste bayésienne* va d'abord attribuer des scores aux différentes conséquences. Si RIP est vraiment catastrophique, on peut ainsi imaginer que $\mathcal{M}(\text{☀}) = 0$, $\mathcal{M}(\text{⊗}) = -1$ et $\mathcal{M}(\text{◻}) = -10^9$. Pour déterminer laquelle des actions ✗ ou \checkmark entreprendre, elle va alors calculer les scores espérés. Ne rien faire conduit à $\mathbb{E}_x [\mathcal{M}(x)|\text{✗}] = \mathcal{M}(\text{⊗}) = -1$.

Qu'en est-il de faire \checkmark ? La réponse de la *conséquentialiste bayésienne* tient alors dans le calcul du score espéré lorsqu'on fait \checkmark :

$$\mathbb{E}_x [\mathcal{M}(x)|\checkmark] = \mathcal{M}(\text{☀})\mathbb{P}[\text{☀}|\checkmark] + \mathcal{M}(\text{◻})\mathbb{P}[\text{◻}|\checkmark] = -10^9 P[\text{◻}|\checkmark].$$

En particulier, la *conséquentialiste bayésienne* voudra faire \checkmark si et seulement si $\mathbb{E}_x [\mathcal{M}(x)|\checkmark] \geq \mathbb{E}_x [\mathcal{M}(x)|\text{✗}]$, ce qui revient à $\mathbb{P}[\text{◻}|\checkmark] \leq 10^{-9}$. Autrement dit, sa prise de décision sera entièrement déterminée par la probabilité que l'action \checkmark cause ◻ . Plutôt que de se lancer dans des débats sans fin, la priorité de la *conséquentialiste bayésienne* sera l'estimation de cette probabilité — mais aussi celle de l'incertitude sur cette estimation, voire des actions à entreprendre pour réduire cette incertitude sur l'estimation et des coûts que la réduction des incertitudes nécessitent.

Je pense personnellement que l'on a tous beaucoup à apprendre de cette *conséquentialiste bayésienne*, même si l'on n'aspire qu'à devenir partiellement conséquentialiste. En particulier, pour parvenir à partiellement lui ressembler, il faut absolument prendre la mesure de l'étendue notre ignorance. Malheureusement, quand il s'agit de morale, c'est un effort qui ne nous est pas habituel. J'ai ainsi eu l'occasion d'organiser un débat public sur la « morale des IA³¹ ». Alors que ce thème me semble requérir une grande expertise, aucun participant n'a posé de question. Dans ce genre de débat, on s'empresse à imposer son point de vue, voire à mettre en valeur ses vertus. Et nous avons tendance à conclure bien avant de prendre la mesure de l'étendue de notre ignorance.

Or, comme on l'a vu dans ce livre, notre morale intuitive est constamment en flagrant excès de confiance. Même si l'on a un fondement utilitariste en nous, l'inadéquation de nos préjugés non-bayésiens nous conduit souvent à des comportements qui n'augmentent pas le bonheur global. Pire, notre excès de confiance nous empêche de corriger ces préjugés. Pour devenir de meilleurs êtres moraux, la lutte contre notre excès de confiance semble être une priorité. Puis, dans l'idéal, ceci devra être suivi par une familiarisation avec les sciences sociales et la formule de Bayes, pour affiner nos crédences en les moyens efficaces à nos fins, utilitaristes ou autres.

En particulier, ce livre nous amène à l'étonnante conclusion suivante : *nul n'est moral s'il n'est bon bayésien*.

³¹https://twitter.com/science__4__all/status/983798135431581697

Le mot de la fin

Je ne peux que vous inviter à longuement méditer cette conclusion étonnante. De façon plus générale, j'ose espérer que ce livre aura remis en cause ce que vous croyiez savoir de la morale, de la logique et du savoir. J'espère qu'il vous aura aidé à mieux cerner les limites de la méthode scientifique. J'espère qu'il vous aura aussi aidé à remettre en cause l'excès de confiance dont vous êtes très probablement victime. Et j'espère qu'il vous aura permis d'entrevoir une meilleure façon d'apprendre et de savoir.

Il arrive si souvent que l'on rejette la pertinence des mathématiques et de la philosophie pour adresser le monde « réel ». Il arrive tout aussi souvent que l'on parle du quotidien et du concret comme d'un domaine qui ne nécessite pas une thèse en mathématiques pour être compris. Il s'agit là d'un grave excès de confiance. Comme le disait John von Neumann, « si les gens ne croient pas que les mathématiques sont simples, c'est parce qu'ils ne se rendent pas compte à quel point la vie est compliquée ». Au contraire, notre incapacité à comprendre le paradoxe du corbeau noir, le choix de l'hôpital plutôt que la clinique et la folie de la croissance exponentielle devraient nous forcer à douter de tout ce que l'on pense avoir compris du monde « réel ».

En particulier, si votre morale est partiellement conséquentialiste, le bayésianisme devrait avoir bouleversé votre attachement à vos principes moraux. En effet, si la bonté d'une action dépend (même partiellement) de ses conséquences, alors il nous faut absolument prédire ces conséquences, ainsi que les conséquences des actions alternatives. Or, « il est difficile de faire des prédictions, surtout concernant le futur », dit l'adage. J'espère vous avoir montré à quel point ça l'était.

On a vu que même Paul Erdős était incapable d'appliquer la formule de Bayes dans des cas pourtant ultra-simplistes, que le démon de Solomonoff devait violer les lois de la physique pour effectuer ses calculs bayésiens et que le *bayésien pragmatique* devait jongler entre la théorie de la complexité, la gestion optimale de sa mémoire et l'échantillonnage MCMC pour effectuer une prédition raisonnablement similaire à celle de la *pure bayésienne*. Ceci devrait nous forcer à davantage d'humilité, et à fuir l'excès de confiance bâtant dont on fait si souvent preuve. Surtout concernant les questions morales, « il faut se hâter de ne pas conclure ».

Mais là n'est pas l'objectif principal de ce livre. Comme annoncé lors du premier chapitre, mes longues réflexions épistémologiques au cours des dernières années m'ont fait renoncer à la méthode scientifique et au fréquentisme. Elles m'ont ensuite transformé en bayésien, puis, notamment suite à ma rencontre avec le démon de Solomonoff, en bayésien extrémiste. J'espère vous avoir convaincu que les raisons de cette conversion ne sont pas complètement irrationnelles. Et j'espère vous avoir aidé à entrevoir les grandes lignes de la philosophie du savoir qui, aujourd'hui, a gagné la quasi-totalité de mes crédences.

Mais surtout, j'espère que vous aurez apprécié le périple qu'il nous aura fallu vivre pour explorer les fondements et les conséquences du bayésianisme. J'espère que vous aurez savouré la découverte des nombreuses sciences utiles à la compréhension et à l'illustration du bayésianisme, de l'informatique théorique aux sciences cognitives, en passant par la biologie évolutionniste et la physique statistique. J'espère que vous vous êtes délectés de la démonstration du rasoir d'Ockham, la décortication du scandale de l'induction et la remise en cause du réalisme. Et j'espère que la lecture de ce livre aura été pour vous un voyage exotique, voire initiatique, dont vous garderez des souvenirs impérissables.

Avant tout, j'espère vous avoir submergé d'enthousiasme, de fascination et de questionnements. Tel fut l'objectif premier de ce livre.

Références en français

 *Le temps des algorithmes* | Le Pommier | S. Abiteboul et G. Dowek (2017)

▶ "Calculer le bonheur" - *L'utilitarisme classique* | Politikon | K. Piriou (2018)
 ▶ *La rationalisation* | La Tronche en Biais | V. Tapas et T. Durand (2015)

▶ *Si Dieu n'existe pas, tout est permis ? Religion et morale (1/2)* | Grain de Philo | Monsieur Phi | T. Giraud (2016)

▶ *La religion nous empêche-t-elle d'être morale ? Religion et morale (2/2)* | Grain de Philo | Monsieur Phi | T. Giraud (2016)

▶ *Nietzsche - La généalogie de la morale* | De Dicto | Politikon | K. Piriou (2016)

▶ *Nietzsche - La morale des winners ! Généalogie de la morale (1/2)* | Grain de Philo | Monsieur Phi | T. Giroud (2017)

▶ *Nietzsche et les méchants - Généalogie de la morale (2/2)* | Grain de Philo | Monsieur Phi | T. Giraud (2017)

▶ *À chacun sa morale ? Relativisme et réalisme* | Grain de Philo | Monsieur Phi | T. Giraud (2017)

▶ *Conséquentialisme - Quel est le but de la morale ?* | Grain de Philo | Monsieur Phi | T. Giraud (2017)

▶ *Les synonymes à connotations opposées* | My4Cents (Mayen) | Science4All | L.N. Hoang (2016)

▶ *L'hooliganisme politique a gâché ma marche pour les sciences* | My4Cents (Genève) | L.N. Hoang (2017)

▶ *Le scrutin de Condorcet randomisé (mon préféré !!)* | Démocratie 5 | Science4All | L.N. Hoang (2017)

▶ *Êtes-vous un hooligan politique ?* | Démocratie 10 | Science4All | L.N. Hoang (2017)

- ▶ *Rationnellement irrationnels* | Démocratie 11 | Science4All | L.N. Hoang (2017)
- ▶ *Le paradoxe de la morale* | Démocratie 25 | Science4All | S. Debove et L.N. Hoang (2017)
- ▶ *La grande Histoire des petites histoires* | Démocratie 26 | Science4All | S. Mombo et L.N. Hoang (2017)
- ▶ *L'instinct tribal* | Démocratie 27 | Science4All | L.N. Hoang (2017)
- ▶ *L'utilitarisme* | Démocratie 28 | Science4All | T. Giroud et L.N. Hoang (2017)
- ▶ *7 arguments CONTRE la démocratie* | Démocratie 30 | Science4All | L.N. Hoang (2017)
- ▶ *Wikipedia et l'épistocratie* | Démocratie 31 | Science4All | L.N. Hoang (2017)
- ▶ *Fat Tony et Dr John (biais-variance)* | IA 12 | Science4All | G. Mitteau et L.N. Hoang (2018)

- ⌚ *Informatique et éthique* | Podcast Science 266 | G. Dowek et S. Abiteboul (2016)
- ⌚ *Utilitarisme artificiel* | Axiome 3 | T. Giraud et L.N. Hoang (2017)

Références en anglais

- 📘 *The Myth of the Rational Voter: Why Democracy Always Chooses Bad Policy* | Princeton University Press | B. Caplan (2007)
- 📘 *Thinking Fast and Slow* | SpringerFarrar, Straus and Giroux | D. Kahneman (2013)
- 📘 *The Righteous Mind: Why Good People are Divided by Politics and Religion* | Vintage | J. Haidt (2013)
- 📘 *Superintelligence: Paths, Dangers, Strategies* | Oxford University Press | N. Bostrom (2014)
- 📘 *Against Democracy* | Princeton University Press | J. Brennan (2016)
- 📘 *The Big Picture: On the Origin of Life, Meaning and the Universe Itself* | Dutton | S. Carroll (2016)

- 📘 *Motivated Numeracy and Enlightened Self-Government* | Behavioural Public Policy | Dan Kahan, Ellen Peters, Erica Cantrell Dawson, Paul Slovic (2017)
- 📘 *Human-level control through deep reinforcement learning* | Nature | V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. Bellemare, A. Graves, M. Riedmiller, A. Fidjeland, G. Ostrovski, S. Peterson, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg et D. Hassabis (2015)

- ▶ *New Coke - A Complete Disaster?* Company Man (2017)
- ▶ *Why you think you're right — even when you're wrong* | TED | J. Galef (2016)

- ▶ *Politics and Numbers* | Numberphile | J. Grime (2013)
- ▶ *Why Asimov's Laws of Robotics Don't Work* | Computerphile | R. Miles (2015)
- ▶ *Can You Solve This?* Veritasium | D. Muller (2014)
- ▶ *The Illusion of Truth* | Veritasium | D. Muller (2016)

Remerciements

Écrire ce livre a été un périple incroyable, avec beaucoup de hauts et de bas. Je n'aurais pas pu achever ce périple sans l'aide, le soutien et l'intelligence de personnes remarquables envers lesquelles je suis très redevable.

Je suis particulièrement redevable envers Thibaut Giraud, Julien Fageot, Maxime Maillot, David Loureiro et Marie Maury, pour leur relecture détaillée et leurs retours très utiles. Plus généralement, je les remercie grandement de m'avoir accompagné dans mes réflexions bayésiennes. À ce titre, je remercie également grandement Peva Blanchard, Rachid Guerraoui, El Mahdi El Mhamdi, Alexandre Maurer, Julien Stainer, Hadrien Hendrikx, Sébastien Rouault, Matej Pavlovic, Clément Hongler, Christophe Michel et Sébastien Carassou parmi tant d'autres, qui ont questionné, stimulé et aiguisé mes méditations bayésiennes. En particulier, l'environnement intellectuel stimulant de mon université, l'École Polytechnique Fédérale de Lausanne (EPFL), m'aura énormément aidé. Je remercie aussi grandement EDP Science pour l'édition de ce livre, et Gilles Dowek pour avoir eu l'amabilité de prendre le temps d'écrire la préface de cet ouvrage.

Plus généralement encore, je remercie tous ceux qui me suivent, de près ou de loin, notamment sur les réseaux sociaux et sur YouTube. Ils y sont beaucoup pour l'envie que j'ai, au jour le jour, de vouloir comprendre et clarifier autant que possible ce que je pense savoir, et ce que je m'apprête à découvrir. Je remercie aussi les nombreux copains du Café des Sciences et du YouTube culturel. Leur amitié, leur bienveillance et la qualité sublime de leur vulgarisation sont une source d'inspiration inépuisable.

Mais surtout, je vous remercie vous, chers lecteurs. L'idée de pouvoir partager avec vous mes aventures bayésiennes est une joie et une motivation grandiose.