# Invitation à la théorie de l'information, Emmanuel Dion

Notes de lecture

Arabella Brayer

September 15, 2016

# Contents

# 1

# Introduction

## Le concept d'information

Le terme "information" désigne une notion difficile à décrire de façon simple et sans emphase, ou sans user d'évidences qui n'apportent aucune information utile. Pour ce faire, on peut s'inspirer de l'analogie entre l'information et l'énergie, notion aux multiples formes également. D'autre part, remarquons que de tout temps, la plupart des inventions ont servi à maîtriser l'une ou l'autre : énergie, information. Quelques exemples : la radio, le téléphone, l'informatique, etc.

## Épistémologie

Du point de vue de l'épistémologie, on peut également rapprocher l'information de l'énergie. On constatera alors que les deux ont été employées avant de savoir les définir de façon formelle. C'est avec la théorie de Shannon que l'information a acquis un sens précis, ainsi qu'une unité de mesure : le bit. C'est la parution du livre de Shannon en 1948 qui marque ce tournant, et qui restera dans l'histoire des sciences du XX$^o$. Dès ce moment, un nom-

bre important de publications sortent à ce sujet, et la recherche clarifie son discours.

Actuellement, la densité de travaux s'est certes un peu tarie, néanmoins l'ensemble de ces travaux sont rassemblés derrière l'expression "théorie de l'information" (ainsi que "théorie de la communication" [1] ) et est largement reconnue.

Parmi les théories existantes en sciences, on pourrait trouver des éléments similaires entre la théorie de l'information et la théorie des jeux : double composante mathématique et conceptuelle, ainsi qu'une large diffusion. D'ailleurs, même si le lien entre ces deux théories ne saute pas à la conscience, elles entretiennent des relations, qui seront détaillées plus tard.

## Utilisations de la théorie

La théorie de l'information a été vue de façon différente dans la science : ainsi a-t-elle apporté à plusieurs domaines, tels que la biologie, la psychologie, etc. Mais son caractère "généraliste" lui a "permis" d'être largement citée en philosophie. Il s'agirait plutôt d'un emploi abusif. On pourrait tenter de réduire la théorie de l'information à quelques opérateurs mathématiques, déjà connus, mais réunis dans cette théorie. On peut également la voir comme une théorie primordiale pour le XX$^o$ siècle.

**Problématique** : Ce débat a-t-il lieu d'être ou pourrait-on imaginer que ces deux propositions ne se rassemblent ?

[1] "Théorie de la communication" est une expression qui désigne la même chose strictement, contrairement à ce que laisse entendre son nom. Shannon lui-même aurait préféré l'usage de l'expression "théorie de l'information".

# 2

# La théorie de l'information : une théorie transversale au cœur de la science moderne

## 2.1 Section heading

La théorie de l'information ne s'intéresse absolument pas à la signification, au sens, contrairement aux autres théories en communication, focalisées sur cet aspect. Weaver et Shannon n'ont jamais souhaité donner une aura autre que technique à cette théorie, rappelons que cette époque est celle où l'on souhaite améliorer la qualité des transmissions. Les débordements sémantiques n'ont sans doute pas lieu d'être et surtout, ne sont pas du fait de ces deux personnes.

## 2.2 Les racines de la théorie

L'origine de la théorie vient du besoin de délimiter les capacités de transmission d'un message, soit par l'intermédiaire du canal de communication

directement, soit par son système de codage. Différents systèmes binaires avaient déjà vu le jour à divers endroits du globe. Ces systèmes possèdent des caractéristiques intéressantes, comme la possibilité d'employer les combinatoires, et d'avoir des propriétés au codage. Le morse est "efficace" à 85%, bien qu'inventé vers 1830, ce qui est très bien. Construire un code efficace nécessite une théorie sur les fréquences d'apparition des lettres ( 1300), des digrammes ( 1600), et celles-ci n'étaient pas encore réunies.

## 2.3 L'approche statistique : l'information de Fisher

Fisher a commencé à considérer l'information comme une quantité mesurable, vers 1920. Il la définit comme étant la valeur moyenne du carré de la dérivée du logarithme de la loi de probabilité étudiée.

## 2.4 L'approche des ingénieurs : les travaux de Nyquist et Hartley

Parallèlement, en 1922, on trouve des premières pistes pour améliorer la qualité et vitesse de transmission des signaux radio. La formule
$W = K \times \log M$ résume ici que l'on considère le caractère comme unité, K étant une constante dépendant de la qualité de la ligne. On note le log, dont on reparlera. Il faut attendre 1948 pour que Shannon fasse progresser la matière.

## 2.5 L'apport de Shannon

L'objectif de Shannon est avant tout d'améliorer les rendements des lignes de télégraphe. Shannon n'est pas un grand érudit mathématique, il résout

magnifiquement des problèmes complexes mais pratiques, plus qu'abstraits. C'est un homme humble, honnête intellectuel, scientifique. Son article déclenche de grands mouvements scientifiques, mais il reste tel qu'il est, préoccupé par des problèmes d'une priorité discutable.

## 2.6 Le MIT, plaque tournante du développement des sciences de l'information

Shannon est d'abord élève, puis professeur au MIT, ce qui va beaucoup l'influencer. Il rencontre Wiener, et les deux se citent régulièrement l'un l'autre dans leurs travaux. Wiener et Shannon arrivent à des conclusions similaires en partant de deux problématiques légèrement différentes. Wiener arrive à quantifier la quantité d'information par $log_2 \frac{quantité-a-priori}{quantité-a-posteriori}$ Wiener étend sa définition, et la rapproche de Von Neumann, de distribution continue de probabilité : $\int f(x) \times \log_2 f(x) \times dx$

Il ne faut pas oublier qu'à l'époque, on écrit déjà des programmes sur cartes perforées, et l'on dispose d'appareils déjà évolués capable de résoudre des problèmes complexes comme extraction d'une racine carrée, etc. Finalement, la technique était en avance sur la théorie.

À cette époque, la logique de Boole n'est pas associée à l'informatique, ni même Turing. La référence était plutôt Von Neumann (lié à Goldstein), directeur du secteur mathématique d'IBM.

Les logiciens de l'époque ne s'intéressent pas à l'informatique.

On peut citer David Slepian, comme contributeur important également : créateur des codes correcteurs) Peter Elias, David Huffman, Warren McCullough (Research Lboratory of Electronics).

## 2.7 Un débat scientifique animé et ouvert

Le livre de Shannon donne lieu à des controverses. D'abord, sur l'emploi des mots : Information, entropie, bruit, cybernétique... L'usage du terme "information" concernant la théorie de l'information peut laisser supposer que l'on s'intéresse à la communication du sens, mais il n'en est rien. On peut dire autrement : "le logarithme du maximum de vraisemblance d'une distribution multinomiale", on s'aperçoit qu'il n'y a pas lieu d'en faire de la philosophie. Cela a pourtant été fait, mais on ne peut pas le reprocher à Shannon, qui n'a eu de cesse de rappeler que ce n'était pas là l'ambition de sa théorie. Notons que les critiques ont toujours été faites quant à la théorie elle-même et non à l'endroit de Shannon.

Concernant l'entropie, le choix du mot renvoie forcément à un terme précis en thermo-dynamique, et l'on peut se demander si ce choix de mot est judicieux. Précisons que Shannon a écrit sa théorie sans mesurer l'impact qu'elle pourrait avoir par la suite, et a peut-être sous-évalué l'importance du choix des mots.

D'autres mots s'ajoutent, comme le bruit, la redondance, etc. ouvrant la porte à toutes sortes d'idées plus large. Mais insistons bien : la théorie de l'information de Shannon porte sur quelque chose d'assez technique finalement, qui est la communication dans un canal "physique". On peut se demander si sa théorie aurait eu le succès qu'elle a eu si les mots choisis n'avaient pas permis toutes ces extrapolations extravagantes.

Citons deux "camps" : W. Weaver, L. Brillouin, E. Jaynes, M. Tribus, E. Schoffeniels, T. Stonier, très "favorables".

L. Cronbach, H. Quastler, B. Mandelbrot, A. Lwoff, D. McKay, C. Waddington, R. Thom, etc. vont - quant à eux - contester les interprétations "abusives".

## 2.8 Une opposition qui porte sur des thèmes fondamentaux

Au regard des critiques, on pourrait penser que cette querelle est stérile, mais il n'en est rien. En réalité, ces critiques apportent beaucoup du point de vue de l'épistémologie. On peut bien voir un rapport entre l'information et l'ADN/ARN, etc.

La théorie de l'information rassemble également deux courants qui s'affrontent chez les probabilistes : le mouvement des fréquentistes, et des subjectivistes. Les fréquentistes croient qu'une expérience se doit d'être répétée afin de pouvoir modéliser les statistiques, l'autre pense que l'on peut théoriser le hasard pour analyser un modèle. On comprend que cela suscite alors des discussions animées.

## 2.9 Les aspects épistémologiques du problème

Entre la naissance de la théorie et les recherches, il s'est écoulé environ 5 à 6 ans. Depuis, quelques personnes continuent de s'intéresser au sujet, mais l'âge d'or est passé. Références : F. Resa, J. Wolfowitz, P. Elias, A. Kolmogorov (utilise l'entropie comme concept de base pour la classification des systèmes dynamiques).

La définition même de l'entropie n'a que peu d'importance dans le cadre rhétorique. Que l'on puisse en parler - point de vue rhétorique - n'apporte rien à la compréhension mathématique du sujet. La chose se complique lorsqu'il faut quantifier.

On peut rapprocher la théorie de l'information de ce point de vue avec la théorie des jeux, car les deux trouvent difficilement des opportunités physiques réelles de réalisation : la théorie des jeux suppose de posséder

des matrices de choix... Ce que concrètement, l'on n'a jamais. La théorie de l'information peut éventuellement porter sur quelque chose de physique, comme la génétique, mais au delà, on dispose peu de conditions compatibles avec la théorie. La preuve ne peut donc passer pour justification physique. D'où le rapprochement avec Godel, Von Neumann, etc., dans la percée de ces théories dont la déduction est à l'origine d'une réponse efficace/inefficace au monde physique.

# 3

# Des opérateurs mathématiques d'une grande élégance

## 3.1 La mesure de l'information : pourquoi le logarithme ?

La théorie ne comporte rien de très complexe sur le plan de la formalisation mathématique. Il faut maîtriser deux définitions : la quantité d'information, et l'entropie. On peut - après ça, s'intéresser à la redondance ou au bruit mais ces premières sont plus importantes pour la compréhension du sens de la théorie.

L'information, chez Shannon, désigne un (ensemble d') événement(s) parmi un ensemble d'événements possibles. Toutes les mesures qui caractérisent ces événements sont probabilistes.

Prenons comme exemple la recherche d'un livre dans une bibliothèque. Si l'on connaît une information "pertinente", cela peut réduire le temps de

recherche. Par exemple, si l'on sait qu'on recherche un livre avec une couverture bleue, et qu'il y en a 1/4 dans la bibliothèque, c'est une information importante.

On cherche donc à quantifier l'information comme ce qui réduit l'incertitude. Dans l'exemple cité avant, on a l'intuition que les nombres des livres concernés peuvent modifier l'information : plus de livres au total, plus de livres bleus, etc. Mais Shannon utilise pourtant le logarithme :

$qté\_d\_infos = I = log(\frac{N}{n})$, ce qui permet de conserver les propriétés additives du logarithme. (rappel : $\log(a \times b) = \log a + \log b$ ) Concernant la base du logarithme, si la base 2 a été choisie, c'est pour une raison principalement arbitraire : $\log(2) = 1$ en base 2, or, il a été fixé que l'information valait 1 lorsqu'il y avait dichotomie parfaite. Le logarithme est un choix judicieux : positivité, additivité, et la base 2 pour indiquer la dichotomie parfaite.

On voit cependant que toutes les informations ne sont peut-être pas équiprobables : en langue française, la fréquence des lettres est inégales. Dans ce cas, l'information sera $I = \log \frac{1}{p}$ où p est le degré d'apparition, soit $I = -\log p$.

Cette unité sera appelée le *bit* par Shannon (porte d'autres noms pour d'autres). Un bit peut se définir de cette façon :

*la quantité d'information qui correspond à la réduction de moitié de l'incertitude sur un problème donné*

Reprenons l'exemple de la bibliothèque et étudions l'information : Mettons qu'il y ait 4000 livres, et 500 bleus. L'information "le livre recherché est bleu" devient : $\log(\frac{4000}{500}) = \log 8 = 3$ : 3 bits. On peut expliquer ça comme ça : on a divisé 4000 par 8 ($\frac{4000}{500} = 8$) or pour écrire de 0 à 7, il faut 3 bits en binaire. Pour savoir dans quel tas chercher, on a donc besoin de 3

bits.

Si on avait eu 1000 livres bleus, on aurait eu besoin de 2 bits, car $\log 4 = 2$. L'information est de moindre "valeur" dans ce cas, car les "tas" seront de 1000 livres...

On le voit dans cet exemple : la théorie de l'information est purement quantitative. On peut aussi se demander la quantité d'information I(bleu clair) contenue dans l'affirmation "le livre cherché est bleu clair". Contrairement à ce qu'on pourrait penser de façon intuitive, la réponse n'est pas $I(bleu) + I(clair) => 2+3 = 5bits$ mais bien : $I(bleu) = \log \frac{4000}{250} = \log 16 = 4$. La différence provient du fait que les informations sont dépendantes.

Si l'on fait le même exercice mais avec I(rouge clair), cette fois-ci, on obtient $I(rougeclair) = I(rouge) + I(clair)$, et on peut en déduire que les informations sont indépendantes. En probabilités, on formule cela de cette façon : $P(rougeclair) = P(rouge) \times P(clair)$.

On peut regrouper des informations ensemble. Attention, car selon qu'elles sont dépendantes ou indépendantes, on obtient pas la même chose.

On peut dénombrer ainsi 3 cas de figure :

1. L'information totale est inférieure à la somme de ses parties. Se produit quand il y a dépendance, une information rend l'autre moins importante.

2. L'information totale est égale à la somme de ses parties : Les informations sont indépendantes

3. L'information totale est supérieure à a somme des parties : Il y a dépendance, une information rend l'autre plus importante.

Ceci permet d'en tirer une propriété multiplicative : Prenons le cas d'un alphabet binaire, avec équiprobabilité de 0 et de 1. Chaque symbole

est porteur de $\log 2$ d'information, soit 1. Si le message est composé de n symboles, alors on obtient : $I = \log 2^n = n \times \log 2 = n$ bits d'information.

En conclusion, on peut calculer l'information avec de longs messages aussi bien qu'avec des courts. Cette distance nous amène aussi à nous intéresser à un autre concept primordial : **l'entropie**.

## 3.2 L'entropie revisitée

L'entropie est un concept aussi fondamental pour la théorie de l'information que l'information. Du fait de son utilisation dans cette théorie, il se retrouve présent dans énormément d'autres domaines. Ainsi peut-on parler de l'entropie d'un style musical, d'une langue étrangère, etc.

L'information mesure plutôt la quantité "transmise", une production. L'entropie, elle, se concentre plutôt sur le potentiel *avant* la transmission du message, ce qui permet de comparer différents canaux, différentes sources, récepteurs, en fonction de leurs propriétés.

On note : $H = \sum_i p_i \times \log \frac{1}{p_i}$ où $p_i$ désigne la probabilité de l'événement i.

Elle peut sembler fort abstraite, mais appliquée à un exemple, elle prend tout son sens : Prenons le morse, avec $P(trait) = 0.75$ et $P(point) = 0.25$. La quantité d'information vaut $I(trait) = -\log 0.75 = 0.415$ bits, $I(point) = -\log 0.25 = 2$ bits. L'apparition d'un point pèse plus lourd que celle d'un trait. Par ailleurs, l'information d'un trait vaut moins qu'une unité, celle d'un point, deux.

Maintenant, si l'on prend un peu de hauteur, on imagine que cette expérience sera reproduite un nombre important de fois. Alors on obtient : $H = 0.75 \times 0.415 + 0.25 \times 2 = 0.811$ bits. Intuitivement, cela représente la *propension* d'un canal à émettre une certaine quantité d'information. C'est

une information moyenne.

Dans le cas où il y a indépendance des probabilités, la somme se simplifie : $H = \log \frac{N}{n}$ bits. On peut en déduire que plus une distribution est équiprobable, plus l'entropie est forte.

## 3.3 L'envers de l'information : la redondance

La redondance diffère de la répétition, qui est un cas particulier de celle-ci. Un terme pour la décrire serait plutôt "corrélation". En guise d'exemple, dans la langue française, un q est quasi systématiquement suivi d'un u, même si ces lettres ne sont pas les mêmes : ce n'est pas de la répétition, mais la lettre suivant q est moins difficile à deviner qu'après un a, ou un e. L'information du u après le q est donc peu informative. La redondance évoque ce caractère peu informatif.

Afin d'optimiser l'information, on comprend que cette redondance doit être le plus faible possible. Cela peut être le cas dans un alphabet où chaque caractère est équiprobable. Dans ce cas, la redondance est nulle, et l'entropie atteint $\log(n)$ par symbole émis. La redondance d'une source est définie comme la différence entre l'entropie de la source et $\log(n)$.

Cependant, on ne cherche pas toujours à l'éviter. Ainsi, sur un canal bruyant, la redondance peut permettre de vérifier l'intégrité du message.

Les langues sont rarement les plus "efficaces" possibles. Ceci explique pourquoi chaque fois que la transmission d'un message est coûteuse, l'on cherche à mettre en place des systèmes plus courts, résumés, comme la sténographie, les abréviations, etc.

### 3.3.1 Un nouveau sens pour les notions de bruit, d'équivoque et d'ambiguïté

Le bruit est le mot qui définit la modification d'un message entre son émission et sa réception. Dans le cas d'un canal bruyant, on ne peut pas être sûr que le symbole reçu soit celui qui a été envoyé. On peut avoir une mesure du doute en probabilités. Dès lors, tout échange devient probabiliste. L'émetteur et le récepteur ne vont pas avoir accès aux mêmes informations. Ainsi, le récepteur ne peut calculer l'entropie qu'à partir des symboles reçus. C'est alors qu'interviennent les probabilités sous forme de "a sachant b", c'est à dire la probabilité que a ait été émis quand b a été reçu.

Le récepteur peut calculer l'entropie de cette façon : $H(a|b_j)$. Si le canal n'est pas bruyant, alors l'entropie est nulle car b correspond toujours à a. Dès lors, on peut calculer $H(A|B)$, ce qui est appelé "l'*équivoque* du canal".

La formule exacte de l'équivoque est : $H(A|B) = -\sum_{i,j}[p(a_i b_j).log(p(a_i|b_j))]$

(il y a une démonstration dans le livre, mais je me dispense d'en faire le résumé ici)

De façon symétrique, l'émetteur peut calculer l'incertitude que son message arrive à son destinataire. Cette mesure s'appelle l'ambiguïté, et sa formule est

$H(B|A) = -\sum_{i,j}[p(a_i, b_j).log(p(a_i, bj))]$

et on peut montrer que $H(A|B) \leq H(A) + H(B)$ (il y a aussi une démonstration dans le livre dont je me dispense également)

Un cas particulier concerne le cas où ces deux quantités sont égales :

$H(A|B) = H(A) + H(B)$.
Par conséquent, une autre quantité (la différence entre ces deux valeurs) a été définie : $T(A, B) = H(A) + H(B) - H(A, B)$ appelée *transinformation*.

Cette valeur varie en fonction de la dépendance entre A et B. Si ces deux valeurs sont dépendantes, la transinformation augmente.

On peut en déduire :
$H(A, B) = H(A) + H(B|A)$
$H(A, B) = H(B) + H(A|B)$
$T(A, B) = H(A) - H(B|A)$
et les démonstrations sont dans l'ouvrage.

Dans le cas où la liaison entre A et B est parfaite, on a $T(A, B) = H(A) = H(B)$ car l'ambiguïté entre A et B est nulle.

Voici quelques illustrations afin de comprendre ce que signifient ces formules. Mettons un émetteur A et un récepteur B.

Dans le premier cas, lorsque A essaie de transmettre un message 1, B reçoit avec une certaine probabilité ce même message. Idem pour un autre message. Dans ce cas, le bruit est nul.

Ce calcul se complique lorsque A essaie de transmettre un message, mais que B reçoit celui-ci, ou un autre. Dans ce cas, la transinformation diminue, et on dit que le canal est bruyant. Dans le pire des cas, on ne peut jamais être sûr du message reçu, la transinformation est de 0.

Un autre cas est celui où A essaie d'envoyer le msg 1, et B reçoit soit le msg 1 soit 2 avec une probabilité de 0.5. Si A envoie le msg 2, il n'est jamais reçu par B. Dans ce cas, on perçoit que quelle que soit l'idée exprimée par A, B l'interprète de la même façon. L'équivoque est donc très élevée, et l'ambiguïté peut prendre n'importe quelle valeur.

Dans le cas inverse, si A ne transmet que le msg 1 et que B reçoit soit le msg 1 soit 2, ces valeurs sont inversées : l'équivoque est quelconque, et l'ambiguïté est élevée.

Globalement, il y a équivoque lorsque le récepteur n'est pas aussi fin que l'émetteur, et il y a ambiguïté lorsque c'est l'émetteur qui n'est pas assez fin pour le récepteur.

*Équivoque* : deux messages peuvent être compris de la même manière.
*Ambiguïté* : un message peut être compris de plusieurs façons.

La capacité d'un canal est la transinformation maximale qu'on peut obtenir avec la loi de probabilité de la source la plus avantageuse possible. Des théorèmes issus de la théorie de l'information ont montré qu'une telle valeur limite est très utile, notamment car il existe toujours des codes permettant d'utiliser la totalité de la capacité d'un canal, même bruyant. Ces résultats ont été la source de nombreux développements de la théorie de l'information dans la communauté mathématique.

Malgré la force de ces valeurs, ce ne sont pas ces éléments qui ont rendu célèbre la théorie de l'information, mais des thèmes plus philosophiques

comme la surprise, l'ordre, et la complexité.

## 3.4 À la croisée de plusieurs concepts psychologiques et philosophiques essentiels

### 3.4.1 L'information comme réduction de l'incertitude

Quoi que la théorie soit appelée "théorie de l'information", l'incertitude est un de ses thèmes centraux, si bien que certains ont même axé leur étude autour de celui-ci. Ils sont exactement opposés. En effet, le gain d'incertitude fait baisser l'information. Il faut comprendre que la théorie de l'information de Shannon n'est valable qu'en considérant un ensemble fini et probabilisé. Dans le cas d'un ensemble infini, la théorie n'a aucun sens.

Un point soulevé concerne la mauvaise perception des probabilités par les humains. *Ref citée : Harold W. Hake, The perception of frequency of occurrence and the development of "Expectancy" in humain experimental subjects"*, on y rappelle la très mauvaise intuition en la matière par les humains. Ceci permet d'en arriver à la notion psychologique de la "surprise". Celle-ci s'analyse en effet très bien à l'aide de la théorie de l'information.

### 3.4.2 L'information comme résultat de la surprise

Prenons le cas d'un professeur qui distribue un polycopié à une classe. À chaque distribution, les élèves répondent "merci". Cette intervention est attendue et très peu surprenante.

Mettons qu'à la fin du cours, un seul de ces élèves viennent dire "merci"

au professeur. L'essence de cette intervention est bien plus lourde de sens ici, car elle n'est pas attendue, et le message est différent : dans le premier cas, on remercie par usage.

Dans le second, on exprime son remerciement pour la prestation. Les chercheurs séparent deux notions pour exprimer la surprise : *surprisal* est fonction de la quantité d'information telle que définie par la théorie de l'information. *surprise* mesure quant à elle la composante psychologique de la surprise.

En guise d'exemple : si l'on lance un dé 5 fois, obtenir 5 faces a la même valeur de surprise que pile-face-pile-face, et pourtant, elle a une valeur de surprise.

Les événements surprenants sont plus riches d'information que les événements routiniers. Une faute d'orthographe pourra en effet soulever un questionnement : faute volontaire ? que signifie le message, etc. L'information est plus importante. Le fait de demander un message diminue également la surprise, ou l'information.

Bref, tout ceci se trouve à la limite de la théorie de l'information voire au delà. On voit qu'ici, les opérateurs mathématiques ne sont plus efficaces, et le vocabulaire de la théorie sert de réservoir de vocabulaire, mais ne sert pas vraiment d'outil. Cela n'a pas été souhaité par Shannon, et cela dépasse le cadre de la théorie.

Si l'on revient à la théorie en elle-même pour se recentrer sur son sens

mathématique, on s'aperçoit que la théorie de l'information donne des outils pour décrire une notion primordiale dans les sciences dures et sociales : la complexité.

### 3.4.3 L'information comme mesure de la complexité

La théorie de la complexité est intuitive, elle est pourtant difficile à calculer. C'est Kolmogorov, et Gregory Chaitin qui sont considérés comme ses fondateurs (quoi que cette parenté soit discutée).

En guise d'exemple, prenons une chaîne composée de 0 et de 1, et qui alternent, ou une autre :
01010101010101010101
01110010011011010010
Si l'on souhaite prévoir une suite à ces chaînes, pour la première, on trouvera aisément, quant à la deuxième, on ne pourra rien prévoir. Mais sur le plan de l'information, elles sont deux chaînes de 20 caractères qui ont la même chance de tirer avec un tirage aléatoire de 0 et de 1. Il faut chercher plus loin avec le concept d'aléa pour voir ce qui les différencie.

Ici intervient la notion de compressibilité : une suite qui peut se réduire à une suite plus courte sans perdre d'information n'est pas une suite aléatoire. La suite qui alterne les 0 et les 1 est compressible, la seconde ne l'est pas. Cela étant, le problème est moins simple qu'il n'y paraît. Mais dans tous les cas, il est possible de décrire la suite en un nombre minimal de bits. La *complexité* est le nombre minimal de bits qui est nécessaire à la description de la suite, dans une machine de turing, par exemple.

En mathématique, globalement, ce qui est novateur est complexe, ce qui

est répétitif est trivial. Si l'on énonce deux exercices similaires, il faudra donner des détails pour le premier voire expliquer la procédure. Pour le deuxième, il suffit d'écrire "comme le premier".

**Evariste Galois** est un mathématicien célèbre du XXème siècle. qui a posé les bases de la théorie des groupes, on raconte qu'il a écrit sa théorie en une nuit avant sa mort. Cela est sans doute romancé, toujours est-il qu'il apparaît ici que des choses primordiales et importantes peuvent s'énoncer rapidement et sans redondance. Dans cet ouvrage, les démonstrations ne sont pas faites, et si le lecteur peut les faire lui-même, cela signifie qu'il aurait été redondant de les écrire. Cela signifie que l'on se mette d'accord sur le contenu de l'alphabet de l'échange. Par exemple, pour transmettre une configuration de jeu d'échec, on peut tout décrire, ou décrire par une référence à une partie, et cela si l'on présuppose que le destinataire du message connaît cette référence.

### 3.4.4 L'information dans la problématique de l'ordre et du désordre

Au début de cet ouvrage était mentionnée la proximité entre l'information et l'énergie. Nous allons voir que ce lien est à manier avec précaution, mais qu'il s'est trouvé au centre de la théorie de l'intérêt lorsqu'une proposition de résolution du paradoxe de Maxwel a été proposé vers 1950.
Afin d'exposer ce paradoxe, il faut rappeler les bases de la thermodynamique.

### 3.4.5 Second principe de la thermodynamique et démon de Maxwel

Le second principe énonce qu'il y a une valeur (l'entropie) qui peut varier, mais toujours dans le même sens. Le second principe est directionnel, irréversible. Pour illustrer : prenons un réfrigérateur dont on ouvre la porte. Les températures se mélangent naturellement, sans apport d'énergie. Pour retrouver la fraîcheur initiale, il faut un apport d'énergie.

**L'entropie** est la mesure physique du désordre. On postule que l'univers tend vers une entropie maximale, et qu'elle sera atteinte lorsqu'il y aura homogénéité.

Pour en revenir au démon de Maxwel, il s'agit d'une histoire où dans un grand récipient se trouvent deux compartiments. Le démon possède la possibilité d'ouvrir durant des temps brefs une porte entre les deux et de faire passer les molécules qu'il sélectionne, de sorte que le compartiment A va toujours se réchauffer, et le B refroidir. On voit que cette situation est un paradoxe si on accepte le fait qu'il ne coûte pas d'énergie d'ouvrir la petite porte au démon. Les lois qui régissent la répartition des particules de gaz sont probabilistes, et on peut résumer l'entropie par la mesure des possibilités de mouvements des particules.

$S = k.\ln(W)$ est la formule physique de l'entropie. En fait, la formule est très proche de cette de l'entropie dans la théorie de l'information si bien que certains ont postulé que c'était l'information qui était nécessaire au démon pour savoir quand ouvrir la porte, idée postulée par Léon Brillouin. Mais les concetps restent différents, aussi a-t-il proposé d'adopter une autre terminologie : le terme **néguentropie** pour l'entropie qui concerne la théorie de l'information.

### 3.4.6 Notion de néguentropie et paradoxe de l'information négative

L'entropie a tendance à augmenter. Dans le meilleur des cas, elle stagne. La néguentropie, quant à elle ne fait que diminuer.

Dans l'ouvrage, il y a des explications plus détaillées sur la différence entre les deux que je choisis de passer car les détails sont un peu ardus et ne me paraissent pas indispensables. On y comprend que la théorie de l'information peut déboucher sur des choses bien plus vastes que la théorie de base, et que c'est sans doute cela qui lui a permis son si grand succès.

## 3.5 La théorie de l'information : Pour quoi faire ?

### 3.5.1 Une vocation d'origine toujours actuelle : La compression de données

L'objectif atteint en premier par la théorie de Shannon est le développement informatique des télécommunications, ce qui est logique au regard de l'époque, et de la formation d'ingénieur de Shannon. Or, le coût des communications est élevé, donc l'enjeu de réduire ce coût est important. Réduire le coût est directement relié aux quantités de données envoyées.

Mettons un alphabet $a, b$ et un canal par lequel on peut envoyer des signaux binaires $0, 1$. La première idée qui vient est de faire correspondre et $a$ et 0, et les $b$ avec les 1. Mais est-ce efficace ? Mettons que notre langage soit muni de la loi de probabilité suivante : $p(a) = 0,7$,
$p(b) = 0,3$
Pour ces données, l'entropie est de :
$H_{source} = 0,7.\log(\frac{1}{0,7}) + 0,3.\log(\frac{1}{0,3}) = 0,881 bits.$

On voit qu'il y a une grande différence avec la réalité si l'on garde l'idée de transmission intuitive alors qu'en réfléchissant en terme de probabilités :
$p(aa) = 0,49$,
$p(ab) = p(ba) = 0,21$,
$p(b) = 0,09$

Par exemple, le code suivant est nettement plus efficace :

- aa = 0
- ab = 11
- ba = 100
- bb = 101

Ce code est non ambigu, ce qui n'est pas le cas de tous les codes. En effet, si l'on lit une suite de caractères, on ne pourra pas confondre deux suites équivalentes (le morse est un code ambigu). Comment s'assurer que ce code est plus efficace que le premier ? C'est là qu'intervient la notion de *longueur*.

$L = \sum\_ip\_i.l\_i$

*L'efficacité* d'un code est le nombre de bits transmis par symboles de l'alphabet, soit :
$y = \frac{H}{L}$, où H est l'entropie, et L la longueur calculée précédemment. Le pire des cas est atteint dans le cas de l'équiprobabilité entre tous les symboles envoyés. Dans ce cas, $y = \log(n)$.

La redondance est simplement la différence entre la longueur maximale et la longueur réelle du code :
$r = y\_max - y$.

Cette notion, discutée précédemment dans ce texte, prend ici son sens.

Appliqué à l'exemple précédent, le premier code intuitif donnerait :
$L = 0,7.1 + 0,3.1 = 1$
L'efficacité est : $y = \frac{0,881}{1} = 0,881$, la redondance : $r = 1 - 0,881 = 0,119 bits$.

Le second code quant à lui :
$L = 0,49.1 + 0,21.2 + 0,21.3 + 0,09.3 = 1,81$ symboles binaires.
L'entropie : $H = 0,49.\log(\frac{1}{0,49}) + 2 \times 0,21.\log(\frac{1}{0,21}) + 0,09.\log(\frac{1}{0,09}) = 1,76 bits$ par message.
Efficacité : $y = \frac{1,76}{1,81} = 0,97 bits$.
Autrement dit, l'efficacité est augmentée, la redondance diminuée, ce qui rejoint le premier théorème de Shannon (*il est toujours possible de trouver un code voisin de l'efficacité maximale*). Des méthodes afin de trouver ces codes ont été proposées, ainsi **Huffman** et **Fano**, etc.
La problématique pour le transport de code est parfois différent : transmission d'image, vidéo, qui n'ont pas les mêmes prérequis, car en général, dans une vidéo, deux images sont très proches, donc on ara intérêt à décrire une image par rapport à celle qui la précède. Ici, la théorie de l'information a joué un rôle important, car elle permet de formaliser ces techniques. On aurait pu s'en passer, mais c'est mieux avec.

### 3.5.2 Une nouvelle approche possible de certains problèmes de logique

La théorie de l'information peut donner un éclairage nouveau à certains problèmes de logique, c'est ce qu'a essayé de montrer **Georges Cullman**, par exemple le problème des deux prisonniers.

### 3.5.3 Le problème du condamné à mort

Un condamné a deux portes pour s'échapper. L'une mène à la mort, l'une permet de s'échapper. Deux gardiens montent la garde et l'un dit toujours la vérité, l'autre ment toujours. L'analyse avec la théorie de l'information donne cela :

$H\_choix\_avant\_question = 0,5.\log(2) + 0,5.\log(2) = 1bit$ car le choix est au hasard.

Mais il souhaiterait que l'entropie du choix après question soit de 0, car il n'y aurait pas d'ambiguïté. Or, on a :

$H\_choix\_avant\_question = H\_quesiton + H\_choix\_après\_question.$

Or, si l'on remplace la lettre $H$ par le mot *incertitude*, cette équation se comprend intuitivement : on suppose que le prisonnier saura tirer tout l'enseignement possible de la réponse à la question. Donc $H(question) = 1$. La théorie ne peut pas nous aider à formuler une question pertinente. Mais on sait que la question "que me répondrait l'autre prisonnier ?" répond à ce critère.

### 3.5.4 L'avare

Un avare possède 26 pièces en or dont l'une est moins lourde que les autres. Pour la trouver, le premier réflexe intuitif est de les chercher partout. Mais

en fait, il y a une solution bien plus efficace qui consiste à diviser le tas en trois groupes : un de 8 et deux de 9. On compare les deux de 9. Celui le plus léger comporte la bille. Dans le cas de l'égalité, c'est l'autre tas restant. On procède toujours de la même façon, et il n'est plus nécessaire que de faire 3 coups pour trouver la pièce.

### 3.5.5 Un outil supplémentaire pour la statistique descriptive

G. A. Miller affirme que la quantité d'information est tout à fait similaire à ce qui est appelé *variance*. Les équations sont différentes mais les deux quantités évoluent de la même façon. *Toutes les formules classiques de l'analyse de la variance ont d'ailleurs leur homologue en analyse de l'incertitude* (citation originale). La différence réside dans le fait que la variance dépend d'unité de référence, là où la quantité d'information n'a pas de dimension. Elle est donc plus pratique dans certains cas.

### 3.5.6 Un instrument de mesure centrale en psychologie expérimentale

Ce n'est pas en statistique que la théorie de l'information s'est avérée la plus utile. Elle s'est répandue à toute vitesse dans les années 50 dans le domaine de la psychologie expérimentale, avant d'en disparaître totalement quelques années plus tard. La question a été : "Quelle quantité d'information un cerveau humain est-il capable de traiter" ? mais elle était trop complexe à aborder avec la théorie. Donc est venue une autre : "quelle quantité d'information un cerveau humain est-il capable de traiter en provenance de ses sens ?". Pour étudier cela, on a conduit beaucoup d'expériences afin de mesurer la capacité humaine à distinguer des catégories de couleurs, etc. Mais à un certain stade, les résultats se recoupaient avec ceux de la

mémoire de travail. La théorie de l'information a permis de procéder à toutes ces expérimentation et de trouver ces résultats, et on ne le cite que très rarement.

### 3.5.7 L'utilisation du concept de néguentropie pour caractériser la nature de la vie

La recherche en biologie a aussi utilisé la théorie de l'information. Cela prend peut-être son origine dans un ouvrage de **Jacques Monod** : *Le Hasard et la nécessité* où il parle beaucoup d'information et de bilan énergétique.

À ce moment, le principe de la vie apparaissait déjà en contradiction avec le deuxième principe de la thermodynamique, car la nature aurait dû tendre vers plus d'uniformité, or l'évolution du vivant fait émerger des structures de plus en plus complexes.

Or, la théorie de l'information offrait une autre voie : la spécificité du vivant pouvait consister à transformer l'énergie en information. (ceci est un peu développé dans le livre, pas ici). On ne peut pas répondre à la question "pourquoi le vivant créé-t-il de l'ordre" mais bien "comment", par contre.

### 3.5.8 Une méthode d'analyse utile pour la linguistique

L'entropie s'est avéré un outil utile pour l'étude des langues. Évidemment, les alphabets et autres éléments de la théorie s'y appliquent bien. Mais ce qui rend la théorie applicable, c'est en considérant les langues comme des systèmes "*ergodiques*" (il existe une certaine constante statistique dans ce système).

La théorie a ainsi pu permettre d'étudier le niveau de redondance des langues (l'anglais a une redondance d'environ 75%). Ainsi, une langue entropique est une langue qui dispose d'un vocabulaire riche, aux mots dif-

férenciés. Il devient possible de comparer des langues à ce niveau, à l'aide d'outils d'analyse.

### 3.5.9 De nouveaux concepts pour la réflexion sur le fait artistique

Il n'y a pas qu'avec le langage parlé que l'on communique. La création artistique est aussi considérée comme communication. Dès lors, on peut appliquer ces idées à de la peinture (étude des informations liées à la couleur et à ses changements), à la poésie (alexandrins vs prose), L'harmonie serait peut-être le juste équilibre entre une quantité suffisante d'information et de redondance, car si tout est information, cela n'est ni très lisible ni très agréable. S'ensuite un peu de dissertation sur l'histoire de l'art que je me permets d'écourter, dont la conclusion est en quelque sorte que la théorie de l'information peut apporter des outils d'analyse, mais en aucun cas d'évaluation).

Signalons tout de même qu'au delà des thèmes cités, la théorie de l'information noue également des liens avec la théorie des jeux, et la théorie générale des systèmes (et dans ce cas, le lien se fait par les instruments de mesure).

En bref, on voit bien que cette théorie a pu être utilisée dans des domaines très riches et variés, en vertu de quoi elle peut être placée parmi les théories les plus transdisciplinaires, et les plus débattues dans le monde scientifique.

## 3.6 Conclusion

La théorie de l'information constitue bien un événement central de la science du XXème siècle. Cinquante ans plus tard, on s'aperçoit qu'elle a autant déçu que fait rêver, car on l'a vu, les attentes furent nombreuses. Toutefois,

n'oublions pas que ce n'est en aucun cas de la responsabilité de son père fondateur, les opérateurs de la théorie ont servi des domaines tels que la psychologie ou la statistique, même si d'autres auraient pu être préférés, la théorie de l'information a contribué à la science du XXème. La question reste ouverte sur l'usage du terme "information". Est-ce à lui que l'on doit le succès de cette théorie qui finalement démocratise l'usage d'un opérateur connu auparavant : le logarithme comme maximum de vraisemblance ? Ce à quoi la théorie en elle-même ne peut bien entendu pas répondre.