Invitation à la théorie de l'information, Emmanuel Dion

Notes de lecture

Arabella Brayer

June 21, 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