

Lê Nguyễn Hoang - El Mahdi El Mhamdi

Le fabuleux chantier

Rendre l'**intelligence artificielle**
robustement bénéfique

Du même auteur :

Lê Nguyễn Hoàng, « La formule du savoir », EDP Sciences, juin 2018,
ISBN : 978-2-7598-2260-7

Imprimé en France

ISBN (papier) : 978-2-7598-2361-1 – ISBN (ebook) : 978-2-7598-2430-4

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 destinées 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, 2019

Table des matières

1	Introduction	9
	L'IA nous a envahis	9
	La première thèse du livre	10
	La deuxième thèse du livre	11
	La conclusion du livre	12
	Fantasmes et catastrophismes	13
	Point sémantique	15
	Bienveillance, nuances et réflexion	18
	Plan du livre	20
I	Rendre l'IA bénéfique est une urgence	23
2	L'IA est déjà partout	25
	Le mirage de l'IA	25
	Fiabilité	26
	Vérification	27
	Surveillance	28
	Automatisation	29
	Aide à la décision	30
	Personnalisation	31
	Analyse surhumaine	32
3	L'IA pose déjà problème	37
	Une ampleur planétaire	37
	L'attention est le nouveau pétrole	40
	Données personnelles	41
	Biais algorithmiques	42
	Polarisation idéologique	44
	Bouleversements sociaux	47
	La démocratisation de la cyber-guerre	49
	L'addiction	52
	La malinformation	53
	Les <i>mute news</i>	54
	L'infobésité	56
	Santé mentale	58

La viralité de la virulence	59
Une force invisible	61
Les victimes des IA	63
4 Une brève histoire de l'information	71
De l'importance de l'information	71
Matière, énergie... Information!	72
La flèche du temps	74
Une histoire informatique de la physique	75
La quantification de l'information	78
Une histoire informatique de la biologie	79
L'évolution des supports de l'information	80
Une histoire informatique de l'évolution culturelle	82
Le pouvoir de l'information	85
L'échelle logarithmique des temps	87
5 On n'arrête pas le progrès	91
Le temps de la légifération	91
Progrès stupéfiants	92
Le progrès pose problème	94
Intérêts économiques	94
Addiction des consommateurs	95
Urgence morale	95
Vers l'anticipation	98
L'hypothèse du monde vulnérable	100
Rien ne sert de traîner	101
6 Vers une IA de niveau humain ?	105
Une menace existentielle	105
Raisonnement probabiliste	106
Avis des experts	108
Sélection et réfutabilité	110
L'excès de confiance des experts	111
Hardware et software	112
Les performances sont imprévisibles	115
Le niveau humain : une fausse borne	118
II Rendre l'IA bénéfique est un défi monumental	125
7 Les contraintes sur les contraintes des IA	127
Être à la pointe	127
Course à l'IA	128
La nécessité de la maîtrise technique	128
Les solutions trop contraignantes	130
Concurrence	131

Monopole	133
Open source	136
Le fardeau moral	137
8 Peut-on contrôler les IA ?	141
Le bouton d'arrêt	141
L'interruptibilité	142
Boîte noire	143
Impossible à surveiller	146
Impossible à tester	147
Peut-on savoir si une IA est bénéfique ?	148
Quel humain en charge ?	150
L'expérience de pensée de la météorite	151
L'humain est une faille	151
Automatiser la sécurité	153
9 La programmation des IA	155
Le <i>machine learning</i> de Turing	155
Supervisé <i>versus</i> non supervisé	157
Apprentissage par renforcement	158
Incertitudes et facteurs d'escompte	160
Exploration <i>versus</i> exploitation	162
Exploration stratégique	164
AIXI	165
10 Le but des IA	169
Thèse de l'orthogonalité	169
Les effets secondaires de YouTube	170
Proxies	173
Hacker les récompenses	174
Objectifs instrumentaux	175
Convergence instrumentale	176
III Le fabuleux chantier pour rendre l'IA bénéfique	181
11 L'IA doit comprendre le monde	183
En quête de solutions robustes	183
La feuille de route	184
Le rôle des sciences	185
Collecte de données	186
Validité et stockage	187
Authentification et traçabilité	188
Confidentialité	189
Le bayésianisme	190
Approximations pragmatiques	191

Les représentations vectorielles	192
Modèle du monde	193
Attaques adversariales	194
Incertitude	197
12 Agréger des préférences incompatibles	201
On ne sera pas d'accord	201
Désaccords épistémiques et épistémologiques	202
Désaccords moraux	203
La théorie du choix social	206
Préférences cardinales	207
Wikipédia	209
<i>Moral machines</i>	210
Cède-t-on le pouvoir aux machines ?	211
Biais des données	212
La granularité des préférences	213
Apprendre les préférences humaines	214
13 Quelles valeurs pour les IA ?	217
L'argument de la Bugatti	217
Lunatiques et manipulables	219
Préférences orphelines	220
Progrès moral	221
Incertitude morale	222
Vers un moi^+	223
La volition	225
L'IA peut-elle apprendre nos moi^{++} ?	226
Pourra-t-on faire confiance à Charlie ?	227
14 Protéger le circuit de la récompense	231
Récapitulatif	231
Court-circuitage	232
Le court-circuitage est dangereux	232
Donner les bonnes incitations	233
Prendre soin du circuit de la récompense	234
PDG versus travailleur	235
Récompenser l'apprentissage	236
Expliquer les récompenses	237
Le contrôle d'Alice	239
Quel objectif pour Bob ?	240
15 Décentralisation et heuristiques	243
Robustesse	243
Ultra-rapidité	244
Les défis de l'algorithmique répartie	245
Le problème des généraux byzantins	246

Spécialisation	248
Heuristiques et ignorance	249
Récapitulatif global	250
IV Remarques et conclusions	253
16 Philosophie morale calculable	255
Vers une morale algorithmique	255
La thèse de Church-Turing	256
Le mot <i>conscience</i>	257
Les zombies philosophiques	259
Morale modèle-dépendante	261
Le réalisme moral	263
L'anti-réalisme moral	264
La complexité de la morale	265
Le temps de calcul de la morale	266
La philosophie avec une deadline	267
Vers une méta-éthique calculable	268
17 Vous pouvez aider	271
Sensibilisation	271
Respectabilité	273
Mieux débattre	274
Attirer toutes sortes de talents	276
Valoriser l'éthique et la sécurité	278
Aider les mouvements existants	280
Méditez, débitez et expliquez les thèses du livre	281
Joignez-vous au fabuleux chantier !	284

La science et la vie quotidienne ne peuvent pas et ne doivent pas être séparées.

Rosalind Franklin (1920-1958)

Le nouveau printemps de l'IA est le plus important développement algorithmique de ma vie. Chaque mois, il y a de nouvelles applications stupéfiantes et de nouvelles techniques transformatives. Mais de tels puissants outils s'accompagnent aussi de nouvelles questions et responsabilités.

Sergey Brin (1973-)

1

Introduction

L'IA nous a envahis

Jusqu'en 2012, il semble que de nombreuses personnalités académiques brillantes demeuraient extrêmement sceptiques au sujet de l'intelligence artificielle (IA) en général, et des réseaux de neurones¹ en particulier. C'est en tout cas ce que prétendit Geoffrey Hinton au moment de recevoir son prix Turing² 2019 pour ses travaux révolutionnaires dans ce domaine : « [Le fait que] des réseaux de neurones très grands qui partent avec des poids aléatoires et sans aucun savoir préprogrammé peuvent apprendre à réaliser de la traduction automatique, ça semblait être une théorie très, très idiote pour beaucoup de gens [...] Un des relecteurs affirma [autour de 2009] que les articles sur les réseaux de neurones n'avaient pas leur place dans une conférence en *machine learning*. »

Cependant, surtout depuis 2015, les succès spectaculaires des IA firent changer d'avis beaucoup de sceptiques. Même si la théorie sur laquelle ces technologies se fondaient ne semblait toujours pas si solide, les prouesses stupéfiantes dont ces IA étaient tout à coup capables avaient de quoi laisser pantois. Du jeu de go aux *deep fakes*, en passant par la reconnaissance vocale, la traduction automatisée et la génération de textes, les réseaux de neurones ont atteint des performances

1. Les réseaux de neurones désignent des techniques grossièrement inspirées de l'organisation des cellules d'un cerveau, elles sont aujourd'hui au cœur du succès des algorithmes regroupés sous le terme « *deep learning* ».

2. Le prix Turing est la plus importante distinction scientifique en informatique, souvent appelée « le Nobel de l'informatique ».

difficilement prévisibles en 2012. Visiblement, beaucoup de chercheurs avaient grandement sous-estimé l'IA.

C'était clairement notre cas. En fait, même après ces électrochocs, que ce soit en 2016, en 2017, en 2018, ou en 2019, la vitesse du progrès des IA n'a cessé de nous surprendre. Mais surtout, nous nous sommes vite sentis dépassés par la place grandissante que ces IA prenaient dans notre quotidien et dans nos sociétés. De la gestion de nos spams à l'auto-complétion de nos messages, en passant par les réponses à nos recherches Google, l'organisation de nos fils d'actualité Facebook et les recommandations de vidéos YouTube, l'IA semble nous avoir envahis.

La première thèse du livre

Étrangement, il nous aura fallu plusieurs années pour en arriver à cette conclusion. Mais après ces longues années d'enthousiasme et de réflexion, une observation devint de plus en plus pressante. Vu la place grandissante que prennent les IA, il paraît désormais urgent de s'assurer que ces IA soient programmées, non seulement pour ne pas être néfastes, mais aussi et surtout pour être *bénéfiques*.

Bien entendu, *toutes* les IA ne sont pas néfastes. Au contraire, la plupart des IA d'aujourd'hui semblent globalement bénéfiques. Cependant, il nous semble que *toutes* les IA influentes devraient être conçues avec l'objectif d'être au moins partiellement bénéfiques. Mais ce n'est pas tout. Il nous semble aussi que les IA influentes déjà bénéfiques devraient être conçues pour être nettement plus bénéfiques encore³, notamment parce qu'elles auraient ainsi probablement un impact bénéfique énorme sur le bien-être de milliards d'individus.

Telle est la première thèse de ce livre, exprimée ci-dessous de manière très approximative.

Thèse 1. *Rendre les IA bénéfiques est une urgence.*

Cette thèse peut sembler modeste. Nous l'avons formulée de sorte qu'elle en ait l'air. Cependant, cette apparence anecdotique est trompeuse. En particulier, ce que nous n'avons pas pris le temps de spécifier, c'est l'urgence relative de cette tâche, comparativement à l'urgence à résoudre d'autres problèmes, comme le racisme, l'extrême pauvreté ou le changement climatique. Une version plus radicale de cette thèse soutiendrait ainsi que rendre les IA bénéfiques est une urgence comparable, voire supérieure, à ces autres défis — notamment car nous prétendons qu'il s'agit de l'une des approches les plus prometteuses pour résoudre ces autres défis.

3. En particulier, nous insisterons plus tard sur l'importance d'être *robustement* bénéfique, c'est-à-dire demeurer très probablement bénéfique malgré l'imprévisibilité des *effets secondaires*, l'inévitabilité de *bugs* informatiques, l'existence de biais dans les données, les *hacks* d'utilisateurs malveillants, la modification de l'environnement, ou encore l'incertitude et la diversité des préférences morales.

Prise au sérieux, cette thèse semble alors devenir bien plus surprenante qu'elle n'en a l'air. Elle semble ainsi inviter à être critique de quiconque préférant volontairement ignorer les *effets secondaires* indésirables des IA, de la même manière que l'on en vient parfois à blâmer ceux qui choisissent volontairement d'ignorer les risques liés au changement climatique. Ou de façon équivalente, la thèse suggère que l'action la plus efficacement altruiste d'aujourd'hui pourrait peut-être être de chercher à contribuer au fabuleux chantier pour rendre les IA bénéfiques.

Les chapitres 2 à 6 de ce livre s'attarderont plus longuement sur cette thèse. Ces chapitres tenteront de vous convaincre du fait que vous sous-estimez très probablement très largement l'urgence que soutient la thèse. Voilà qui a des conséquences majeures sur ce qu'il nous faut exiger, par exemple, des entreprises du numérique.

La deuxième thèse du livre

Malheureusement, il serait probablement malencontreux de ne faire qu'exiger des entreprises et des gouvernements que leurs IA aient telle ou telle propriété. Ou de simplement protester contre le manque d'effort de leur part pour rendre les IA bénéfiques.

En effet, rendre les IA bénéfiques ne se réduit malheureusement pas à appuyer sur un bouton magique qui résoudrait tout à coup le problème. Au contraire, dans ce livre, on verra que l'urgence soulevée par la première thèse semble en fait correspondre à une tâche herculéenne. Telle est d'ailleurs la deuxième thèse de ce livre.

Thèse 2. *Rendre les IA bénéfiques est un défi monumental.*

Là encore, cette thèse a été formulée de manière relativement modeste. En particulier, cette formulation ne précise pas la difficulté de la tâche. Mais selon une version plus radicale de cette thèse, la difficulté de rendre les IA bénéfiques serait sans doute comparable, voire supérieure, à la difficulté de préserver la paix mondiale, résoudre l'hypothèse de Riemann ou concevoir une IA de niveau humain. Après tout, rendre les IA bénéfiques nécessite au moins de s'accorder sur une définition du mot « bénéfique ». Clairement, ceci n'est pas une maigre tâche.

La majeure partie de ce livre, à savoir les chapitres 7 à 15, défendra cette deuxième thèse. Nous verrons ainsi, encore et encore, que des approches naïves pour parvenir à nos fins semblent en fait vouées à l'échec. Ces chapitres tenteront ainsi de vous convaincre du fait que vous sous-estimez très probablement la difficulté de rendre les IA (robustement) bénéfiques. En fait, ces chapitres suggéreront que la tâche de rendre les IA bénéfiques ne pourra être résolue que si un très grand nombre de très grands talents divers et variés contribuent ensemble

à cet effort.

Dans les chapitres 11 à 15, nous présenterons une sorte de feuille de route pour bien penser le problème de rendre les IA bénéfiques. Même si cette proposition est très probablement très imparfaite, il nous semble qu'elle peut servir de base de travail⁴ et aider à mettre en évidence certaines étapes indispensables pour garantir la sûreté des IA. L'objectif de ces chapitres sera également de vous stimuler intellectuellement et de susciter chez vous une curiosité et un enthousiasme. En effet, nous espérons aussi vous convaincre que rendre les IA bénéfiques est aussi un *fabuleux* chantier. Voire, peut-être, le plus fabuleux des chantiers jamais entrepris par l'humanité.

La conclusion du livre


Le principal objectif du livre est d'en venir à la troisième thèse, qui est une conclusion à laquelle les deux premières thèses semblent conduire. Cette conclusion est particulièrement contre-intuitive. La voici.


Thèse 3. *Il est urgent que toutes sortes de talents soient mis dans les meilleures dispositions pour contribuer à rendre les IA bénéfiques.*

En particulier, une conséquence très étrange de cette conclusion, c'est qu'il peut sembler, à l'inverse, presque « immoral » pour tout talent de ne pas au moins s'intéresser un peu au problème de rendre les IA bénéfiques, tout comme il peut sembler « immoral » pour un politicien influent de ne pas au moins s'intéresser aux problèmes de racisme ou de pauvreté.

Sans aller jusque-là, nous chercherons à vous convaincre que l'aide de tout talent serait extrêmement précieuse. Malheureusement, de nos jours, les impacts sociaux des IA ne semblent pas être une question que la plupart des mathématiciens, philosophes, psychologues, sociologues, ingénieurs et dirigeants se posent régulièrement. Une contribution espérée de ce livre est donc d'inviter une fraction non-négligeable d'entre eux à davantage s'y intéresser. Mais les contributions directes pour rendre les IA bénéfiques ne sont pas les seules qui seront nécessaires pour mener ce chantier à bout. Ce fabuleux chantier nécessitera également de nombreuses contributions indirectes, par exemple en termes de sensibilisation aux défis à relever, de management des ressources humaines nécessaires au chantier ou de gouvernance entre les différentes entités influentes⁵.

4. Une autre feuille de route est publiée en même temps que ce livre par Stuart Russell. Bien que nous n'ayons pas la prétention de nous comparer à un pionnier comme Russell, et que nos feuilles de route aient beaucoup de similitudes, à commencer par le rôle central de l'alignement, la nôtre nous semble avoir des divergences qui gagneraient à être explorées.

 *Human Compatible : Artificial Intelligence and the Problem of Control* | Viking | Stuart Russell (2019)

5.  *Guide to working in AI policy and strategy* | 80,000 Hours | M Brundage (2017)

Pour l'instant, d'un point de vue technique, les efforts pour rendre les IA bénéfiques semblent essentiellement se restreindre à des points de suture pour colmater des hémorragies locales. Typiquement, certaines propositions s'intéressent uniquement à une poignée de propriétés désirables, notamment en termes de protection des données privées et de suppression des biais algorithmiques. Cependant, comme nous le verrons, pour de nombreuses IA qui influencent, volontairement ou non, les convictions de milliards d'utilisateurs, les solutions proposées jusque-là semblent encore très insuffisantes. En particulier, il semble que de telles IA ne peuvent pas se contenter d'être approximativement bénéfiques. Elles se doivent d'être *robustement* bénéfiques.

En particulier, il est important de noter que les IA qui interagissent avec des milliards d'utilisateurs, de créateurs de contenus et d'entreprises, évoluent dans des environnements extrêmement complexes, comme par exemple les réseaux sociaux. C'est aussi cette complexité de l'environnement qui rend la tâche d'être robustement bénéfique difficile. En effet, cet environnement est changeant et dynamique. Pire, il s'adaptera inéluctablement aux modifications des IA. Dans un tel contexte, parvenir à être constamment bénéfique selon le plus grand nombre, et surtout à ne jamais causer de torts majeurs, semble être une tâche monumentale.

Pour ces IA plus influentes, comme bien d'autres avant nous, nous prétendons que l'*alignement des valeurs*, aussi appelée *AI alignment* ou *value-loading* en anglais, est une étape probablement incontournable. Nous chercherons même à montrer qu'il s'agit là d'une condition nécessaire et suffisante pour garantir que les IA agiront de manière robustement bénéfique. Malheureusement, pour l'instant, trop peu de chercheurs semblent s'intéresser à ce problème pourtant crucial.

Néanmoins, il existe bel et bien déjà une littérature académique passionnante à ce sujet. En plus de sensibiliser à l'importance de l'*alignement*, nous espérons ainsi que la lecture de ce livre aidera tout curieux ou curieuse à découvrir les nombreux défis que pose la programmation d'IA robustement bénéfiques, voire à déterminer comment il ou elle peut contribuer au mieux à résoudre ces défis. En particulier, nous essaierons de montrer que le fabuleux chantier pour rendre les IA robustement bénéfiques est rempli de défis élégants, variés et multidisciplinaires. Voilà qui rend ces défis excitants et fascinants.

Fantasmes et catastrophismes

Cela fait maintenant quelques années que l'IA fait énormément parler d'elle. Tant de choses sont dites à son sujet. Chacun y va de son avis. Chacun y va de son qualificatif. Révolution, buzz, catastrophe, tendance passagère, sous-estimée, sur-estimée, dangereuse, hype, fantastique, inéluctable, lubie, imprévisible, magique, mécanique, disruption, risque existentiel⁶.

6. Un risque existentiel est un risque de destruction de toute l'humanité.

En particulier, ce dernier qualificatif est devenu un sujet de débat houleux, notamment suite à la publication du livre *Superintelligence* du philosophe Nick Bostrom en 2014. S'ensuivirent de nombreuses déclarations polémiques. Par exemple, Stephen Hawking, Elon Musk et Bill Gates partagèrent les craintes de Bostrom quant aux risques majeurs que pourrait poser une IA de niveau humain.

Cependant, le 20 septembre 2016, le site [Web TechnologyReview.com](http://WebTechnologyReview.com) du MIT publia une tribune d'Oren Etzioni, professeur d'informatique à l'université de Washington et PDG du Allen Institute for Artificial Intelligence. La tribune fut intitulée *Non, les experts ne pensent pas qu'une IA superintelligente est une menace pour l'humanité*. Le sous-titre ajouta : « demandez aux gens qui savent vraiment. »

Mais un mois plus tard, ce même site Web publia cette fois une tribune co-signée par Allan Dafoe et Stuart Russell, respectivement expert en gouvernance de l'IA et chercheur en IA, cette fois intitulée *Oui, nous sommes inquiets à propos du risque existentiel d'une intelligence artificielle*. Dafoe et Russell défendirent alors la thèse de Bostrom, en soulignant notamment que, même si le risque n'était pas imminent, il demeurerait néanmoins préoccupant.

Le plus étrange dans cette affaire, c'est que le désaccord entre Etzioni, Dafoe et Russell ne concerne même pas la dangerosité d'une IA. Il concerne les avis des experts sur la dangerosité des IA. Plus étonnant encore, Etzioni, Dafoe et Russell s'appuyaient bel et bien sur des sondages des avis des experts — le désaccord concernait en fait l'interprétation de ces sondages.

Mais le plus perturbant, c'est la posture agressive que même les experts adoptent pour débattre du futur de l'IA. Un chercheur anonyme sondé en 2016 et cité par Etzioni écrivait ainsi : « Nick Bostrom est un marchand professionnel de la peur. Le rôle de son institut⁷ est de trouver des menaces existentielles pour l'humanité. Il les voit partout. Je suis tenté de l'appeler le 'Donald Trump' de l'IA. »

Cet exemple n'est malheureusement qu'un exemple parmi tant d'autres. Les chercheurs en IA s'invectivent régulièrement sur les réseaux sociaux à ce sujet⁸. Il semble qu'il faille reconnaître d'importantes divergences concernant les avis des experts. La communauté des experts semble loin d'être une section militaire bien ordonnée chantant au pas⁹.

Ces divergences entre experts sont importantes à prendre en note. Elles nous invitent à plus de prudence quand il s'agit de parler d'IA. Il existe ainsi nécessairement beaucoup d'experts qui sont en excès de confiance. Prenons donc


7. Bostrom a fondé le *Future of Humanity Institute* à l'université d'Oxford.

8. Voir par exemple les commentaires à ce tweet d'OpenAI :

<https://twitter.com/OpenAI/status/1096092704709070851>

Ou cette discussion entre prix Turing sur Facebook :

<https://www.facebook.com/yann.lecun/posts/10156111192797143>

9.  *IA : Clash de prix Turing* | Alexandre Technoprog (2019)

soin du nôtre ! En particulier, il serait malencontreux d'isoler un expert en particulier, et de penser que l'avis de cet expert est le « bon » avis à avoir — et il serait encore plus problématique de considérer que notre avis est clairement plus pertinent que l'avis de cet expert. Quand il s'agit du futur de l'IA, le langage des probabilités et de l'incertitude semble incontournable.

Mais surtout, ces incompréhensions entre experts du domaine montrent à quel point les risques de contre-sens sont énormes. En particulier, l'intelligence n'est clairement pas suffisante pour bien analyser les idées de ce livre. Quand il s'agit d'IA, il semble crucial de surveiller nos nombreux biais cognitifs, à commencer par l'*excès de confiance* et le *biais de confirmation*. Pour éviter les malentendus et les mécompréhensions, il semble désirable de faire des efforts particuliers d'écoute, de bienveillance et de réflexion, surtout dans des situations de débat.

Point sémantique


L'une des raisons de ces nombreuses divergences est l'ambiguïté des terminologies utilisées, y compris par les experts. Il y a ainsi beaucoup de confusions dues au simple fait que même les experts n'assignent pas le même sens aux mêmes mots. Voire qu'un même expert utilise parfois un même mot dans des sens différents.

On peut prendre l'exemple de la notion de conscience pour illustrer cela. On a ainsi souvent tendance à confondre différentes notions pourtant assez distinctes de conscience. Il y a par exemple la *conscience d'accès*, c'est-à-dire la faculté d'une intelligence à accéder à sa propre réflexion. Cette notion algorithmique est relativement simple à implémenter¹⁰. En fait, on peut estimer que de nombreuses IA d'aujourd'hui ont déjà une telle forme de conscience.

Cependant, la conscience qui fascine davantage les philosophes est en fait autre. Il s'agit de l'expérience subjective vécue par une entité donnée. On parle aussi de *qualia* ou de *conscience phénoménale*. Certains philosophes affirment alors que cette conscience phénoménale sort nécessairement du cadre physique, et qu'il s'agit d'une propriété fondamentalement inobservable. Malheureusement, la confusion entre ces deux notions conduit souvent à davantage de confusion encore¹¹. Pire, on les confond avec d'autres usages encore du mot « conscience », comme la conscience morale qui suggère une faculté à comprendre et à se conformer à une morale. Nous reviendrons sur ces difficultés en fin de livre.

Il semble que ce qui rend ce terme particulièrement problématique, c'est la très forte connotation qui lui est associée. Intuitivement, on est tentés de dire

10. Par exemple, un compteur dans une boucle de calcul permet à un algorithme d'accéder à un aspect de sa réflexion, à savoir combien de fois il a effectué cette boucle de calcul. Plus généralement, et contrairement d'ailleurs à l'humain pour l'instant, on peut permettre à un algorithme d'accéder au code qu'il exécute.

11.  [La conscience \(avec Monsieur Phi\)](#) | Science Étonnante | T Giraud & D Louapre (2017)

que la conscience est une propriété fondamentalement désirable. Voilà qui nous pousse, consciemment ou non, à définir la conscience de sorte que cette notion s'applique à ce que l'on considère désirable ou meilleur. Tel est le biais du *raisonnement motivé*, maintes fois mis en évidence par la psychologie empirique¹². Cette remarque s'applique d'ailleurs aussi aux concepts d'*intelligence* ou de *morale*. Malheureusement, cette malléabilité de la sémantique conduit trop souvent à des débats interminables sur des problématiques qui, comparées à l'urgence de rendre les IA bénéfiques, nous semblent secondaires.

Ce livre n'a pas vocation à trancher sur les définitions des mots « intelligence », « conscience » et « morale ». D'ailleurs, ces mots n'apparaîtront quasiment pas dans ce livre. Néanmoins, puisque nous craignons qu'on pourrait nous reprocher de ne pas définir au moins le mot « intelligence », c'est avec une certaine réticence que nous proposons de le faire ici. Nous avons tenté d'opter pour une définition parmi les moins polémiques — malheureusement, toute définition semble polémique.


Intelligence : Capacité à atteindre des objectifs¹³.

Cette définition est très inclusive. Elle s'applique à toute entité qui reçoit, traite, stocke et émet de l'information (par exemple en adoptant un comportement). Selon cette définition, une fourmi a de l'intelligence dans le sens où elle est capable d'atteindre l'objectif « *trouver des graines et les ramener à la colonie* ». Un ordinateur qui joue aux échecs est intelligent dans la mesure où il atteint l'objectif « *battre le champion du monde aux échecs* ». Même une plante est intelligente dans le cadre de cette définition, dans la mesure où elle est capable de répondre à des stimuli environnementaux en changeant sa morphologie, de telle sorte qu'elle puisse capter plus de lumière et atteindre ainsi un objectif.

Néanmoins, selon cette définition, toutes ces entités n'ont pas les mêmes degrés d'intelligence. En s'appuyant sur la cette définition, on peut en effet comparer les intelligences en fonction de leur efficacité et du nombre d'objectifs qu'elles arrivent à atteindre. Ainsi, il semble raisonnable d'affirmer qu'un humain est globalement largement plus intelligent qu'un rat, dans la mesure où il arrive à atteindre un spectre d'objectifs plus large que ceux du rat.

Nous reviendrons davantage sur la capacité des machines à atteindre des objectifs au moment d'aborder la notion d'IA de niveau humain dans le chapitre 6. Nous discuterons même d'une définition formelle de l'intelligence, appelée *intelligence de Legg-Hutter*, au moment de parler des algorithmes d'apprentissage par renforcement dans le chapitre 9. Cependant, nous insistons encore une fois sur le fait que définir le mot « intelligence » n'est pas l'objet de ce livre.

Il y a un autre concept si central aux discussions de ce livre que nous ne pourrions pas non plus faire l'impasse sur sa définition, à savoir le concept d'*IA*.

12.  *Système 1 / Système 2 : Les deux vitesses de la pensée* | Flammarion | D Kahneman (2012)

13. Nous reviendrons longuement sur cette notion d'*objectif* dans le chapitre 10, où l'on verra notamment que cet objectif n'a rien « d'objectif ».

Malheureusement, là encore, aucune définition ne semble capable de satisfaire tous les experts. Dans le cadre de ce livre, nous avons fait le choix d'adopter une définition très englobante, pour désigner en fait l'ensemble des outils du numérique.

IA : Outil de traitement automatique de l'information¹⁴, généralement doté d'un objectif.


Un intérêt de cette définition est qu'elle permet de démystifier la notion d'IA. Une IA n'a pas à être spectaculaire ou superintelligente pour satisfaire notre définition. Elle n'a qu'à collecter, stocker, traiter et émettre de l'information. Selon notre définition, un thermostat par exemple est une IA : c'est un outil qui *traite de l'information* (température) de *manière automatisée* afin de *réaliser un objectif* (maintenir la chambre à une température désirée).


En fait, notre définition n'impose ni un fonctionnement biologique, ni un fonctionnement électronique des IA. Certaines IA ne sont en fait qu'un tas d'objets inertes¹⁵, à l'instar de la machine Menace qui n'est qu'un tas de boîtes d'allumettes et de billes¹⁶. Mieux encore, selon notre définition, certaines IA performantes sont des agrégats de composants biologiques, électroniques et matériels, comme les entreprises, les gouvernements et les économies mondialisées¹⁷. La NASA est ainsi probablement actuellement l'IA la plus performante quand il s'agit d'envoyer des hommes sur la Lune¹⁸. De façon cruciale, toutes ces organisations reçoivent, stockent, traitent et émettent des informations. Leur fonction est en grande partie, sinon exclusivement, du *traitement automatique de l'information*. C'est ce traitement de l'information qui nous intéresse.

En particulier, si nous insistons tant sur le traitement de l'information, c'est pour couper court aux querelles sémantiques inutiles qui polluent une grande partie des débats sur l'IA. Avec notre définition, comme nous le verrons dans le prochain chapitre, nous sommes clairement déjà envahis par les IA — en particulier par les IA électroniques ! En fait, à chaque fois que vous lisez « IA », nous vous invitons à penser à l'algorithme de recommandation de vidéos de YouTube. C'est typiquement cette IA qu'il nous semble urgent de rendre *robustement bénéfique*.

Insistons encore dessus. Les querelles sémantiques ne sont pas l'objet de ce livre. Notre objectif, c'est avant tout de défendre les trois thèses énoncées plus haut. Autrement dit, nous vous supplions de prêter uniquement attention à l'urgence de rendre les *outils de traitement automatique de l'information* bénéfiques, et de réfléchir aux défis qu'il nous faut relever pour y arriver. Ce sont de ces problèmes

14. Autrement dit, selon notre définition, le mot « IA » est synonyme du mot « algorithme ».

15.  *The Game That Learns* | Vsauce2 | K Lieber (2019)

16.  *MENACE : the pile of matchboxes which can learn* | standupmaths | M Parker (2018)

17.  *Le paradoxe de la veste de laine* | Monsieur Phi | T Giraud (2016)

18.  *Conférence sur la SUPER-INTELLIGENCE + quelques suppléments* | Monsieur Phi | T Giraud (2018)

que nous souhaitons parler.

En particulier, il est bon de garder en tête que les objets d'étude de ce livre, ces « IA », n'ont absolument pas à être « intelligentes » pour garantir la validité des arguments de ce livre. En fait, aucun des arguments de ce livre ne devrait perdre de sa validité si vous remplacez systématiquement la terminologie « IA » par « Information Automatiquement traitée », « Infatigable Algorithme » ou « Instrument Arithmétique ». À chaque fois que vous serez gênés par notre utilisation de cette terminologie, nous vous invitons d'ailleurs vivement à faire cet exercice de substitution dans votre tête. Dans le cadre de ce livre, *il n'y a nul besoin de supposer que les IA sont « intelligentes »*. Encore moins qu'elles sont « conscientes ». Au risque de plagier Laplace, nous n'aurons pas besoin de ces hypothèses¹⁹.

Bienveillance, nuances et réflexion

Pour clarifier, le problème souligné par ce livre n'est pas le risque d'IA « consciemment malveillantes ». Ce risque nous semble en fait négligeable²⁰. De façon plus générale, notre préoccupation principale ne sera ni la motivation des IA, ni la motivation des développeurs des IA. Le problème soulevé dans ce livre est celui des *effets secondaires* des IA. En particulier, nous chercherons à montrer qu'une IA influente qui n'est pas conçue pour être *robustement bénéfique* aura certainement des *effets secondaires* difficilement prévisibles et potentiellement très indésirables. Comme nous le verrons, c'est via de tels *effets secondaires* que l'IA tue déjà.

En fait, même si l'urgence à rendre les IA bénéfiques est une préoccupation importante de ce livre, nous insisterons beaucoup plus encore sur la difficulté d'y arriver. En particulier, nous cherchons avant tout à défendre la thèse 3 : *il est urgent que toutes sortes de talents soient mis dans les meilleures conditions pour contribuer à rendre les outils de traitement de l'information bénéfiques*. C'est de cela que nous souhaitons vous convaincre.

Malheureusement, les défis à relever pour rendre les IA bénéfiques sont horriblement complexes et pleins de subtilités et de nuances. Pire, la réflexion poussée autour de ces défis monumentaux conduit souvent à des conclusions très contre-intuitives. Sorties de leur contexte, il peut être horriblement tentant de rejeter, voire de moquer, ces conclusions. Pour éviter ce travers hautement probable, bienveillance, nuances et réflexion semblent être les maîtres mots.

Ainsi, dans ce livre, nous avons fait un énorme effort pour aller dans ce sens. Cependant, nous craignons que nos bonnes intentions aient été très insuffisantes.

19. Selon la légende, après avoir lu l'*Exposition du Système du monde*, le général Bonaparte questionna l'absence de Dieu dans ce livre de Laplace. Laplace aurait répondu : « je n'ai pas eu besoin de cette hypothèse ».

20. Nous reconnaissons toutefois le fait que nous nous trompons peut-être sur ce point. Les armes autonomes pourraient être des IA conçues pour être, en un sens, « consciemment malveillantes ».

De façon ironique, nos discussions à venir sur les *effets secondaires* indésirables des IA auront très certainement elles-mêmes des *effets secondaires* indésirables. Nous en sommes vraiment désolés. Exposer les idées de ce livre avec pédagogie et clarté fut une tâche monumentale elle aussi. Nous sommes conscients de ne l'avoir résolue que bien trop partiellement.

Pour éviter des contre-sens malheureusement hautement probables, y compris chez les experts en IA, nous vous encourageons, cher lecteur ou lectrice, à corriger vous-même les nombreuses notions de ce livre, avec calme, rigueur et ouverture d'esprit. Nous vous invitons à vous saisir du sujet de ce livre, à critiquer ce qui y est écrit, mais aussi à vous exercer à défendre au mieux les thèses qui y sont présentées. De plus, nous vous suggérons d'accueillir avec bienveillance les mécompréhensions des autres (y compris les nôtres !) et d'essayer d'être pédagogique dans l'aide que vous fournirez pour clarifier les notions de ce livre. Quand quelqu'un vous demandera de commenter ce livre, que vous soyez critique ou non, nous vous supplions de ne pas être caricatural.

Pour méditer au mieux les idées de ce livre, et notamment éviter le *biais de confirmation*, il pourrait être une bonne idée d'organiser ou de participer à des groupes de lecture. Par exemple, il serait peut-être très instructif de prendre part, chaque semaine, à une analyse collective d'un chapitre de ce livre²¹. De telles rencontres pourraient ainsi permettre d'envisager des perspectives différentes sur les enjeux de l'IA. Elles pourraient aussi plus simplement aider à entretenir la motivation à réfléchir activement au fabuleux chantier pour rendre les IA bénéfiques²². Si jamais de telles rencontres vous intéressent, et si vous cherchez d'autres lecteurs de ce livre avec qui échanger, nous vous suggérons, par exemple, de contacter l'association *Altruisme Efficace France*²³, pour trouver de potentiels compagnons de lecture. Ou, bien entendu, tout autre cadre en dehors de cette association qui vous paraîtrait approprié.

En particulier, nous espérons que c'est avant tout cette invitation à la bienveillance, à la nuance et à la réflexion qui émergera des discussions autour de ce livre. Et si vous pensez qu'il y a des aspects importants du fabuleux chantier qui ont été omis dans ce livre, nous vous serons très reconnaissants de nous les signaler, si possible avec pédagogie, clarté et bienveillance. Comme on essaiera de vous en convaincre, le jeu semble largement en valoir la chandelle.

21. Prenez garde toutefois à éviter les phénomènes bien connus de *polarisation de groupe*, en apportant régulièrement des contrepoints à l'avis du groupe, surtout si celui-ci semble consensuel.

22.  *Curiosité préoccupée avec Jérémy Perret | Probablement ? | J Perret & LN Hoang (2019)*

23. En fonction de votre localisation géographique, nous vous invitons à contacter *Altruisme Efficace Québec*, *Effective Altruism Geneva* ou autres, voire à monter votre collectif local, en vous coordonnant si possible avec *Altruisme Efficace France* ou encore le *Center for Effective Altruism*.

Plan du livre

Le reste du livre se décompose comme suit. Dans un premier temps, des chapitres 2 à 6, nous insisterons sur la première thèse du livre, à savoir l'urgence à rendre les IA bénéfiques. Le chapitre 2 insistera sur l'omniprésence et la place déjà prépondérante qu'ont les IA d'aujourd'hui, en cherchant au passage à expliquer ce rôle que les IA ont pris. Le chapitre 3 cherchera à montrer que les IA n'ont pas un rôle innocent. Au contraire, la place prépondérante qu'elles ont prise signifie que les actions entreprises par ces IA ont des *effets secondaires* d'ampleur planétaire. Le chapitre 4 prendra du recul et analysera le rôle central de l'information et du traitement de l'information dans l'histoire de la vie et des civilisations. Les chapitres 5 et 6, eux, insisteront sur l'importance d'anticiper le futur, et chercheront à montrer que le progrès des performances des IA est à la fois inévitable et très imprévisible. Voilà qui rend le problème de rendre les IA bénéfiques d'autant plus urgent.

La deuxième partie du livre, elle, s'intéressera à la deuxième thèse de ce livre, à savoir la difficulté à rendre l'IA bénéfique. Cette seconde partie s'étendra des chapitres 7 à 10. Dans un premier temps, les chapitres 7 et 8 montreront que des idées naïves qui consisteraient à contraindre ou contrôler les IA semblent en fait très peu prometteuses. Puis, les chapitres 9 et 10 introduiront une compréhension conceptuelle des algorithmes des IA du présent et de l'architecture probable des IA du futur, qui semble indispensable à maîtriser pour rendre les IA bénéfiques.

Puis la troisième partie du livre, des chapitres 11 à 15, proposera une esquisse de *feuille de route* pour mieux structurer la réflexion autour des solutions techniques pour rendre les IA bénéfiques. Cette feuille de route vise à découper le fabuleux chantier pour rendre les IA bénéfiques en un très grand nombre de sous-problèmes plus simples, de la fiabilité des données à l'inférence de l'état du monde à partir de ces données, en passant par l'alignement des objectifs des IA et la conception d'un système de récompense adéquat pour ces IA.

Les deux derniers chapitres seront quelque peu à part. Le chapitre 16 sera une digression sur les implications de l'approche algorithmique de ce livre sur la philosophie morale. Nous chercherons notamment à montrer que certaines notions semblent en fait incalculables, ce qui suggère qu'il pourrait s'agir de distractions qu'il serait alors souhaitable de moins mettre en avant. Nous insisterons aussi sur la pertinence de la théorie de la complexité algorithmique pour la philosophie morale, ainsi que sur l'importance de la méta-éthique algorithmique.


Enfin, le chapitre 17 évoquera les nombreux défis non techniques qui sont indispensables à relever pour mettre toutes sortes de talents dans les meilleures prédispositions pour réfléchir au mieux au vaste et fabuleux défi de rendre les IA bénéfiques.


*Practical AI podcast*⁶² de Changelog ou le *AI alignment podcast*⁶³ du Future of Life Institute.


Dans ses activités d'enseignement, il a construit et enseigné un cours sur le machine learning donné au doctorants de l'université Mohammed VI fraîchement créée au Maroc. Il est aussi chargé de TD de plusieurs cours de niveau Master sur les algorithmes, les systèmes distribués et le machine learning à l'EPFL en Suisse.

Il a d'abord évolué dans la physique de la matière condensée. Son travail sur la robustesse des matériaux en silicium amorphe⁶⁴ est apparu dans le journal *Applied Physics Letters*, considéré comme l'un des deux plus importants journaux revus par les pairs en physique appliquée, en matière condensée et en physique des semi-conducteurs. Durant ce bref début de carrière en physique, son intérêt pour le Web et la dissémination d'information l'a poussé à co-fonder *Mamfakinch*, un média marocain ayant été primé par le *Breaking Borders Award*⁶⁵, décerné par Google et Global Voices en 2012.


Après l'expérience *Mamfakinch*, il quitte son travail d'ingénieur en physique pour se dédier aux projets de pédagogie et de dissémination d'information sur le web. Convaincu par cette expérience que le format vidéo allait l'emporter sur le format texte, il décide d'expérimenter le tutorat scientifique sous format vidéo et lance *Wandida*, une chaîne YouTube qui propose des explications concises de concepts scientifiques de niveau universitaire. Wandida convainc Google qui finance son lancement puis l'EPFL qui finance sa durabilité et l'incorpore à son offre éducative en ligne. Durant la période Wandida / Mamfakinch, El Mahdi El Mhamdi a été convié à de nombreux événements sur la portée éducative ou journalistique du Web. Il a présenté ses méthodes à la conférence annuelle de l'*Association for Learning Technologies (ALT)* à l'université de Manchester au Royaume-Uni, à la conférence *e-Learning Africa* en Ouganda puis en Éthiopie. Il a aussi été coach et membre du jury du Hackathon sur l'audiovisuel et l'éducation *Hack'Xplor*, tenu à Liège en Belgique puis au congrès de l'Organisation internationale de la francophonie tenu en 2014 à Dakar au Sénégal. En plus de son rôle à Mamfakinch, il a aussi rédigé des articles pour d'autres médias comme *Médias24.com*, le principal média économique marocain, *Future-Challenges.Org* de la *Bertelsman Foundation* en Allemagne, ou encore le média français *Le Monde*⁶⁶.

62.  *Staving off disaster through AI safety research* | Practical AI | EM El Mhamdi & C Benson (2019)

63.  *AI Alignment Podcast : The Byzantine Generals' Problem, Poisoning, and Distributed Machine Learning with El Mahdi El Mhamdi (Beneficial AGI 2019)* | FLI Podcast | EM El Mhamdi & L Perry (2019)


64.  *Is light-induced degradation of a-Si:H/c-Si interfaces reversible ?* | Applied Physics Letters | EM El Mhamdi et al. (2014)

65.  *Breaking border for free expression* | Google Public Policy | Bob Boorstin (2012)

66.  *Quel avenir pour la francophonie numérique ?* | LeMonde.fr | El Mahdi El Mhamdi (2015)

L'expérience Web, notamment celle de Wandida, fut une période d'interaction privilégiée avec les chercheurs en informatique et en intelligence artificielle. Cette interaction lui a notamment permis de se rendre compte qu'au delà des chamboullements sociaux, l'informatique recèle surtout des questions scientifiques et épistémologiques fondamentales. Il décide alors de revenir à la recherche. Fin 2015, l'EPFL accepte de financer un poste pour que Wandida (et les initiatives pédagogiques du même genre) puissent être poursuivies, ce qui lui permet de se libérer pour entamer sa thèse de doctorat.

Convaincu de la portée épistémologie de l'informatique, El Mahdi El Mhamdi veut œuvrer pour que l'apport de l'informatique dépasse celui de la technologie. Dans ce sens, il a co-écrit un travail épistémologique⁶⁷ avec le sociologue Dominique Boullier de Science Po. Ils y expliquent comment la science algorithmique, au delà des outils technologiques qu'elle offre aux autres sciences, notamment les sciences sociales, recèle des outils conceptuels encore peu connus et exploités par les autres disciplines, à savoir la théorie de la complexité algorithmique et la théorie de l'apprentissage. Dans le même esprit, il a été invité à donner une conférence en session plénière lors de la cinquième rencontre internationale des psychologues et psychiatres francophones experts en TDAH titrée « *Ce que l'intelligence artificielle doit aux sciences cognitives et ce qu'elle peut leur rendre* ». Le but de cette intervention (et d'autres du même type) est de sensibiliser l'audience au fait que, au-delà des « gadgets » et des « logiciels » d'aide à la décision médicale qui semblent intéresser les praticiens de la santé, l'IA est surtout une occasion de faire progresser la réflexion conceptuelle et épistémologique sur des questions comme « qu'est-ce que "réfléchir" ? », « qu'est-ce que "apprendre" ? » et comment, en revenant aux origines des méthodes actuelles en IA, qui trouvent leur sources en sciences cognitives et non en informatique traditionnelle, on pourrait amorcer de meilleures discussions entre informaticiens et experts du cerveau.

67.  *Des modèles aux pratiques : le machine learning à l'épreuve des échelles de complexité algorithmique* | Revue d'anthropologie des connaissances | D Boullier & E.M. El Mhamdi (2019)