

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 <i>moi</i> ⁺	223
La volition	225
L'IA peut-elle apprendre nos <i>moi</i> ⁺⁺ ?	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