Habilitation à diriger les recherches

Université Paris VII

par

Alexei Grinbaum

Volume I. Philosophie de la physique.

L'OBSERVATEUR DANS LA THÉORIE QUANTIQUE

Volume II. Éthique des sciences.

TECHNOLOGIES NOUVELLES ET RÉCITS ANCIENS

Soutenue le 28 novembre 2018 devant le jury composé de :

M. Jeffrey Bub	*Rapporteur*	Professeur à l'Université du Maryland
M. Raja Chatila	*Rapporteur*	Professeur à Sorbonne Université
M. Gilles Dowek	*Rapporteur*	Directeur de recherche à l'INRIA
M. Jean-Jacques Szczeciniarz	*Rapporteur*	Professeur à l'Université Paris VII
M. Michel Bitbol	*Examinateur*	Directeur de recherche au CNRS
M. Olivier Darrigol	*Examinateur*	Directeur de recherche au CNRS
M. Nicolas Gisin	*Examinateur*	Professeur à l'Université de Genève
Mme Michèle Leduc	*Examinatrice*	Directeur de recherche émerite au CNRS

Table des matières

Chapitre 1

Curriculum vitæ

Expérience professionelle

2006–	Chercheur au laboratoire Larsim du CEA-Saclay
2007–	Enseignant-vacataire à l'ENSTA, CentraleSupélec, Université d'Évry, École du Val-de-Grâce, Institut Pasteur, INRIA, INSTN, Institut Gustave Roussy, Université Paris-Saclay. Cours invités à l'École Polytechnique, Sciences Po et l'Université européenne de Saint-Pétersbourg.
2007-2009	Directeur adjoint du Groupe de recherche et d'intervention sur la science et l'éthique (GRISE), École Polytechnique
2006	Postdoc au Perimeter Institute of Theoretical Physics (Canada)
2005	Postdoc aux Archives Henri Poincaré (CNRS, Nancy)

Éducation

2004	Doctorat en philosophie des sciences (École Polytechnique)
2003	M.Sc. de physique théorique (Université d'État de Saint-Pétersbourg)
2001	DEA de Sciences cognitives (École Polytechnique)
2000	Programme international de l'École Polytechnique
1999	B.Sc. de physique (Université d'État de Saint-Pétersbourg)

— Chercheur invité à l'Institute for Quantum Optics and Quantum Information (Vienne, Autriche), à l'Université de Genève (Suisse) et à l'Université de Pavie (Italie).
— Membre de la CERNA (Commission d'éthique de la recherche en sciences et technologies du Numérique d'Allistene), 2012-2018.
— Membre du Groupe de travail sur les impacts économiques et sociaux de l'intelligence artificielle du Gouvernement français (2017).

— Membre du General Principles Committee of the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (depuis 2016).
— Expert de la Commission Européenne pour l'évaluation éthique des projets de recherche du programme H2020 (depuis 2016).
— Membre du Groupe de travail éthique, juridique, normalisation et réglementaire, France Robot Initiative, 2014-2015.
— Member of Advisory Board of research project "Russian Computer Scientists at home and abroad" (STS Center of the European University in St Petersburg), 2014.
— Membre du Groupe de travail « L'impact de la technologie sur la vie des hommes », Centre d'analyse stratégique, Secrétariat d'État chargé de la Prospective, 2008.

Projets de recherche

— EC H2020 research project RRI-Practice : Responsible Research and Innovation in Practice (2016-2019)
— EC FP7 research project observatoryNano : European observatory for science-based and economic expert analysis of nanotechnologies (2008-2012)
— EC FP7 research project NanoCode : European Code of Conduct for Nanosciences and Nanotechnologies Research (2010-2011)
— ANR research project « FoundPhys : Fondements de la physique » (2006-2010)
— Member of ESF Research Network for Philosophical and Foundational Problems of Modern Physics (2002-2005)

Referee pour Nature, British Journal for the Philosophy of Science, Studies in the History and Philosophy of Modern Physics, Foundations of Physics, Foundations of Science, Proceedings of Royal Society A, Annales Henri Poincaré, Philosophia Scientiæ, Comptes Rendus de l'Académie des Sciences – Geosciences, NanoEthics, Science and Engineering Ethics, Mind and Matter, Minds and Machines, Revue d'histoire des sciences, Axioms, PLOS, Journal of Responsible Innovation, Entropy, European Journal for Philosophy of Science, Life Sciences, Society and Policy, Oxford University Press, Routledge, Editions du Seuil, Flammarion.

Près de 250 conférences données depuis 2003.

Chapitre 2

Liste des publications

Les publications qui apparaissent en gras sont incluses dans les annexes.

2.1 Livres

« **Les robots et le mal** » **(Desclée de Brouwer, à paraître en 2019)**.

« Mécanique des étreintes : intrication quantique » (Encre Marine, 2014).

2.2 Publications en philosophie de la physique

1. **"The Effectiveness of Mathematics in Physics of the Unknown," Synthese, 2017. DOI 10.1007/ s11229-017-1490-0. phil-sci/12950**
2. **"Narratives of Quantum Theory in the Age of Quantum Technologies," Ethics and Information Technology, 19, 295–306, 2017. arxiv:1702.03001.**
3. **"How device-independent approaches change the meaning of Physics", Studies in the History and Philosophy of Modern Physics, 58, 22-30, 2017. arXiv: 1512.01035**
4. "Quantum correlations: Challenging the Tsirelson bound", in: Quantum Interaction, eds. H. Atmanspacher, Th. Filk and E. Pothos, Springer, 2016, pp. 3-11.
5. **"Quantum theory as a critical regime of language dynamics", Foundations of Physics 45, 1341-1350, 2015.**
6. "Information-theoretic constraints on correlations with indefinite causal order", Phys. Rev. A 92, 042124, 2015 (avec I. Ibnouhsein).
7. "Renormalized entropy of entanglement in relativistic field theory", Phys. Rev. D 90, 065032, 2014 (avec I. Ibnouhsein et F. Costa).
8. "Quantum observer, Information Theory and Kolmogorov Complexity," in New Challenges to Philosophy of Science, eds. H. Andersen, D. Dieks, W. J. Gonzalez, T. Uebel and G. Wheeler, Springer, 2013, pp. 59-72.
9. **"Quantum Observer and Kolmogorov Complexity", in A. Wüthrich and T. Sauer, eds., New Vistas on Old Problems: Recent Approaches to the Foundations of Quantum Mechanics, Edition Open Access, Max Planck Research Library for the History and Development of Knowledge Proceedings vol. 3, 2013, pp. 13-34.**

10. "Twin quantum Cheshire photons" (avec I. Ibnouhsein), arXiv:1202.4894.
11. "Quantum observer and Kolmogorov complexity: a model that can be tested", 2011. Older version: "A Mathematical Criterion of 'Element of Reality'", arXiv:1007.2756.
12. **"Which fine-tuning arguments are fine?" Foundations of physics, 42, 2012, pp. 615-631, arXiv:0903.4055.**
13. **"On Epistemological Modesty", Philosophica, 83, 2010, pp. 139-150.**
14. "On the eve of the LHC: conceptual questions in high-energy physics", CEA report, arXiv: 0806.4268, PhilSci/4088.
15. "Reconstructing instead of interpreting quantum theory", Philosophy of Science, 74, 2007, pp. 761–774. Also quant-ph/0509104.
16. **"Reconstruction of quantum theory," British Journal for the Philosophy of Science, 58, 2007, pp. 387–408.**
17. **"Information-theoretic principle entails orthomodularity of a lattice," Foundations of Physics Letters 18 (6), 2005, pp. 563–572.**
18. "Elements of information-theoretic derivation of the formalism of quantum theory", International Journal of Quantum Information 1(3), 2003, pp. 289–300.
19. "Information-theoretic derivation of the formalism of quantum theory", in: A. Khrennikov (ed.), Proceedings of International Conference "Quantum theory: Reconsideration of Foundations–2", Växjo University Press, Växjo, Sweden, 2004, pp. 205-217.
20. "O filosofii fiziki" (On the philosophy of physics), Zvezda, 10, 2003, pp. 217-224 (en russe).

2.3 Publications en éthique des sciences

1. "Quelques questions éthiques des nanotechnologies", in: Traité de bioéthique. Tome 4: Quelques territoires nouveaux de bioéthique. Paris : Éditions érès, 2018. P. 319-327.
2. "Chance as a value for artificial intelligence", Journal of Responsible Innovation, 2018. DOI 10.1080/23299460.2018.1495032
3. « Contre la transparence : La Valeur du hasard pour une machine apprenante », Revue française d'éthique appliquée, 5, 47-53, 2018.
4. **"Ethics in Robotics Research," IEEE Robotics and Automation Magazine 24, 139-145, 2017 (avec R. Chatila, L. Devillers, J.-G. Ganascia, C. Tessier et M. Dauchet).**
5. « Responsabilité des êtres calculants », Revue française d'éthique appliquée 3, 117-123, 2017.
6. « La responsabilité du chercheur est-elle seulement morale ? », Responsabilité éthique face aux biotechnologies. Swiss Philosophical Preprint Series, 126, 8-9, 2016.
7. « L'Irrévérence de Hans Jonas », Esprit, 426, 213-216, 2016.
8. **"Uncanny valley explained by Girard's theory", IEEE Robotics and Automation Magazine 22(1), 152-150, 2015.**
9. « **Faibles doses : quelques questions philosophiques** », **ERS, 5, 2014, pp. 420-423.**
10. "The Old-New Meaning of Researcher's Responsibility", Etica e Politica / Ethics and Politics, XV, 2013, 1, pp. 236-250.
11. « Le vivant à l'aune de la biologie de synthèse », Biofutur. 339 (janvier 2013), pp. 41-42.
12. « **Biologie de synthèse : questions de société** », **Annales des Mines : Réalités industrielles, février 2013, pp. 91-95.**

13. **"What is 'responsible' about responsible innovation? Understanding the Ethical Issues", in Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society (eds R. Owen, J. Bessant and M. Heintz), Wiley, 2013, pp. 119-142 (avec C. Groves).**
14. « Nom et image comme rites de la technoscience », Visible, 8, 2012, pp. 203-214.
15. "Nanotechnological Icons", Nanoethics, 5, 2011, pp. 195-202.
16. **"The Nanotechnological Golem", Nanoethics, 4, 2010, pp. 191-198.**
17. « Le défi social des nanotechnologies », Agir, septembre 2010, pp. 29-36.
18. « Au-delà des risques », in : Risques émergents, Paris : Economica, 2010, pp. 155-157. English version in: Emerging Risks, ed. C. A. Raimbault and A. Barre, Gower, 2012, pp. 209-212.
19. "Toolkit for ethical reflection and communication on nanoscience and nanotechnology". Publication of the European project observatoryNano, 2010 (avec M. Pavlopoulos and V. Bontems).
20. « Les dimensions éthiques des nanotechnologies », in : Nanotechnologies : Les guides de l'innovation, Paris : Techniques de l'ingénieur, 2010, pp. 10-12 (avec M. Pavlopoulos)
21. « La catastrophe nanotechnologique », Le Mensuel de l'Université, 2009.
22. « Nanosciences : les enjeux du débat », Le Débat, janvier-février 2008, pp, 65-79 (avec E. Klein et V. Bontems).
23. « Barrières cognitives dans la perception des nanotechnologies », Annales des Mines : Réalités industrielles, mai 2007, pp. 47–53.
24. "Cognitive Barriers in Perception of Nanotechnology," Journal of Law, Medicine and Ethics, 34(4), 2006, 689–694.
25. "Living With Uncertainty: From the Precautionary Principle to the Methodology of Ongoing Normative Assessment", C. R. Geoscience 337(4), 2005, 457–474 (with J.-P. Dupuy).
26. **"Living with Uncertainty: Toward a Normative Assessment of Nanotechnology", Techné 8 (2), 2004, 4-25 (with J.-P. Dupuy).**
27. « La condition de l'homme moderne et les nanotechnologies », in : G. Nivat (ed.), Les limites de l'humain. 39ème Rencontres Internationales de Genève, L'Age d'Homme, Geneva, 2004, p. 141.

Chapitre 3

Activités d'encadrement

J'ai co-dirigé deux thèses de doctorat : en philosophie des sciences, celle de Maël Pégny, soutenue en 2013 [171], et, en physique, celle d'Issam Ibnouhsein, soutenue en 2014. La thèse de M. Pégny, intitulée "Sur les limites empiriques du calcul. Calculabilité, complexité et physique", a été consacrée aux liens entre logique et physique, notamment à la thèse de Church-Turing physique. Le travail d'Issam Ibnouhsein avait pour sujet les ordres causaux indéfinis, notion inspirée par les développements en mécanique quantique et introduite dans le débat conceptuel sur la causalité à la fin des années 2000. Ce travail a notamment été pionnier dans l'utilisation des méthodes entropiques pour l'analyse des superpositions non-classiques des ordres causaux [132].

J'ai encadré trois postdoctorants en philosophie des sciences, un en philosophie de la physique et deux en éthique des sciences.

J'ai encadré quatre stages de fin d'étude en philosophie et fondements de la physique, sur des sujets aussi variés que les approches *background-free* en gravité quantique, la contextualité quantique ou l'utilisation de la théorie des catégories afin de décrire les structures fondamentales de la théorie quantique.

Depuis 2007, j'ai organisé 38 colloques et séminaires en philosophie et fondements de la physique ainsi qu'en éthique des sciences.

J'ai été partenaire des projet ANR, FP7 et H2020, en philosophie de la physique et en éthique des sciences, notamment coordinateur du premier projet ANR consacré aux fondements de la physique (2006-2009), du volet éthique et sociétal de l'Observatoire européen des nanotechnologies (FP7 osbervatoryNano, 2008-2012) et du projet phare du programme H2020 sur la notion d'innovation et recherche responsables (RRI-Practice, 2016-2019).

Chapitre 4

Originalité des recherches

4.1 Deux axes de recherche

Depuis plus de quinze ans, je poursuis des recherches dans deux domaines différents de la philosophie des sciences : l'un est consacré à philosophie et aux fondements de la physique, l'autre à l'éthique. Dans mon esprit, ils sont autonomes. Le choix des sujets d'étude en éthique se fait indépendamment de mes préoccupations liées à la physique quantique.

La juxtaposition, s'il en est une, n'existe que dans les zones obscures de ma conscience. La rendre explicite me paraît incommode et assez peu éclairant, sauf sur le plan autobiographique. Si je devais le faire, je dirais que l'origine de ce dédoublement réside dans un double besoin esthétique, au sens qu'a donné à ce mot Alexeï Lossev, extraordinaire philosophe russe du XX^{e} siècle :

> Ce qu'on peut qualifier d' « esthétique » présuppose une vie intérieure de l'objet, donnée, en même temps, à son extérieur, et une forme extérieure de l'objet donnant, en même temps, un accès direct à sa vie intérieure. Cela signifie que l'esthétique est, avant tout, ce qui est expressif, ou l'expressivité en soi. [147, p.309]

D'un côté, élevé dès l'âge de onze ans dans le monde des mathématiques, j'éprouve le besoin de sentir leur rigueur. Ainsi, au début des années 2000, je cherchais, après avoir quitté la physique des hautes énergies, un sujet pour le futur travail de thèse. Mon regard se tournait alors vers les sciences cognitives. Mais il n'en fut rien : le désir des mathématiques l'emporta et je m'engageai dans les études des fondements de la physique quantique. La compréhension ne vint que plus tard que la redoutable force qui me tenait et me guidait était, par nature, une astreinte intérieure à la beauté mathématique.

De l'autre côté, j'éprouve tout aussi vivement le besoin de me plonger dans l'histoire intellectuelle du monde issu de l'Empire romain. Une histoire, à qui veut la saisir, impose de chercher des structures et des formes récurrentes de la pensée, que j'appelle des *motifs*. L'historien Peter Brown, dont la formation initiale était centrée sur le Moyen Âge, l'évoque en ses propres termes lorsqu'il explique l'émergence de

son intérêt pour l'antiquité tardive :

> Institutions and powerful bodies of ideas, that I had known only in the medieval and post-Reformation periods — and many of which, in their modern form, still hung, like chill clouds, above the heart of any Irish boy, Catholic or Protestant — were shown to have originated first in a very distant, ancient world. [38]

Ce besoin de dégager des motifs atemporels en tout problème philosophique, notamment éthique, je l'ai acquis assez tard, après avoir achevé ma thèse de doctorat. Cela vint en conséquence des découvertes méthodologiques en philosophie et en sciences humaines, qui avaient produit sur moi un effet de révélation. Je fus fasciné par les possibilités de systématisation, non-mathématique mais non-triviale, qui s'appuie sur la récurrence d'un motif caché, souvent anthropologique, et s'applique à des réalités aussi différentes que le mythe ou les nouvelles technologies.

La force esthétique possède ainsi, comme dirait un mathématicien, un domaine de définition comprenant deux parties disjointes. Mais, au fond, ce n'est qu'une seule force, l'unique attrait à la beauté des structures abstraites. Impuissant à m'y opposer, je soumets à l'appréciation du jury un mémoire en deux volumes, consacrés à l'une et l'autre de mes thématiques de recherche.

4.2 Philosophie et fondements de la physique

La question de la place de l'observateur en mécanique quantique provoque aujourd'hui, comme il y a quatre-vingt ans, un malaise aisément reconnaissable. Ses causes sont multiples ; j'en choisis trois. Elles correspondent à trois voies que j'emprunte pour étudier l'observateur quantique avec des moyens philosophiques et mathématiques. L'espoir et l'ambition qui fondent ce programme de recherche en philosophie de la physique consistent à pouvoir tirer, à partir de moyens formels inédits, des leçons conceptuelles tout aussi inédites au sujet de l'observateur. Le recours aux mathématiques nouvelles permet d'intégrer ce concept, historiquement demeuré métathéorique, au corpus scientifique.

La première voie est philosophique et historique ; elle remonte à l'interprétation de la mécanique quantique proposée par Niels Bohr. Les écrits de Bohr, comme ses conférences orales, ont le mérite d'être concis et de paraître clairs. Or, leur sens n'est jamais facile à saisir. De Bohr à Wheeler en passant par Everett, je présente une lignée de la pensée épistémologique sur le concept d'observateur. Sa caractéristique principale réside dans la nécessité de formuler cette réflexion philosophique dans le langage rigoureux d'une logique ou d'une physique mathématiques. Mais, historiquement parlant, c'est la théorie de l'information qui permet de concevoir l'observateur en tant que notion scientifique.

La deuxième voie est celle des reconstructions axiomatiques de la mécanique quantique. Son point de départ se situe en 1935, année de la fondation de la logique quantique par von Neumann et Birkhoff. Un demi-siècle plus tard, après l'extinction des approches logiques, les reconstructions prennent la forme des modèles opérationnels.

L'axiomatique de Hardy, publiée au début des années 2000, marquera une ligne de partage entre deux époques.

Des travaux que j'ai pu mener dans le domaine des reconstructions, je tire deux enseignements. Le premier : ne sont véritablement intéressantes que les reconstructions partielles, plus éclairantes que les tentatives monstrueusement compliquées des reconstructions complètes. Les reconstructions partielles permettent de jeter une lumière sur un principe ou une caractéristique fondamentale de la mécanique quantique ; une seule à la fois, mais cela est bien plus instructif. La deuxième leçon concerne le choix de deux caractéristiques qui occupent une place centrale dans toutes les reconstructions. D'abord, il s'agit du postulat de continuité, qui peut prendre différentes formes mathématiques ; vient ensuite la règle de composition des systèmes, qui donne la borne de Tsirelson de l'inégalité de Bell sous forme CHSH. Certaines études de la contextualité quantique nous ont montré, à l'aide d'exemples édifiants, l'insuffisance de ces deux contraintes pour la reconstruction complète de la mécanique quantique. Malgré cela, elles y sont nécessaires. Je suggère quelques pistes pour mieux comprendre cette situation.

La troisième voie suit le chemin des développements des quinze dernières années en information quantique et en fondements de la physique. Nous avons été témoins de l'émergence des approches indépendantes du dispositif, dont les exemples, très variés, incluent des techniques de cryptographie quantique ou encore la recherche sur les ordres causaux superposés. Sur le plan philosophique, ces développements viennent en appui d'une conclusion contre-intuitive : il est concevable que la notion de système ne soit pas donnée *a priori* dans une théorie physique. Autrement dit, une théorie physique peut ne pas être construite à partir de systèmes. Il existe des cadres théoriques qu'il est impossible d'interpréter dans ces termes. Cela m'amène à une interrogation du sens de « théorie physique » dans les modèles indépendants du dispositif.

J'adopte un langage formel des entrées et des sorties, qui a la forme des suites de symboles dans un alphabet fini. C'est dans ce cadre mathématique abstrait que je propose d'élaborer, à l'aide de divers arguments empruntés à la théorie de la complexité de Kolmogorov, une approche qui aboutit à des conclusions nouvelles, notamment à une conception formelle d'observateur. Certaines parmi ces conclusions sont vérifiables expérimentalement ; à ce jour, elles n'ont pas été falsifiées. Quant à d'autres, elles ouvrent une perspective nouvelle sur l'étude de la contextualité et sur l'émergence de la description probabiliste en mécanique quantique. Somme toute, cela inaugure un nouveau point de vue sur l'observateur.

4.3 Éthique des sciences

Dans les mythes, récits fondateurs qui fécondent le temps de l'histoire, on découvre d'éventuelles solutions à des questions éthiques apparemment insolubles, posées par l'avènement des nouvelles technologies. Sous une forme ou sous une autre, ces dilemmes ont toujours été sources de tourment pour l'humanité. Bien avant l'invention

des voitures autonomes ou des premiers assistants robotiques, des récits associaient déjà le mal et la pure fonctionnalité des démons, ou le mal et l'innovation trop rapide, irréfléchie. Aujourd'hui, ce type de comparaison, que j'appelle « homologie » et qui fonde ma méthode d'analyse, nous permet de sortir d'une impasse bien connue des historiens des sciences. Il s'agit du paradoxe dit « de la nouveauté en éthique » : si une technologie disruptive, radicalement nouvelle, couvre un champ d'action dont l'étendue est incomparable à tout ce qu'on a connu antérieurement, alors les questions éthiques qu'elle pose nous semblent tout aussi nouvelles et incommensurables avec la réflexion morale qui l'a précédée.

Croire qu'il existe pareilles nouveautés en éthique, c'est être assuré de ne pas aller bien loin dans la réflexion. Les fonctionnalités qu'offrent aux utilisateurs les systèmes informatiques sont certes incomparables aux capacités d'action que nous avions à l'époque pré-numérique, mais le bien et le mal qu'elles introduisent dans le monde ne sont pas des espèces de « Big Bien » ou de « mal nanotechnologique » inconnues à l'échelle humaine. Si un élément chimique du tableau de Mendeleïev, par exemple, peut être réellement nouveau, ce n'est pas le cas des notions de bien et de mal. Les concepts moraux n'existent pas dans la nature : on les trouve dans les récits fondamentaux que composent les hommes et qu'ils se racontent. Je tente de dégager quelques motifs de ces récits communs aux mythes et à la réalité technologique.

Le motif central, le plus important pour penser le mal numérique, réside dans la pure fonctionnalité des systèmes informatiques. Chaque machine, qui accueille un individu particulier, que j'appelle « individu numérique », n'est ni libre ni dotée d'une volonté. Elle ne peut définir ses propres objectifs. Les individus numériques apprennent à partir de données dépourvues, pour eux, de toute signification. Leur puissance de calcul leur permet d'y trouver des corrélations qu'aucun homme ne pourra jamais mettre au jour avec la même célérité ni la même habileté. Toutefois, seul l'homme assigne aux machines des buts et attribue des causalités aux informations qu'elles communiquent.

Quel est le rapport entre un individu fonctionnel et le mal ? J'étudie d'abord quelques exemples : des assistants domestiques devenus délateurs, des voitures autonomes soudainement meurtrières ou des agents conversationnels qui injurient leurs interlocuteurs. Ces systèmes participent à des conflits humains, les provoquent parfois. Nous bénéficions de possibilités inouïes grâce aux nouvelles fonctionnalités utiles qu'offre le numérique, mais avec elles, apparaissent de nouveaux dangers. Comment peut-on penser ces risques ? C'est là que les mythes autour des anges et des démons — qui sont des êtres purement fonctionnels tout comme les machines — peuvent nous être précieux. J'en tire des enseignements éthiques pour les appliquer aux systèmes informatiques.

Cet aller-retour entre le numérique et le mythe sera répété, en variant sujets et récits, jusqu'à ce que le paysage éthique du monde des nouvelles technologies nous devienne suffisamment familier. Ce n'est que lorsque son caractère nouveau s'estompe qu'on sait qu'une technologie a été totalement intégrée dans notre existence, après avoir évolué, s'étant conformée aux contraintes matérielles comme aux exigences morales de ses utilisateurs. Or, lorsque la nouveauté disparaît, ce n'est pas seulement la

technologie qui a changé. Nous aussi, nous avons évolué. Comment penser la nouvelle condition humaine ?

Cette question est centrale dans mon travail en éthique des sciences. Une étude de la notion de responsabilité du chercheur, que j'ai pu mener au début des années 2010, a notamment suscité beaucoup d'intérêt [113]. Mais, dans ce mémoire, le thème de la responsabilité n'apparaîtra qu'au second plan. Avant tout, c'est la méthode de réflexion éthique, élaborée d'une manière que j'espère originale, que je soumets à l'appréciation du jury.

Volume I. Philosophie de la physique

Chapitre 5

Synthèse des travaux

5.1 L'observateur en mécanique quantique

John Wheeler souligne la difficulté de définir les limites de la « communauté des observateurs-participants » [218]. La théorie quantique ne donne aucune information au sujet de la composition physique de l'observateur : ce terme métathéorique ne bénéficie d'aucune description *constructive*. En particulier, on ne peut déduire des mesures quantiques si l'observateur est un humain, une machine, une molécule ou l'Univers tout entier. Cependant, la mécanique quantique décrit l'information que possède l'observateur. La seule exigence théorique est qu'il enregistre une information lorsqu'il obtient un résultat de mesure. Cette brève synthèse, a fortiori incomplète, de quelques approches historiques de l'observateur quantique montre le chemin vers sa conception informationnelle, formelle et mathématique. Pourtant, tous les points de vue que nous évoquons, de Bohr à Everett, sont antérieurs à la « deuxième révolution quantique », initiée par John Bell en 1964.

La conférence de Como, prononcée par Niels Bohr en 1927, pose le fondement de l'interprétation de Copenhague de la mécanique quantique [12]. Même si ce terme ne désigne qu'un ensemble disparate de points de vue exprimés par des élèves et des collègues de Bohr [87], on retrouve son idée centrale dans cette position :

> Ce n'est qu'à l'aide des idées classiques qu'il est possible d'associer une signification non-ambiguë aux résultats de l'observation... La nature même de l'observation physique exige que toute expérience soit ultimement exprimée dans les termes classiques. [34, p. 94]

Il existe deux façons d'interpréter cette citation, qui se distinguent par le sens qu'elles attribuent au mot « classique ». La première est une inférence assez brutale à la nécessité de la mécanique classique qui serait indissociable de la mécanique quantique :

> Il est impossible en principe de formuler les notions de base de la mécanique quantique sans utiliser la mécanique classique. [142, p. 2]

La seconde interprétation, plus subtile, dit que la mécanique quantique aurait besoin, non d'une mécanique mais d'un langage classique :

> Bohr went on to say that the terms of discussion of the experimental conditions and of the experimental results are *necessarily* those of 'everyday language', suitably 'refined' where necessary, so as to take the form of classical dynamics. It was apparently Bohr's belief that this was the only possible language for the *unambiguous communication* of the results of an experiment. [32, p. 38].

Cette opposition entre deux interprétations de Bohr est conceptuellement importante. Or, elle n'est pas faite systématiquement. Le premier point de vue, mécaniste, présuppose que le monde est fait de systèmes physiques, quantiques ou classiques, et qu'aucune notion d'observateur n'y est nécessaire. La seconde lecture, linguistique, insiste sur le « langage classique », les « concepts classiques » ou les « termes classiques ». Elle semble exiger la présence de quelqu'un qui les maîtrise : un observateur. L'observateur prépare le système et le mesure, mais il possède également, selon Bohr, une faculté linguistique.

Lorsque Bohr formule ses idées à la fin des années 1920, il utilise systématiquement des phrases à double sens, mécaniste et linguistique. Dans les décennies suivantes, sa manière de s'exprimer évolue. Sans toutefois nier l'importance du langage classique, Bohr focalise son analyse philosophique, dans les années 1940 et 1950, sur la capacité à communiquer les résultats d'une mesure entre plusieurs observateurs. La communicabilité devient son exigence première ; le recours au langage classique n'est qu'un moyen de l'obtenir, et il n'est point une condition *a priori*. À cet argument, Bohr en ajoute un autre, en insistant sur un traitement symbolique, donc mathématique, que l'on doit donner au problème de communicabilité.

L'attitude de Bohr envers le formalisme mathématique de la théorie quantique n'a jamais été simple. Vers la fin de sa vie, il devient « de plus en plus persuadé du besoin de recourir à une symbolisation pourvu que l'on souhaite exprimer les résultats physiques les plus récents » [133]. Cette insistance sur la symbolisation, censée fonder l'objectivité de la description physique, traverse, en effet, toute sa vie. En 1958, Bohr la relie au choix d'un formalisme mathématique pour la théorie quantique : « L'utilisation des symboles mathématiques garantit la non-ambiguité de définition exigée par la description objective » [35]. Ce que Bohr entend par là n'est pas encore clair [87, 119]. De notre point de vue, il est peu probable que Bohr se réfère aux espaces de Hilbert ; rappelons que, dans le même texte, il se montre très critique envers le formalisme de la fonction d'onde de Schrödinger. Pour mieux saisir la signification de cette phrase, lisons la correspondance de Bohr des années 1920 :

> Regarding the question discussed in your letter about what was meant, when I in my article in *Naturwissenschaften*, emphasized so strongly the quantum-theoretical method's symbolic character, I am naturally in complete agreement with you that every description of natural phenomena must be based on symbols. I merely sought to emphasize the fact, that this circumstance — that in quantum theory, we typically use the same symbols we use in the classical theory — doesn't justify our ignoring the large difference between these theories, and in particular necessitates the

> greatest caution in the use of the intuitive concepts [*anskuels-former*] to which the classical symbols are connected. Naturally, one doesn't easily run this danger with the matrix formulation, where the calculation rules, which diverge so greatly from the previously standard algebraic ones, hold quantum theory's special nature before our eyes. Furthermore, to use the word "symbolic" for non-commutative algebra is a way of speaking that goes back long before quantum theory, and which has entered into standard mathematical terminology. When one thinks about the wave theory, it is precisely its "visualizability" [*anskuelighed*] which is simultaneously its strength and its snare, and here by emphasizing the approach's [*behandlingens*] symbolic character, I was trying to bring to mind the differences — required by the quantum postulate — from classical theories, which are hardly ever sufficiently heeded. [33]

Une première interprétation, quasiment spontanée, de cette longue citation serait de voir dans la « symbolisation » un argument portant sur le caractère abstrait de la représentation des quantités observables. Ce caractère abstrait serait alors opposé à une représentation plus directe et moins abstraite que l'on trouve en physique classique. Dans son fin commentaire, Hans Halvorson fait cependant intervenir une autre distinction catégorique, proposée par Cassirer dans « La philosophie des formes symboliques », dont le premier tome paraît en 1923. Bohr l'a assurément lu. Il correspond avec Cassirer ; celui-ci parle du degré de représentation : « une vague exigence de ressemblance entre la chose et son image [*Darstellingsfunktion*] » laisse sa place à « une relation logique hautement complexe [*Bedeutungsfunktion*] » en physique mathématique [52]. Cette seconde conception revêt un « caractère symbolique », le même terme qu'utilise Bohr dans les années qui suivent la parution de l'ouvrage de Cassirer.

À cette interprétation, il convient d'ajouter un élément visant, non à opposer le degré d'abstraction des symboles mathématiques utilisés en physique classique et en mécanique quantique, mais à appuyer la thèse de leur caractère abstrait. De 1928 à 1958, Bohr ne se sent pas satisfait des niveaux d'abstraction qu'il trouve dans les différents outils mathématiques, comme la fonction d'onde ou l'espace de Hilbert, proposés pour la construction du formalisme mathématique de la mécanique quantique. Il est concevable que, de son point de vue, même vers 1958, il était encore nécessaire de découvrir un niveau de symbolisation et d'abstraction adéquat. Ces « symboles » devraient alors fournir à la théorie physique un fondement de « langage commun ». Nous proposons de recourir, à cette fin, à la théorie de l'information. En prolongeant et en modifiant l'idée de Bohr, le rôle que nous espérons donner aux mathématiques serait celui d'une élaboration d'une description formelle de la base linguistique, symbolique et abstraite, des théories physiques.

Sur le plan historique, ce rôle des mathématiques a été tout sauf évident. La pensée philosophique de la notion d'observateur a d'abord emprunté un chemin très différent, centré sur la conscience de l'observateur. Ainsi, en 1939, Fritz London et Edmond Bauer soutiennent, pour la première fois dans l'histoire des interprétations

relationnelles de la mécanique quantique, que les propriétés des systèmes dépendent de l'observateur, mais que celui-ci doit être représenté par une personne humaine. Ils excluent toute éventualité d'une description objective de la réalité, en faveur d'un point de vue subjectif : "It seems that the result of measurement is intimately linked to the consciousness of the person making it" [145, p. 48]. Or, différents observateurs s'accordent sur le résultat d'une mesure ; ceci nécessite aussi une explication. Pour London et Bauer, l'accord entre plusieurs observateurs conscients provient d'un autre niveau, rapporté à la communauté des scientifiques : dans cette « communauté de conscience scientifique », il existe « un accord sur ce qui constitue l'objet d'une étude » [145, p. 49]. Malgré l'évocation de cette mystérieuse « conscience scientifique », qui provoquera par la suite un large débat [67, 101], l'ouvrage de London et Bauer constitue la première tentative d'analyser spécifiquement la notion d'observateur quantique.

Le lien entre la pensée de Bohr et les propositions de London et Bauer est assuré par Eugene Wigner, qui remplace la notion de « conscience scientifique » par celle de la communicabilité de l'information. Dans la droite ligne de London et Bauer, Wigner stipule que la conscience de l'observateur « entre dans la théorie d'une manière inévitable et inalterable », car elle est la seule à pouvoir capter l'impression produite sur l'observateur par le système mesuré [223]. Ce point de vue se conjugue bien avec la fidélité à la tradition de Bohr, notamment lorsque Wigner propose d'établir un lien entre l'existence d'une fonction d'onde et le fait que « l'information fournie par la fonction d'onde soit communicable ». La notion d'information est ainsi posée, pour la première fois dans l'histoire de la mécanique quantique, au fondement d'une interprétation réaliste du vecteur d'état. Cependant, la communicabilité de l'information, pour Wigner, ne dépend pas de l'usage des symboles mathématiques. Il ne reprend pas à son compte l'insistance de Bohr sur ce sujet, en reliant la communicabilité à la réduction de la fonction d'onde et au problème dit d'« ami de Wigner » :

> The communicability of information means that if someone else looks at time t and tells us whether he saw a flash, we can look at time $t + 1$ and observe a flash with the same probabilities as if we had seen or not seen the flash at time t ourselves. [222]

Comment se passe en pratique la communication entre observateurs ? Wigner ne donne à cette question qu'une réponse vague : « Si quelqu'un d'autre détermine d'une certaine manière l'état d'un système, alors il peut m'en parler… ». Ici, l'usage du « peut » suggère un mode seulement potentiel de la communication ; en même temps, le verbe « parler » pointe vers la suffisance du langage commun pour communiquer le résultat de la mesure en induisant la réduction de la fonction d'onde. La conscience, selon Wigner, en serait chargée : c'est en elle qu'un énoncé linguistique « agit » sur la fonction d'onde.

Le mysticisme wignerien à propos de la conscience humaine, lieu de rencontre entre le langage commun et le formalisme de la mécanique quantique, a été critiqué, à juste titre, par plusieurs auteurs. Léon Brillouin affirme que l'information doit être

définie avec l'exclusion de tout élément humain [95, p. 360]. Satosi Watanabe attribue l'origine de l'accord entre observateurs à la direction du temps qu'ils possèdent tous en partage : "The past-to-future directions of all observers coincide. This statement has a well-defined physical meaning, for 'positive time direction' is a Lorentz-invariant concept" [95, p. 387].

Créateur de l'interprétation éponyme de la mécanique quantique, Hugh Everett est aussi un fin penseur de l'observateur quantique. Comme Brilloin, il croit que la conscience humaine ne doit y jouer aucun rôle. L'observateur se définit, selon Everett, en tant que système possédant une mémoire : une partie telle que « son état entre en correspondance avec l'expérience passée » [86]. Les systèmes dotés d'une mémoire ne sont pas nécessairement des êtres humains ; sont aussi qualifiés les « machines à fonctionnement automatique, qui possèdent un appareil sensoriel et qui sont couplés aux dispositifs d'enregistrement ». Ainsi l'observation repose essentiellement, pour Everett, sur la présence d'une mémoire. Les enregistrements dans la mémoire doivent maintenir la distinction entre les états propres d'un système :

> If we are to able to call the interaction an observation at all, the requirement that the observer's state change in a manner which is different for each eigenfunction is necessary.

Cette exigence, formulée dans le langage formel de la mécanique quantique, est mathématique : elle porte, non sur le support physique de la mémoire, mais sur son fonctionnement. Différents observateurs peuvent utiliser des supports différents ; la seule condition commune est relative à la taille de la mémoire, suffisamment grande pour maintenir la distinction entre les états propres du système observé. Ainsi, avec Everett, nous rencontrons pour la première fois une conception abstraite de l'observateur quantique, qui repose sur des conditions mathématiques, plutôt que sur l'utilisation du langage commun ou sur la conscience humaine.

5.2 Reconstruction de la théorie quantique

5.2.1 Définition

Une interprétation vise à donner à la mécanique quantique une signification claire. La difficulté principale de cette tâche réside dans le « problème de la mesure » : l'évolution, unitaire et réversible, de la fonction d'onde se voit remplacée, au moment de l'observation, par une réduction irréversible et non-unitaire [213]. Les interprétations proposent différentes stratégies pour donner un sens physique à cette réduction, soit en postulant que la mesure projective découle naturellement des variables fondamentales de la théorie (Bohm), soit en reniant la réduction complètement (Everett).

Dans le travail de doctorat [101], nous avons montré l'insuffisance conceptuelle des interprétations, car elles ne font qu'adjoindre au formalisme mathématique une couche externe de « signification », tout en considérant ce formalisme comme fixe. Nous avons

supputé qu'une signification satisfaisante ne peut qu'émerger avec le formalisme lui-même, dans une démarche de dérivation à partir d'axiomes fondamentaux. Ce travail, publié sous forme de quelques articles, a suscité beaucoup d'intérêt [103].

La démarche de reconstruction que nous avons proposée n'est pas tout à fait équivalente au schéma de reconstruction rationnelle de Carnap [51]. Elle procède en trois étapes. D'abord, les théorèmes et les principaux résultats d'une théorie physique sont formellement dérivés des présupposés mathématiques plus simples. À leur tour, ces présupposés, ou axiomes, apparaissent en tant que traductions, dans le formalisme mathématique, d'un ensemble de principes physiques. Ce sont ces principes qui donnent, dans un troisième temps mais pas le moindre, un sens à la théorie. Ces trois étapes sont celles d'un travail de recherche. Dans sa forme achevée, une reconstruction procède dans le sens opposé : d'abord, un ensemble de principes physiques fondamentaux ; ensuite, leur représentation dans un formalisme mathématique ; enfin, une dérivation rigoureuse du formalisme de la théorie.

Par rapport aux interprétations, les reconstructions bénéficient d'une force persuasive supplémentaire, puisée dans les mathématiques. Les théorèmes de la théorie sont établis formellement ; leur signification devient transparente grâce aux axiomes dont ils sont dérivés. Le problème de signification porte alors, uniquement sur les principes fondamentaux. En particulier, le problème de la mesure perd son caractère central, jusqu'au point de devenir un « pseudo-problème », comme l'appelle le philosophe Jeffrey Bub dans un récent ouvrage consacré aux approches théorético-informationnelles de la théorie quantique [47, p. 223]. La force persuasive d'une reconstruction ne dépend pas du traitement qu'elle fait du problème de la mesure, mais émane seulement des principles qui fondent la structure de la théorie.

Le point de départ d'une reconstruction consiste en un ensemble de principes, représentés sous forme d'axiomes mathématiques. Puisque ces principes portent le sens de la théorie, ils doivent être compréhensibles. Mais d'où les tire-t-on ? Quels sont les bons candidats pour tenir le rôle de principe et quels autres doivent être rejetés ?

Avant de sélectionner un principe fondamental pour une théorie, il est nécessaire de développer une intuition à propos de ce qui est plausible, ou au contraire invraisemblable, dans cette théorie. Dans le cas de la mécanique quantique, comme dans d'autres, cela n'est possible qu'au travers de la pratique de cette théorie : il s'agit d'employer le formalisme standard de la mécanique quantique pour la résolution de problèmes concrets. Lorsqu'on aura acquis une bonne expérience pratique, il sera légitime de se demander quels théorèmes ou quels résultats observés correspondent, dans le formalisme, à des traits profonds et répétitifs de la théorie. Ce choix de principes candidats ne peut s'effectuer de manière logique, comme le rappelle Einstein dans un exposé informel mais profond sur son épistémologie [81]. À une autre occasion, Einstein suggère l'« élévation » [92] des faits empiriques au statut d'axiomes, opération dont on ne peut rendre compte purement logiquement. Le choix des principes marque ainsi une irruption de l'élément heuristique dans la tâche de reconstruction.

Dans un article exposant les fondements épistémologiques de notre point de vue

[108], nous avons souligné que les principes fondamentaux d'une théorie ne sont pas des vérités ultimes à propos de la nature. Indépendamment des croyances ou des convictions ontologiques, qui restent personnelles et subjectives, les principes possèdent un statut épistémologique minimal et objectif : ils rendent compte de la signification de la théorie physique particulière qui en découle. Cependant, une autre théorie pourrait s'appuyer sur des principes différents. Les deux ensembles de principes demeurent valides, chacun pour sa tâche bien circonscrite, même malgré une éventuelle contradiction entre eux. Il n'est nullement question d'établir des principes universels, qui s'appliqueraient à toutes les théories physiques.

En physique théorique, comme dans les mathématiques du XIX^{e} siècle, la méthode axiomatique doit être séparée d'une attitude, qui remonte aux Grecs, visant à donner aux axiomes le sens de vérités absolues à propos du monde réel. Beaucoup de progrès en mathématiques sont dus à cette séparation épistémologique, qui permet de voir dans les axiomes uniquement des éléments structurels d'une théorie. Étendue à la physique théorique, cette attitude aboutit à une prescription d'économie épistémologique, se gardant de toute affirmation théorique péremptoire invoquant « le réel ». Notre attitude de modestie épistémologique consiste à limiter la portée ontologique des principes physiques en restreignant leur apport philosophique aux seules les théories qu'ils fondent.

Ainsi, les principes ne peuvent être fondamentaux que *hic et nunc*, dans une reconstruction donnée. L'intuition, l'heuristique ou les croyances, qui nous ont amenés à élever des faits empiriques au rang de principes, ne nous informent ni sur le réel ni sur l'ontologie, car elles sont inévitablement *theory-laden* : imprégnées de théorie. Mais, une fois les principes sélectionnés, il est parfaitement rationnel de mettre entre parenthèses tous ces facteurs imprécis qui avaient contribué à leur choix. Les croyances restent subjectives et ne peuvent, en cette qualité, que fonder un point de vue personnel sur le réel. La théorie, quant à elle, acquiert un sens en vertu de ses principes.

Einstein, qui soutient la méthode axiomatique car elle « fait disparaître l'obscurité qui entoure actuellement les principes », reste néanmoins conscient du lien problématique entre les « schèmes conceptuels » et les « objets réels » [78]. Cette réserve correspond bien à l'attitude de reconstruction épistémologiquement modeste.

En allant encore plus loin, nous sommes tentés de dire que le problème du réel ne peut qu'être métaphysique, formulé et résolu au-delà de ce que contient la théorie physique.

5.2.2 De la logique quantique aux approches opérationnelles

Le premier article proposant un traitement axiomatique de la mécanique quantique paraît peu de temps après la création de la théorie : en 1927, von Neumann et Nordheim affirment qu'en mécanique quantique, « l'appareil analytique de la théorie et les quantités arithmétiques qui y figurent reçoivent une inteprétation physique fondée sur des postulats physiques. Ici, le but est de formuler des exigences physiques de manière si complète que l'appareil analytique en sorte déterminé de façon unique. Cette voie est celle d'axiomatisation » [126]. Une vision de physique mathématique,

issue du programme de Hilbert, donne ainsi naissance au traitement axiomatique de la mécanique quantique.

La logique quantique voit le jour en 1935. Dans un texte fondateur [30], von Neumann et Birkhoff montrent un exemple du changement d'attitude, selon lequel la théorie physique devrait opérer, non avec des objets réels, comme il était stipulé dans le célèbre article EPR contemporain [82], mais avec des entités théoriques faisant partie d'une dérivation mathématique. Dans un geste d'adieu à la formulation de la mécanique quantique dans l'espace de Hilbert, qu'il avait lui-même proposée quelques années auparavant, von Neumann livre dans une lettre à Birkhoff une « confession » sur le fait qu'il « ne croyait plus aux espaces de Hilbert » [215]. Afin de décrire les systèmes physiques différemment, von Neumann souhaite établir une correspondance entre les mesures et une structure de géométrie projective, isomorphe à un treillis non-booléen.

Suivant le premier travail de von Neumann et Birkhoff, la logique quantique donne naissance au bon nombre de reconstructions axiomatiques, notamment celles proposées par Zieler [225], Varadarajan [210, 211], Piron [176, 177], Kochen et Specker [138], Guenin [115], Gunson [117], Jauch [134], Pool [183, 184], Plymen [181], Marlow [155], Beltrametti et Casinelli [26], Holland [128] ou Ludwig [148]. Sur une voie d'axiomatisation alternative, mais compatible avec la voie logiquee, on vit se développer des approches algébriques, initialement conçues par Jordan, von Neumann et Wigner [135], puis par Segal [193, 194], Haag et Kastler [118], Plymen [180] ou Emch [85] ; ces recherches sont bien résumées dans les ouvrages postérieurs [26, 44]. Un système axiomatique contenait typiquement, à cette époque, une longue liste de postulats formulés mathématiquement. Leur motivation physique restait obscure ; on cherchait, par-dessus tout, la rigueur mathématique.

Citons, à titre d'exemple, le système axiomatique pour la mécanique quantique, proposé au début des années 1950 par George Mackey [149, 150]. Un ensemble désigné $\mathfrak{B}$ est constitué de tous les sous-ensembles de Borel des nombres réels ; un ensemble abstrait $\mathcal{O}$ est traité comme l'espace d'observables et un autre ensemble, $\mathcal{S}$, comme celui des états. On suppose qu'une fonction p, qui met un nombre réel $0 \leq p(x, f, M) \leq 1$ en correspondance avec chaque triplet x, f, M, l'élément x étant dans $\mathcal{O}$, f dans $\mathcal{S}$ et M dans $\mathfrak{B}$, obéisse aux axiomes suivants :

M1 La fonction p est une mesure de probabilité. Mathématiquement, on pose $p(x, f, \emptyset) = 0$, $p(x, f, \mathbb{R}) = 1$ et $p(x, f, M_1 \cup M_2 \cup M_3 \ldots) = \sum_{n=1}^{\infty} p(x, f, M_n)$ si M_n sont des ensembles de Borel disjoints par paire.

M2 Pour que deux états soient différents, ils doivent attribuer des distributions de probabilité différentes au moins à une observable. Pour que deux observables soient différentes, elles doivent posséder des distributions de probabilité différentes au moins dans un état. Mathématiquement, si $p(x, f, M) = p(x', f, M)$ pour tout f dans $\mathcal{S}$ et pour tout M dans $\mathfrak{B}$, alors $x = x'$; et si $p(x, f, M) = p(x, f', M)$ pour tout x dans $\mathcal{O}$ et pour tout M dans $\mathfrak{B}$, alors $f = f'$.

M3 Soit x un membre de $\mathcal{O}$ et soit u une fonction de Borel sur les réels, elle-

même réelle et bornée. Alors il existe un y dans $\mathcal{O}$, tel que $p(y, f, M) = p(x, f, u^{-1}(M))$ pour tout f dans $\mathcal{S}$ et pour tout M dans $\mathfrak{B}$.

M4 Si f_1, f_2,... sont membres de $\mathcal{S}$ et si $\lambda_1 + \lambda_2 + \ldots = 1$, où $0 \leq \lambda_n \leq 1$, alors il existe f dans $\mathcal{S}$, tel que $p(x, f, M) = \sum_{n=1}^{\infty} \lambda_n p(x, f_n, M)$ pour tout x dans $\mathcal{O}$ et pour tout M dans $\mathfrak{B}$.

M5 Soit une *question*, une observable e dans $\mathcal{O}$ telle que $p(e, f, \{0, 1\}) = 1$ for all f in $\mathcal{S}$. Les questions e et e' sont disjointes si $e \leq 1 - e'$. Alors une question $\sum_{n=1}^{\infty} e_n$ existe pour toute séquence (e_n) de questions telle que e_m et e_n sont disjointes si et seulement si $n \neq m$.

M6 Si E est une mesure compacte dans l'ensemble des questions, alors il existe une observable x dans $\mathcal{O}$, telle que $\chi_M(E) = E(M)$ pour tout M dans $\mathfrak{B}$, où χ_M est la fonction caractéristique de M.

M7 L'ensemble partiellement ordonné de toutes les questions en mécanique quantique est isomorphe à l'ensemble partiellement ordonné des sous-espaces fermés d'un espace de Hilbert séparable de dimension infinie.

M8 Si e est une question différente de 0, alors il existe un état f dans $\mathcal{S}$, tel que $m_f(e) = 1$.

M9 Pour toute séquence (f_n) des membres de $\mathcal{S}$ et pour toute séquence (λ_n) de nombres réels non-négatifs, dont la somme est égale à 1, le groupe d'évolution temporelle à un paramètre $V_t : \mathcal{S} \mapsto \mathcal{S}$ agit comme : $V_t\left(\sum_{n=1}^{\infty} \lambda_n f_n\right) = \sum_{n=1}^{\infty} \lambda_n V_t(f_n)$ pour tout $t \geq 0$; et pour tout x dans $\mathcal{O}$, f dans $\mathcal{S}$, et M dans $\mathfrak{B}$, $t \to p(x, V_t(f), M)$ est continu.

De ces axiomes, couchés dans le langage de la logique quantique, Mackey déduit toutes les composantes essentielles du formalisme de la mécanique quantique : la structure de l'espace de Hilbert à partir de M5-M8, l'espace des états et l'interprétation probabiliste à partir de M1-M4, et l'évolution temporelle à partir de M9. Cette approche purement mathématique, qui obscurcit la signification et l'origine physique des principes fondamentaux, se trouve à l'opposée des démarche de reconstruction modernes. Pour apprécier cette évolution, comparons le travail de Mackey avec celui mené, un demi-siècle plus tard, par Lucien Hardy [121].

Le système axiomatique de Hardy représente une approche nouvelle appelée « opérationnelle ». Elle considère comme fondamentaux uniquement les paramètres des mesures et leurs résultats, seules données disponibles à l'observateur. Ces éléments forment typiquement un ensemble convexe. Sur le plan opérationnel, Hardy commence sa dérivation par deux nombres entiers, K et N. K est le nombre des degrés de liberté du système, défini comme le nombre minimal de mesures nécessaires pour caractériser son état. N, la dimension, est définie par le nombre maximal d'états que l'on peut fidèlement distinguer en une seule mesure. Les axiomes ne font intervenir que ces deux quantités :

H1 *Probabilités.* Dans la limite de n tendant vers infini, les fréquences relatives (que l'on mesure par la proportion des occurrences d'une sortie donnée) tendent

vers la même valeur quel que soit le cadre dans lequel on effectue une mesure donnée sur un ensemble de n systèmes préparés selon une procédure donnée.

H2 *Simplicité.* K est déterminé par une fonction de N, avec $N = 1, 2, \ldots$, et pour tout N donné, K prend la valeur minimale qui est cohérente avec les axiomes.

H3 *Sous-espaces.* Un système dont l'état appartient, suite à des contraintes, à un sous-espace de dimension M se comporte comme un système de dimension M.

H4 *Systèmes composés.* Un système composé de deux sous-systèmes, A et B, satisfait à $N = N_A N_B$ et à $K = K_A K_B$.

H5 *Continuité.* Pour toute paire d'états purs d'un système, il existe une transformation continue et réversible qui les relie.

On s'aperçoit immédiatement que ce système axiomatique est plus bref et plus simple à appréhender que celui de Mackey. Le sens physique des axiomes est assez facile à saisir, grâce aux intitulés qui en font des principes physiques.

Toutefois, Hardy insiste que le choix d'une philosophie instrumentaliste n'est pas un prérequis essentiel à toute reconstruction opérationnelle. Certains verraient dans ce choix d'axiomes une approche réaliste, d'autres préféreraient une lecture en termes de variables cachées ou des interprétations de la réduction spontanée du paquet d'onde. La reconstruction de Hardy ne dépend pas de pareilles convictions. C'est même son atout ; il a pour origine l'emploi des méthodes de dérivation mathématique, appliquées à des principles physiques simples.

La démarche de reconstruction opérationnelle éclaire donc la mécanique quantique d'une manière nouvelle. Toutefois, le sens physique de l'axiome H5 reste obscur. Son origine physique relève presque du mystère. Hardy reconnaît dans le postulat de continuité une caractéristique fondamentale de la théorie quantique, qui reste à être appréhendée.

5.2.3 Une reconstruction fondée sur les principes informationnels

En 1996, Carlo Rovelli propose une interprétation relationnelle de la mécanique quantique, fondée sur deux principes informationnels [191] :

R1 Il existe une quantité maximale d'information pertinente que l'on peut extraire d'un système.

R2 Il est toujours possible d'obtenir une information nouvelle à propos d'un système.

Nous avons utilisé ces principes dans une reconstruction de la mécanique quantique visant à dériver la structure de treillis orthomodulaire [100, 102]. Dans ce cadre, l'information est toujours définie relativement à un observateur ; pour Rovelli, l'état d'un système, loin d'être objectif, est nécessairement indexé par l'observateur qui le décrit. On peut alors poser le problème de l'accord intersubjectif entre deux ou plusieurs observateurs, dont l'origine demeure à ce jour un des points d'achoppement des approches relationnelles [42].

En lien avec R1, Rovelli introduit l'idée de « capacité informationnelle maximale » d'un système. Cette quantité, selon lui, est responsable de l'apparition en physique de la constante de Planck. Nous en avons proposé une interprétation différente : la capacité informationnelle maximale pourrait varier selon la taille de la mémoire des observateurs.

Rappelons brièvement les étapes de la démonstration de l'orthomodularité du treillis. Cette preuve repose sur la conceptualisation en logique quantique de la notion d'information pertinente, introduite dans l'axiome R1. Historiquement, l'orthomodularité n'apparaît qu'au stade avancé du développement de la logique quantique. Avant elle, la première innovation mathématique fut celle des algèbres de von Neumann [161] ; puis, celle des treillis modulaires [30] ; et seulement après, suite au travail de Husimi [131], consacré principalement à la dérivation d'une logique non-booléenne à partir de faits expérimentaux, on introduisit le concept de treillis orthomodulaire [146]. Pour la reconstruction complète de la théorie quantique, celui-ci est nécessaire, mais pas suffisant [136]. Il est typiquement postulé ou dérivé à partir de présupposés formels, portant sur les relations entre les éléments du treills. Le travail de Beltrametti et Cassinelli présente une remarquable exception à cette règle ; ils justifient l'orthomodularité conceptuellement :

> Orthomodularity corresponds to the survival... of a notion of the logical conditional, which takes the place of the classical implication associated with Boolean algebra. [26]

Une motivation similaire se trouve également dans l'œuvre de Jauch et Piron. Ils reformulent l'orthomodularité comme la propriété suivante : si une proposition est plus grande qu'une autre dans l'ordre du treillis, alors ces deux propositions sont compatibles [176, 134]. Drieschner, dont la motivation est encore différente, donne une formulation concise de l'idée sous-jacente à la logique quantique des années 1970 : "If x implies y, they have to be compatible" [74]. Cette interprétation repose essentiellement sur la possibilité d'introduire la compatibilité des propositions d'une manière indépendante de l'introduction de l'orthomodularité. La première fait référence à la philosophie de Bohr et aux relations de Heisenberg, tandis que la seconde est placée au fondement d'une reconstruction de la théorie qui contient ces relations. Le risque de circularité dans un tel argument est apparent.

Notre reconstruction emploie la relation de pertinence pour lier les propositions binaires. Supposons que a fournisse une information à l'observateur. Pour la rendre non-pertinente, celui-ci peut poser une nouvelle question, b, telle que b implique la négation de a : $b \rightarrow \neg a$. Pour motiver cette définition, il convient de remarquer que, sur le plan heuristique, si un observateur s'attend *honnêtement* à pourvoir obtenir les deux réponses possibles à b, 0 ou 1, alors l'information qu'il possède d'après a n'est plus pertinente. Formellement :

Definition 5.1. Question b is called irrelevant with respect to question a if $b \wedge a^{\perp} \neq 0$. Otherwise question b is called relevant with respect to question a.

Dans un treillis hilbertien, il est vrai que si $b > a$, alors b est toujours non-pertinent par rapport à a. Tout sous-espace fermé $c \subseteq a$ fournit un élément de treillis pertinent

par rapport à a ; tous les autres élément sont, dans le cas hilbertien, non-pertinents. L'intérêt de notre notion de pertinence provient du cas générique, car elle ne s'y réduit pas à une simple inclusion ensembliste. Elle nous permet de proposer une signification informationnelle de l'orthomodularité [100].

5.2.4 Reconstructions partielles

De 1935 et jusqu'à la fin des années 1990, toutes les reconstructions de la mécanique quantiques visaient à réunir un ensemble d'axiomes permettant de dériver l'intégrité de son formalisme. Peu d'entre elles y sont parvenues, et le sens physique de ces « réussites »était le plus souvent éparpillé entre les principes et les postulats de natures et d'origines différentes. Certains servaient à dériver la structure algébrique de l'espace des états, d'autres la fonction de probabilité, ou d'autres encore, l'évolution temporelle. Un type nouveau de reconstruction théorético-informationnelle a vu le jour au début des années 2000 : les reconstructions partielles. En partant des principes fondamentaux au nombre réduit, voire d'un seul principe, on souhaite dériver, non l'ensemble des composantes du formalisme mathématique de la mécanique quantique, mais un de ses traits particulièrement saillant. Les deux caractéristiques centrales sont, comme nous verrons, la continuité des transformations entre les états purs du système et la règle de composition des sous-systèmes.

Afin d'opposer les significations de la mécanique quantique avec et sans la propriété que l'on souhaite reconstruire, on invente des modèles « quasi quantiques » ou « post-quantiques ». Cette idée de modifier la mécanique quantique n'est pas nouvelle en soi ; par exemple, des extensions non-linéaires de l'équation de Schrödinger furent étudiés dans les années 1980 [29, 96, 216, 162, 205]. Toutefois, le but des reconstructions partielles n'est pas de remplacer la mécanique quantique par une autre théorie. Ces modèles ne visent pas à décrire le monde réel. Leur ambition est de permettre la comparaison entre les cas d'absence et de présence, au sein d'un modèle, d'une propriété cruciale de la théorie quantique. L'objectif est d'explorer les conséquences du manque de cette propriété ; puis, dans la limite du possible, de chercher des candidats à l'élévation au rang de principe fondamental, responsable de l'apparition en mécanique quantique de la propriété en question. Cette démarche prolonge et étend l'épistémologie einsteinienne, évoquée *supra*.

Parmi les modèles proches de la mécanique quantique, mais non-quantiques, il faut mentionner celui de Spekkens, qui contient un seul principe : la quantité d'information disponible à propos d'un système est égale à celle qui n'est pas disponible. Par ailleurs, les « théories probabilistes générales » de Barrett [1, 2, 17, 120, 186, 200, 203], appelés péjorativement « fantasy quantum mechanics » ou « quantum mechanics lite », ouvrent des possibilités inédites pour l'analyse de certaines caractéristiques de la théorie quantique choisies à la carte, par exemple, de la possibilité de signalement supralumineux, du *bit commitment*, de la téléportation, du codage dense, du *remote steering*, etc. Dans les modèles post-quantiques, pareilles propriétés computationnelles sont typiquement différentes du calcul quantique standard ; de là on extrait des enseignements quant à l'origine des propriétés observées dans le monde quantique.

Des propriétés cruciales de la théorie quantique, telles que la non-localité, la contextualité, l'existence d'un espace des états continu, peuvent également être analysées par la voie des reconstructions partielles. Par exemple, le modèle de Spekkens contient des caractéristiques que l'on croyait être typiquement quantiques, comme la non-commutativité, l'interférence, la multiplicité des décompositions convexes d'un état mixte, l'absence de clonage, la téléportation ou le pouvoir computationnel proche du quantique [203, 187, 72]. En comparant ce modèle avec la mécanique quantique, on apprend que l'existence d'un espace continu des états, d'un théorème de Bell ou encore la contextualité, qui sont absentes du modèles de Spekkens, n'empêchent pas la préservation de quelques propriétés quantiques. Elles sont donc, dans une certaine mesure, indépendantes.

Entre autres, le modèle de Spekkens, formé d'un ensemble discret d'états, met en lumière le rôle du continu dans les reconstructions de la mécanique quantique. Les axiomatiques de Mackey ou de Piron, formulés dans le langage logique, contenaient déjà des présupposés dont découlait le caractère continu des transformations entre les états d'un système [149, 176]. Cet élément du formalisme quantique devient crucial à toutes les tentatives d'axiomatisation, à partir du travail de Solèr, qui propose un axiome explicite pour le choix d'un corps numérique de l'espace de Hilbert [201]. L'axiome de Solèr, comme la reconstruction de Zieler [225], vise directement le problème du corps *numérique* et garantit que celui-ci est le corps des réels, complexes ou quarternions. Ce travail important prouve qu'un axiome de continuité est nécessaire, mais pas suffisant, pour toute reconstruction de la mécanique quantique. Son insuffisance sera, d'ailleurs, bien illustrée par une reconstruction fondée sur l'utilisation des algèbres C* [59].

Les postulats de continuité du second type, même s'ils mènent aux mêmes conclusions que les postulats du premier, visent à reconstruire le caractère continu de l'espace des états d'un système. L'axiome H5 de Hardy en fait partie, mais d'autres exemples existent également. On peut citer la non-contextualité de Gleason [98], l'homogénéïté de l'espace des paramètres [43, 65], la "two-sphere property" de Landsmann [143] ou les axiomes C et D proposés par Holland [128].

Cette intrusion du continu en physique, quelle que soit sa forme, ne saurait être sans conséquences quant aux prédictions expérimentales et à leur accord avec les mesures empiriques. Une analyse récente de cette problématique a été proposée par Gisin [97]. De notre côté, dans une critique de l'argument de Wigner sur l'efficacité des mathématiques en physique [110], nous avons discuté de la place à donner en physique à des éléments d'abstraction mathématique, comme l'emploi du continu. Cette question est aussi soulevée dans les débats autour du problème d'hypercalcul et de la thèse de Church-Turing physique, qui confronte les pratiques concrètes du calcul avec ses modèles théoriques. Cette thématique a été particulièrement étudiée par notre doctorant Maël Pégny [171, 172].

Il suffit de mentionner, dans ce court mémoire, un problème fondamental de calculabilité, évoqué pour la première fois par Nielsen [166]. Dans sa formulation standard, utilisant un corps numérique continu, la mécanique quantique permet l'existence des observables qui font intervenir des nombres réels non-calculables. Ces nombres

peuvent apparaître en tant que coefficients des états superposés, des valeurs propres des opérateurs ou des prédictions probabilistes de la théorie. Nielsen construit l'observable $h = \sum_{x=0}^{\infty} h(x)|x\rangle\langle x|$, où la base orthonormale des états $|x\rangle$ décrit l'état d'un système physique concret et $h(x)$ est la fonction d'arrêt. Le problème d'arrêt étant indécidable, cette observable encode une information qui ne peut pas être calculée ; pourtant, la mécanique quantique n'interdit pas de mesurer l'observable de Nielsen. Il est donc clair que la confrontation entre la continuité du formalisme quantique et l'expérience provoque des troubles conceptuels et philosophiques ; de même pour la confrontation entre la mécanique quantique et les résultats logiques. Une adéquation parfaite entre eux reste donc inatteignable.

Les postulats de continuité, quelle que soit leur forme, sont responsables du fait que la reconstruction reproduit fidèlement le formalisme de la mécanique quantique. Sans être suffisants, ils sont nécessaires pour qu'on puisse retrouver la « quanticité » (*quantumness*). Mais il reste à établir leur rôle précis, et ce à quoi la théorie pourrait ressembler en leur absence. Le modèle de Spekkens fournit une première réponse à cette interrogation ; nous y revenons dans notre programme de recherche.

Après la continuité, une autre caractéristique fondamentale de la mécanique quantique, qu'il est possible d'analyser à l'aide des reconstructions partielles, est sa règle de composition, liée à la quantité de la non-localité. On décrit habituellement cette quantité par une limite qu'atteint tel ou tel modèle dans la violation de l'inégalité de Bell. Sous la forme Clauser-Horn-Shimony-Holt (CHSH) [58], l'inégalité de Bell donne une borne quantique bien connue, $2\sqrt{2}$, qui porte le nom du physicien soviéto-israélien Boris Tsirelson [57]. Certains modèles s'approchent de cette borne du haut ; dans notre programme de recherche, nous proposons, pour la première fois, des modèles qui y tendent par le bas.

Vers la fin de l'époque de logique quantique orthodoxe, dans les années 1990, il devint clair que la règle de composition des sous-systèmes est l'une des clefs de voûte de la théorie quantique. La structure de produit tensoriel caractérise la composition dans l'espace de Hilbert ; c'est elle qui fait naître l'intrication quantique. Il est donc nécessaire, selon la démarche de reconstruction, d'identifier les principes physiques dont on dérivera le produit tensoriel. Durant environ vingt ans, cette tâche incomba aux approches catégorielles, usant des catégories monoïdales pour analyser la règle quantique de composition, tout en laissant de côté les autres éléments du formalisme [60, 61, 62]. Aujourd'hui, par exemple, les méthodes catégorielles permettent d'établir une correspondance entre la composition dans l'espace de Hilbert et la composition des mots dans une phrase en langue naturelle. Ces méthodes n'apportent, hélas, pas d'éclairage décisif à la physique, mais elles contribuent à la diffusion de ses méthodes à d'autres domaines scientifiques.

Dans le domaine des reconstructions partielles, une nouvelle approche a vu le jour au début des années 1990, en lien avec les « boîtes » introduites par Popescu et Rohrlich (boîte PR) [186]. Elle peut être décrite dans le langage moderne des approches indépendantes du dispositif [111]. Une boîte PR est un modèle donné par les entrées $x, y \in \{0, 1\}$ et les sorties $a, b \in \{0, 1\}$ de deux parties, Alice et Bob, liées

par une distribution conjointe de probabilité :

$$P(ab|xy) = \begin{cases} 1/2 : & a + b = xy \mod 2 \\ 0 : & \text{autrement.} \end{cases} \tag{5.1}$$

La contrainte de non-signalement, vérifiée par les boîtes PR, implique que, bien que ce modèle ne soit pas quantique, il respecte néanmoins les lois de la relativité. Le caractère post-quantique du modèle devient évident lorsqu'on remarque qu'il permet de violer de façon maximale l'inégalité de Bell sous forme CHSH :

$$|E_{00} + E_{10} + E_{01} - E_{11}| = 4, \tag{5.2}$$

où les corrélateurs sont définis par $E_{xy} = P(a = b|xy) - P(a \neq b|xy)$.

Sur le plan expérimental, il n'est, bien sûr, pas possible de construire une boîte PR, mais seulement de s'en approcher [188]. En effet, la construction d'une boîte PR réelle exige, en vertu de son caractère post-quantique, qu'une quantité de signalement soit présente, sous forme de coordination entre les paramètres des mesures d'Alice et de Bob [114, 208]. Alternativement, on peut opérer une postsélection [154]. Chercher dans les boîtes PR quelque élément empirique serait donc vide de sens.

L'utilité de ce modèle, et son enseignement principal, réside dans la possibilité de comparer les boîtes PR avec la quantité de la non-localité disponible en mécanique quantique, en se demandant de quels principes découle la borne de Tsirelson. Parmi les différentes propositions, on trouve l'existence d'une limite classique, équivalente à un principe de correspondance entre la physique quantique et la physique classique [66]. Rohrlich a étudié l'existence d'un principe de correspondance analogue dans le cas des boîtes PR, que l'on aurait pu, en suivant l'exemple de Masanes et Müller [157], ériger en un postulat ; or, le résultat obtenu par Rohrlich fut négatif [189, 190]. D'autres tentatives, plus heureuses, de dériver la borne de Tsirelson incluent le recours à des principes comme une forme entropique des relations d'incertitude [212, 168], le *swapping* de la non-localité [197, 198], la localité macroscopique [165] ou la causalité informationnelle [170, 6]. Toutes ces reconstructions partielles montrent le chemin pour une interrogation conceptuelle du rôle que joue la non-localité quantique.

5.2.5 Rôle de la dérivation mathématique

Nous avons dit qu'une reconstruction est l'unique voie de réponse aux interrogations philosophiques à propos de la mécanique quantique. Cette idée n'est pas nouvelle ; on la trouve, par exemple, chez Rovelli :

> Quantum mechanics will cease to look puzzling only when we will be able to *derive* the formalism of the theory from a set of simple physical assertions ("postulates," "principles") about the world. Therefore, we should not try to append a reasonable interpretation to the quantum mechanical formalism, but rather to *derive* the formalism from a set of experimentally motivated postulates. [191]

L'élément crucial, présent dans les reconstructions mais pas dans les interprétations de la mécanique quantique, est la dérivation mathématique à partir de principes physiques. Le rôle des mathématique est donc central, et ces mathématiques sont souvent nouvelles ou atypiques, par rapport à celles qu'on employait déjà.

Par un heureux hasard, Max Born, contrairement à la majorité des physiciens de son époque, avait connaissance du calcul matriciel. Il en fit part à son ami, Werner Heisenberg, qui appliqua ce formalisme mathématique aux règles du calcul des spectres atomiques. La mécanique matricielle était née.

Quelques années plus tard, John von Neumann remplaça la mécanique matricielle, alors au cœur de la mécanique quantique en compagnie de la mécanique ondulatoire de Schrödinger, par la théorie des opérateurs dans l'espace de Hilbert. Bien que sa construction mathématique ait pris du temps à diffuser parmi les physiciens, elle l'emporta pour deux raisons. La première est que la théorie de Heisenberg, motivée empiriquement, y est incorporée dans un cadre mathématique général. La seconde tient en ce que ce cadre est apte à produire des prédictions nouvelles aussi en théorie des champs. Mais von Neumann, nous l'avons dit *supra*, ne s'arrêta pas à cette étape de ses recherches mathématiques. Il souhaitait remplacer son propre formalisme hilbertien par un autre, lié aux treillis de la logique quantique. Cette démarche, qui ne connut pas le même succès, s'inscrivait pourtant dans la même démarche d'introduction des outils mathématiques inédits en physique.

On trouve une tentative similaire dans l'histoire de la relativité, exemple paradigmatique s'il en est un de l'application réussie des nouvelles mathématiques. Pour paraphraser un adage célèbre d'Einstein, c'est la géométrie riemannienne qui « décide de ce qui peut être observé » empiriquement.

La formulation de la relativité générale que donne Einstein repose sur le tenseur métrique. Quelques années plus tard, Hermann Weyl publie une formulation mathématique différente, fondée sur la connexion dite « de Weyl ». Cette nouvelle manière de présenter la théorie ne produit pas de changement au niveau des observables. Pourtant, Weyl insiste sur la supériorité de son approche, arguant qu'elle bénéficie d'une force explicative supérieure à celle d'Eistein. Le philosophe Thomas Ryckman souligne : bien que Weyl abandonne graduellement toute ambition de remplacer l'approche d'Einstein par la sienne, il croit toujours que « bon nombre d'arguments esthétiques et philosophiques plaident en faveur de sa théorie plutôt que celle d'Einstein » [192, p. 160].

La leçon à tirer de cet épisode historique est double. Premièrement, les améliorations mathématiques visent à fournir un pouvoir explicatif supplémentaire, pas nécessairement une méthode de calcul plus facile. Partant de la déception de Weyl, il serait sage de ne pas exiger qu'un nouveau formalisme mathématique soit simple à utiliser. Secondement, il convient de se demander pourquoi, malgré la persévérance de Weyl, son approche n'est jamais devenue populaire, car ses mérites esthétiques et conceptuels sont évidents. Une explication sociologique facile serait de dire qu'Einstein, déjà très célèbre, dominait la communauté des physiciens. Mais, on peut aussi remarquer que le formalisme de Weyl était resté trop proche de celui d'Einstein. Il employait les mêmes mathématiques fondamentales, celles de Riemann, et produisait les mêmes

prédictions. La différence n'était ni manifeste ni mesurable. Si les mathématiques nouvelles peuvent un jour jeter une lumière nouvelle sur une théorie physique déjà connue, elles devraient pour cela faire usage d'un formalisme encore inédit.

C'est précisément cela que von Neumann souhaite accomplir lorsqu'il abandonne les espaces de Hilbert. Il choisit d'étudier la logique non-booléenne, encore jamais appliquée à la physique. Cependant, il échoue ; cinquante ans plus tard, on aura enfin saisi toute la difficulté de ce programme. En cause, entre autres, le problème de la composition des treillis orthomodulaires : Weyl, s'il avait vécu assez longtemps, aurait été horrifié par le manque de beauté et de sens esthétique d'une construction aussi artificielle que le crochet de Sasaki. Cet échec suggère que ce ne sont pas les reconstructions complètes qui ont quelque chance de réussir ; seules les reconstructions partielles jettent une lumière utile sur diverses composantes du formalismes quantique, un élément à la fois.

5.3 Naturalité dans les théories quantiques des champs

La section précédente se termine par une rapide évocation des considérations esthétiques en physique mathématique. En mécanique quantique, l'histoire des arguments esthétiques, d'ailleurs, souvent fallacieux, remonte aux travaux menés par Dirac dans les années 1930 [70, 71]. Un intérêt philosophique pour les modèles de la physique des hautes énergies nous a amené à étudier cette thématique dans le cadre des concepts de naturalité (*naturalness*) et du *fine-tuning* en théorie quantique des champs [105]. Cette étude s'inscrivait dans un projet plus vaste, réalisé en 2008 avec deux collègues du CEA-Larsim, consacré aux controverses philosophiques autour de la physique du LHC [104]. Dans le souci de limiter la taille de ce mémoire, nous résumons très brièvement l'argument principal de notre contribution.

La naturalité est souvent comprise — pour de bonnes raisons — comme une heuristique d'ordre esthétique. Dans la théorie des champs, pendant les débats des années 1970, postérieurs à l'invention du groupe de renormalisation mais antérieurs à la découverte des bosons faibles, 't Hooft relie la définition de naturalité à la notion de symétrie [206], tandis que Susskind, partant d'une idée de Wilson, préfère de concevoir la naturalité comme un critère de stabilité contre faible variations des paramètres fondamentaux [204]. À partir de cette époque, et jusqu'à la fin du XX^{e} siècle, trois principales mesures quantitatives voient le jour, permettant de saisir le degré de fine-tuning des modèles au-delà du Modèle standard (BSM). Ces mesures sont proposées par Barbieri et Giudice [15], Anderson et Castaño [7] et, en améliorant légèrement cette dernière conception, par Athron et Miller [11]. Tous ces auteurs visaient à représente par un nombre la plausibilité théorique des différents modèles BSM.

Plusieurs commentateurs du *fine-tuning* en théorie des champs en forment spontanément une probabilité. Des lignes entières d'argumentation sont élaborées à la base d'une telle heuristique probabiliste qui, comme nous l'avons montré, ne peut

être qu'une métaphore. Nous avons critiqué l'utilisation de la naturalité pour guider la recherche, en établissant des limites conceptuelles précises à tout emploi du *fine-tuning* en tant que probabilité.

La source principale du trouble réside dans l'impossibilité de parvenir à une normalisation de la mesure prétendument probabiliste. Nous avons proposé, à la place de cette lecture mal fondée, une interprétation du fine-tuning en tant que *Gedankenfrequenz*. Cela rejoint, d'ailleurs, les débats actuels sur la ressemblance entre la naturalité en cosmologie et en théorie des champs. Il nous semble que les deux concepts, sans se ressembler formellement, suivent tout de même une heuristique intellectuelle analogue.

Notre insistance qu'en théorie quantique des champs, la naturalité ne peut être nullement comprise comme un argument rigoureux, a été revendiquée par les évolutions des années 2010, à la suite de la découverte du boson de Higgs au LHC. À notre surprise, notre étude datant de 2008 connaît aujourd'hui un certain succès parmi les philosophes des sciences qui travaillent sur cette thématique.

5.4 Modèles indépendants du dispositif

5.4.1 Définition

Un modèle indépendant du dispositif est défini comme un ensemble de n « parties » ou « laboratoires locaux ». Chacune de ces parties « sélectionne » un paramètre de mesure ou « choisit » une entrée $x_1 \in \mathcal{X}_1, \ldots, x_n \in \mathcal{X}_n$ respectivement ; « ensuite », elle « obtient » une sortie ou un résultat de mesure $a_1 \in \mathcal{A}_1, \ldots, a_n \in \mathcal{A}_n$. Les ensembles $\mathcal{X}_1, \ldots, \mathcal{X}_n$ et $\mathcal{A}_1, \ldots, \mathcal{A}_n$ sont des alphabets finis. Les verbes utilisés dans ces expressions communiquent une signification purement opérationnelle, sans préjuger de la nature humaine d'une partie, de son libre arbitre ou d'un choix conscient. Le mot « ensuite » introduit une flèche du temps locale, propre à chaque laboratoire, qui pointe de son entrée vers sa sortie. Bien que l'existence d'une telle flèche paraisse intuitive, il est concevable qu'en toute généralité, elle ne soit pas postulée. Le cadre, par conséquent, ne pose aucune condition physique sur le modèle, sauf une : les données de types « entrée » et « sortie » doivent être bien distinctes. La physique est décrite par une distribution conjointe de probabilité $\mathbf{p} = P(a_1, \ldots, a_n | x_1, \ldots, x_n)$ (Figure 5.1).

Les modèles indépendants du dispositif que l'on trouve dans la littérature introduisent des contraintes supplémentaires sur $\mathbf{p}$, dont la plus fréquente est le principe de *non-signalement*, déjà évoqué plus haut. Ce principe signifie que le choix des paramètres de mesure par une partie n'a pas d'influence sur les statistiques de sortie d'une partie différente. Son importance est tellement grande qu'il est possible de démontrer son équivalence avec un principe manifestement différent, celui de l'existence d'une limite inférieure dans l'algorithme de recherche de Grover [14]. Sur le plan mathématique, une distribution de probabilité ne permet pas le signalement si et seulement si toutes les probabilités marginales, pour une seule partie, sont des

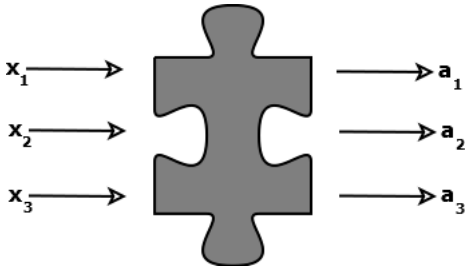

FIGURE 5.1 – Dans le cas de $n = 3$ parties, la physique est entièrement décrite par la distribution des probabilités $\mathbf{p} = P(a_1a_2a_3|x_1x_2x_3)$.

fonctions des entrées des parties respectives x_i :

$$P(a_i|x_1, \ldots, x_n) = P(a_i|x_i). \tag{5.3}$$

La contrainte de non-signalement, même si elle est très répandue, n'est pas universelle : par exemple, dans l'étude des ordres causaux indéfinis (section 5.5), elle apparaît en conséquence du respect de certaines inégalités, plutôt que comme un principe fondamental [21]. D'autres contraintes sur $\mathbf{p}$ sont envisageables, par exemple une condition de sécurité du *bit commitment* [4].

À l'instar des reconstructions partielles, les modèles indépendants du dispositif ouvrent un espace d'investigation qui permet de comparer la mécanique quantique avec ses « voisines » légèrement différentes. Cela nous incite — et c'est le principal apport philosophique de ces modèles — à poser des questions inédites à propos de la théorie quantique. Par exemple, bien que chaque partie possède une flèche du temps locale, l'existence d'une flèche du temps globale n'est pas assurée. En suivant Hardy [122, 123], Chiribella [55] a suggéré de prendre ce problème comme point de départ d'un nouveau modèle. Si ses premières constructions théoriques dépendait du choix de la mécanique quantique, plus tard Branciard et d'autres ont reformulé son interrogation sous une forme indépendante du dispositif. Nous discuterons *infra* du formalisme de la matrice de processus, introduit par Oreshkov, Costa et Brukner [169], qui a propulsé cette étude de la causalité quantique sur l'avant-scène des fondements de la physique [9, 40, 41, 8, 132, 20, 88].

Autre exemple, les corrélations « presque quantiques », étudiées par Navascuès *et al.* dans un modèle indépendant du dispositif [164, 163, 209], utilisent ces définitions générales :

> For Alice (respectively for Bob), an experiment is a process or black box to which she feeds an input x from the alphabet $\mathcal{X}$ and from which she receives an output a from the alphabet $\mathcal{A}$. Alphabets $\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}$ are of finite cardinality. [144]

Cette base conceptuelle est exemplaire. Contrairement à l'idée de Bohr sur le rôle de la langue naturelle, elle s'appuie sur un langage formel. Les modèles indépendants du dispositif sont définis par un choix d'alphabets pour les entrées et les sorties ; les

séquences de symboles puisés dans ces alphabets forment leur socle mathématique commun.

Célébré comme un « développement récent de grande importance » en physique quantique [185], les modèles indépendants du dispositif sont caractérisés, à l'instar des boîtes de Popescu-Rohrlich, par l'absence de présupposés à propos des processus internes, cachés dans la boîte. Quoique les définitions soient formulées à l'aide d'un langage formel, l'espace intérieur des boîtes reste physique ; or, la théorie qui le décrit demeure inconnue.

5.4.2 Problème de confiance pour l'observateur

L'introduction par John Bell [25] des inégalités qui quantifient la force des corrélations quantiques fut une véritable « deuxième révolution quantique » [10]. Toutefois, il a fallu attendre quelques décennies pour que les conséquences des inégalités de Bell sur le développement de la théorie quantique deviennent apparentes. Dans un article récent [111], nous suggérons que le sommet de ces développements est atteint, de manière quelque peu inattendue, avec l'avènement des approches indépendantes du dispositif. Elles inaugurent, en effet, un point de vue différent sur la théorie physique en reléguant au second rôle la notion de système, au profit de celle d'observateur.

Dans sa formulation standard, la mécanique quantique décrit l'évolution du système sous l'action d'un Hamiltonien ainsi que les résultats des mesures que l'observateur effectue sur ce système. Le concept d'observateur est extérieur à la théorie : quelle que sa composition matérielle, physique, l'unique rôle qu'il joue en mécanique quantique est de choisir les paramètres de mesure et d'enregistrer son résultat. Une approche à l'observation fondée sur les entrées et les sorties, nous l'avons dit, est une approche *opérationnelle*.

Or, croire que les corrélations entre les choix des paramètres de mesure et ses résultats sont dues à des médiateurs, porteurs de l'information, que l'on appelle « systèmes », n'est qu'intuitif. Une conception possible de ces médiateurs voit en eux des « lignes » ou des « fils » dans les diagrammes symboliques, qui relient différentes opérations avec l'information de l'observateur. Cette conception moderne mène à « des modes nouveaux d'explication des phénomènes physiques » [60, 61, 62]. Cependant, il existe un ancien mode explicatif, qui donne une autre signification aux systèmes. Ainsi, on conçoit les systèmes par la séparation de leurs degrés de liberté de ceux appartenant à des non-systèmes, par exemple aux appareils de mesure ou à l'environnement. Un système est compris comme un bouquet de degrés de liberté pertinents, qui est désigné par un nom unique. L'idée de séparation joue ici un rôle crucial, or elle *ne définit pas* le système ; seulement, elle rend possible une *explication* de ce qu'il est. L'impossibilité d'utiliser une démarche de séparation pour *définir*, mais seulement pour *expliquer*, possède une longue histoire philosophique ($\delta\iota\epsilon\tilde{\iota}\lambda\epsilon\nu$ dérivé de $\delta\iota\alpha\iota\rho\acute{\epsilon}\omega$, séparer, Platon *Timée* 41d).

L'ancien mode explicatif n'est pas l'unique façon de concevoir des systèmes physiques ; il ne s'applique pas aux approches indépendantes du dispositif. Selon le nouveau mode, le concept de système voit son utilité limitée. Pourtant, cela ne nie pas

que notre intuition soit bien fondée dans plusieurs situations physiques, celles que l'on décrit habituellement dans les termes d'évolution des systèmes. Mais, il existe des cas — et c'est l'élément contre-intuitif — dans lesquels il est légitime, au sein même d'une théorie physique, de poser la question des conditions sous lesquelles une description en termes de systèmes devient possible. Ainsi, la formulation d'un problème physique ne commence pas avec la liste des systèmes concernés, mais avec celle des entrées et des sorties ; on cherche à comprendre les contraintes qui donneraient sens à l'interprétation de cette situation comme relevant des systèmes.

En remontant vers l'origine des difficultés que pose la notion de système, on se heurte au problème de confiance en cryptographie quantique. Selon le formalisme standard de la mécanique quantique, il est nécessaire de supposer, au début d'une évolution unitaire, que les paramètres de préparation et de mesure sont sélectionnés « honnêtement ». Cela signifie, pour l'observateur, qu'il peut faire confiance à la préparation du système qu'il va mesurer, et au catalogue des degrés de libertés qui constituent ce système. Pour résumer, l'observateur quantique sait quel est son système, et il n'en doute pas.

Par exemple, une mesure binaire de la polarisation d'un photon présuppose que les systèmes mesurés soient réellement des photons. Un observateur qui effectue ces observations ne s'attend pas à trouver autre chose. D'habitude, cette confiance va de soi et n'est pas questionnée, même si elle pourrait se révéler tout à fait infondée.

En cryptographie quantique, le manque de confiance dans la préparation des systèmes est étudié en tant que problème théorique et empirique. On y trouve des outils pour travailler avec des systèmes à « caractère non-spécifié » [13] dont la « nature [est] inconnue » [16]. Une approche indépendante du dispositif emploie ces mêmes outils : il s'agit de mener une étude théorique, sans présupposer la connaissance des systèmes qui fonctionnent comme les médiateurs dans l'expérience. Le terme de *dispositif* renvoie ici à tout processus ou appareil dont la description est fournie de manière opérationnelle, qu'il soit classique ou quantique. Dans ce sens, les dispositifs incluent, non seulement des appareils spécifiques de cryptographie ou des tables d'optique quantique, mais aussi des objets plus étranges, comme les courbes fermées d'espèce temps [68, 27] ou les espaces-temps de Malament-Hogarth [75, 127].

Le terme « modèle indépendant du dispositif » (*device-independent model*) fut introduit par Mayers et Yao [158] à des fins de cryptographie en présence des sources imparfaites. Depuis, les avancements de la science ont été si nombreux qu'il est désormais possible, par exemple, de poser une question impensable dans le cadre de la mécanique quantique standard : « Alice et Bob devraient-il faire confiance à eux-mêmes ? ». En effet, on démontre qu'il est suffisant, pour que cette confiance en soi soit rétablie, qu'une infime partie des choix d'Alice ou de Bob ne soit pas contrôlée par un adversaire [63, 83]. Si ces choix sont réellement aléatoires — autrement dit, si l'observateur demeure un petit peu libre —, alors des méthodes cryptographiques permettent de restaurer la confiance dans les sources. Le rôle que joue ici le hasard est fondamental ; nous y reviendrons.

Les méthodes initialement développées en cryptographie trouvent plusieurs usages en information quantique ; par exemple, on effectue des tests indépendants de dispo-

sitif des inégalités de Bell. L'apport conceptuel et philosophique de ces approches s'étend aujourd'hui à l'ensemble de la théorie quantique. Quel est cet apport ? En transformant la confiance en un problème théorique, les approches indépendantes du dispositif effacent le dogme central de la mécanique quantique, celui-là même qui place un système prédéfini au cœur de la description théorique. Pour apprécier l'importance de cette évolution, nous la comparons avec un autre glissement paradigmatique : l'avènement, dans la physique du XX^{e} siècle, des théories dites « principielles ».

5.4.3 Qu'est-ce une théorie physique ?

Depuis un siècle, toute discussion du sens de la théorisation en physique ne peut faire abstraction d'une distinction, proposée par Einstein en 1919 [77], entre les théories principielles et constructives. La relativité, exemple paradigmatique de théorie principielle, repose sur quelques principes universels, non sur des lois régissant le comportement d'un type particulier de matière. Einstein dira plus tard que ces principes servent à « limiter le champ du possible » [76]. Une théorie constructive, au contraire, s'applique à une espèce circonscrite d'objets physiques.

La distinction einsteinienne s'inscrit dans la lignée historique d'une approche, proposée par Helmholtz, visant, plutôt que de poser une ontologie de particules chargées, à « contraindre » la théorie électrodynamique de la même manière que les principes contraignent une théorie principielle. Helmholtz souhaitait établir une théorie différente du modèle précédent des interactions électriques et magnétiques. Pour cela, poser l'existence des « atomes de charge » était contre-productif ; il travaille ainsi, uniquement avec la fonction de potentiel, assujettie à la loi de conservation de l'énergie. L'historien Jed Buchwald décrit la méthode de cette « nouvelle physique » [48] :

> Indeed, the longest-lasting and deepest assumption of Helmholtz's work in electrodynamics from 1847 until the late 1880s, and even after, was precisely that Neumann's potential U_d (or something very like it) is a fundamental quantity. There are two general (and connected) reasons for his conviction in the potential's importance. First, and of overriding importance, it has immediate energetic significance. Second, an electrodynamics based upon U_d, unlike Fechner-Weber, remains independent of assumptions concerning electric atoms and their concomitant forces; *it acts as a constraint* upon any model purporting to explain charge and current. It is, therefore, a much more general theory that Fechner-Weber, and, like energy conservation itself, it can survive extreme changes in the higher-order principles of electrodynamics. [49, p. 9, nous soulignons]

Pour interpréter le point de vue de Helmholtz, il est possible de soutenir qu'il refusait entièrement l'idée d'une ontologie : "The idea was that this framework could serve as a kind of book-keeping device for recording the results of his experiments without committing to any particular ontology regarding the ultimate nature of electrodynamic action" [89]. Ce point de vue nous semble tout de même osé pour son époque. Mais, il s'agit indubitablement d'une attitude méthodologique nouvelle qui a probablement

contribué à la formation épistémologique du jeune Einstein.

La relativité restreinte n'est pas une théorie constructive, car elle ne dit mot au sujet de la constitution matérielle des horloges et des règles, qui fonctionnent comme ses instruments de mesure. Cela ne nous choque guère. Or, Einstein, créateur de la théorie, croyait que ce manque de caractère constructif était un défaut. Selon son point de vue des années 1920, les théories principielles ne pouvaient pas, à elles seules, fournir une compréhension satisfaisante de la physique [37]. Il aurait fallu les remplacer par une théorie constructive :

> When we say we have succeeded in understanding a group of natural processes, we invariably mean that a constructive theory has been found which covers the processes in question. [77]

Toutefois, le désir d'Einstein de remplacer, pour des raisons philosophiques, la relativité restreinte par une théorie constructive n'a jamais été réalisé. Il est tentant de suggérer qu'une approche indépendante du dispositif suivra le même sort. Il est concevable que le souhait d' « ouvrir la boîte », pour décrire ce qui se passe à l'intérieur, soit une illusion épistémologique, dans la même mesure où le fut l'insistance d'Einstein : aujourd'hui, nous voyons bien qu'avec le temps, sa croyance en la force explicative des seules théories constructives s'est révélé illusoire. Il est possible que l'approche indépendante du dispositif soit un jour fermement ancrée comme une façon légitime de construire une théorie physique, sans qu'il soit nécessaire de « remplir la boîte ».

Cependant, les approches indépendantes du dispositif inaugurent un départ plus radical encore de la physique « matérielle » que les théories principielles. Ces dernières présupposent, tout comme les théories constructives, qu'on élabore une théorie à l'aide de systèmes physiques ; seulement, dans les théories principielles, ces entités sont théoriques, et non matérielles. Aucun des deux types des théories n'envisage la possibilité qu'une théorie puisse être construite sans que soient d'emblée posées des entités constitutives, ni qu'une théorie puisse permettre de dériver les conditions de se propre interprétation comme étant « à propos de systèmes ». Einstein n'a certainement pas pensé à une telle possibilité. En 1949, il remarque que les impressions sensorielles, d'abord perçues et ensuite conceptualisées, sont de mauvaises candidates pour « meubler » la réalité, tout en ajoutant que l'existence des entités relève, elle, d'une nécessité théorique :

> A basic conceptual distinction, which is a necessary prerequisite of scientific and pre-scientific thinking, is the distinction between "sense-impressions" on the one hand and mere ideas on the other. There is no such thing as a conceptual definition of this distinction (aside from circular definitions). Nor can it be maintained that at the base of this distinction there is a type of evidence, such as underlies, for example, the distinction between red and blue. Yet, one needs this distinction in order to be able to overcome solipsism. Solution: we shall make use of this distinction unconcerned with the reproach that, in doing so, we are guilty of the metaphysical "original sin." [80]

Les approches indépendantes du dispositif vont plus loin que cela. Les « impressions sensorielles », sous forme d'entrées et de sorties, ne nécessitent pas qu'une notion de système soit posé *a priori*.

Dans le même temps, ces remarques d'Einstein contiennent une pensée profonde : il n'est pas possible de donner une *définition* rigoureuse de système. Les notions théoriques fondamentales ne peuvent être définies à partir de données phénoménales qu'au prix d'une certaine circularité, parce que notre interprétation de ces données fait déjà intervenir des concepts théoriques. Il est alors nécessaire de renoncer à l'ambition de *définir* formellement un système. Toutefois, il est possible d'*expliquer* ce qu'est un système, sans qu'une telle explication devienne une définition. Einstein pensait qu'une explication serait un prérequis de toute théorie ; les approches indépendantes du dispositif y donnent un contre-exemple. Cet argument einsteinien, risqué, est explicitement kantien :

> It is also the presupposition of every kind of physical thinking. Here too, the only justification lies in its usefulness. We are here concerned with "categories" or schemes of thought, the selection of which is, in principle, entirely open to us and whose qualification can only be judged by the degree to which its use contributes to making the totality of the contents of consciousness "intelligible." [80]

Toute l'histoire des interprétations kantiennes des théories physiques [67, 31, 175] montre que, souvent, un présupposé catégorique perd son caractère *a priori* suite à l'évolution des théories physiques. Il devient alors objet légitime d'enquête théorique. Michael Friedman parle dans ce sens d'un *apriori relativisé* [92].

C'est cette relativisation de la notion métathéorique de système, et conjointement de celle d'observateur, que nous effectuons dans notre programme de recherche. L'évolution concernée : la possibilité de construire une théorie physique sans que soit d'emblée supposée l'existence des systèmes. L'origine de cette évolution : les modèles indépendants du dispositif, dont la nature des données initiales, suites symboliques d'entrée et de sortie, nous incite à élaborer une conception formelle de l'observateur.

5.4.4 Une conception formelle de l'observateur

Dans un argument bien connu, Einstein proclame que la mécanique quantique ne saurait être complète à moins que la fonction d'onde ψ ne décrive « l'état réel du système réel » [79]. Les inégalités de Bell font planer un grand doute, définitivement confirmé par les expériences d'Alain Aspect, sur la possibilité d'adhérer à la position einsteinienne. Les approches indépendantes du dispositif l'évacuent définitevement.

Or, au début des années 1930, peu de scientifiques en dehors de Copenhague pensaient qu'Einstein pourrait avoir tort en insistant sur la réalité des états du système. Ainsi, Boris Podolsky ouvre l'article EPR par la célèbre introduction des « éléments de réalité », ce qui, d'ailleurs, ne satisfait pas pleinement son co-auteur Einstein [130] :

> Any serious consideration of a physical theory must take into account

> the distinction between the objective reality, which is independent of any theory, and the physical concepts with which the theory operates. These concepts are intended to correspond with the objective reality, and by means of these concepts we picture this reality to ourselves. [82]

À la même époque, Dirac avance une idée similaire, à ceci près qu'à la place de l'ambitieuse « réalité objective », il parle d' « entités physiques », conceptuellement plus modestes :

> "The most powerful advance would be to perfect and generalize the mathematical formalism that forms the existing basis of theoretical physics, and after each success in this direction, to try to interpret the new mathematical features in terms of physical entities." [69]

Les approches indépendantes du dispositif rendent ces anciennes positions obsolètes, car elles ouvrent une possibilité nouvelle : une théorie peut ne pas contenir de systèmes. Rien dans le formalisme ne permet alors de concevoir la réalité comme étant composée d'entités physiques.

Cela ne veut pas dire pour autant que l'approche indépendante du dispositif cesse de nous renseigner sur la physique. Seulement, dans un revirement radical, elle propose de construire une théorie physique formée exclusivement d'entrées et de sorties, sans qu'elle contienne une entité chargée de faire la médiation entre les deux.

Ce revirement est plus puissant que le traditionnel rejet du réalisme des entités. Ce dernier désigne un point de vue selon lequel les entités physiques existent réellement et objectivement ; Steven French constate, déjà en 1998, son « withering away » [91]). Dans une approche indépendante du dispositif, il est concevable que les systèmes n'existent ni dans la « réalité » ni dans la théorie. Leur place dérivée, parfois impossible à dériver, représente une innovation.

Ce n'est pas uniquement le réalisme des entités que rejettent les approches indépendantes du dispositif. Un autre point de vue, bien ancré dans le travail quotidien des physiciens, perd également de son attrait. Il s'agit d'un réalisme dit « naïf » : une heuristique non-savante, qui stipule que les objets de toute enquête expérimentale, tels électrons ou photons, sont réels, parce que l'homme forme cette croyance de façon empirique et qu'elle est cohérente avec son travail de laboratoire.

De cette prétendue cohérence, plusieurs développement au sein de la mécanique quantique permettent cependant d'en émettre quelques réserves. La dualité onde-particule, ensuite les relations d'indétermination d'Heisenberg et enfin la contextualité de Kochen-Specker interdisent toute éventualité d'un cadre philosophique réfléchi qui soit compatible avec le réalisme naïf. Les approches indépendantes du dispositif l'achèvent formellement.

Bien que cette nouveauté épistémologique, que représente le manque de systèmes, contribue à rendre ces modèles intéressants, cela complique davantage leur réalisation expérimentale. On ne peut construire une « boîte » dans un laboratoire qu'avec des systèmes connus, tels photons. Effacer cette connaissance requerrait un esprit assez tortueux. Cependant, le modèle l'exige. Il faut mettre entre parenthèses toute connaissance des systèmes pour peu que l'approche ne souffre d'une dépendance du dispositif

utilisé. Par exemple, il est interdit de se servir même des détails insignifiants du dispositif quantique pour régler les paramètres de mesure, car cela pourrait engendrer une violation du principe de non-signalement [182, 199]. Pourtant, cette démarche est naturelle pour un expérimentateur ; mais elle est prescrite par la philosophie des approches indépendantes du dispositif.

La position philosophique la plus proche des modèles indépendants du dispositif, pourtant antérieure à leur création, est probablement celle exposée par John Wheeler. Il écrit dans son journal privé :

> On this view physics is not machinery. Logic is not oil occasionally applied to that machinery. Instead everything, physics included, derives from two parents, and is nothing but cathode-tube image of the interplay between them. One is the "participant." The other is the complex of undecidable propositions of mathematical logic. ... The propositions are not propositions about anything. They are the abstract building blocks... [217]

La « machinerie » correspond ici, sans doute, à la notion einsteinienne de théorie constructive. À première vue, Wheeler ne fait que l'attaquer. On aurait pu se contenter d'une telle lecture en faveur des théories principielles, si ce n'était pour la phrase suivante : selon Wheeler, les propositions ne se réfèrent à « rien ». Cela signifique qu'elles ne contiennent pas de sémantique ; toutefois, elles sont des briques élémentaires d'une conception de la réalité. Bien que celle-ci soit "as full-blown as anyone could want" [94], elle naît d'un modèle purement formel et abstrait, en conformité avec la philosophie d' « observateurs-participants » de Wheeler [219, 218].

Cela représente la première occurrence d'un point de vue qui réunit ces deux composantes : d'un côté, les propositions sont des éléments basiques, de l'autre, ces mêmes propositions ne se réfèrent à « rien ». Il n'existe derrière elles aucune entité, empirique *ou théorique*. Les approches indépendantes du dispositif, sur le plan philosophique, ne font que prolonger la vision de Wheeler ; seulement, son expression « proposition indécidable » sera remplacée par les suites d'entrée et de sortie.

Reprenons, cette fois avec Wheeler, l'idée, dont les origines remontent à von Neumann, à savoir que l'on doit placer, au fondement de la théorie quantique, un modèle mathématique abstrait. Wheeler souligne la nécessité de respecter une certaine économie conceptuelle en gardant seulement un petit nombre de présupposés. Les autres, y compris la structure de l'espace de Hilbert, devraient émerger d'un « traitement » des axiomes fondamentaux :

> 2. Start with what formal system?
> Take a formal system. Enlarge it to a new formal system, and that again to a new formal system, and so on, by resolving undecidable propositions ("act of participation"). Will the system become so complex that it can and must be treated by statistical means? Will such a treatment make it irrelevant, or largely irrelevant, with what particular formal system one started? [217]

Le point de vue philosophique promulgué par les approches indépendantes du dispositif appartient à cette lignée méthodologique. Dans une théorie physique, il s'agit

d'abord d'une couche opérationnelle et, mathématiquement parlant, algébrique. À partir d'elle peuvent être constituées — mais pas toujours —, des entités théoriques.

Wheeler n'utilise pas le mot « physique » dans ses notes : le sens de ce mot implique usuellement une interprétation sémantique et une référence. En transposant son point de vue que les « propositions are not about anything », suggérons que les modèles indépendants du dispositif ne se réfèrent à aucune entité, même théorique. L'interprétation sémantique, sans être fausse, n'est pas requise ; dans la mesure du possible, on l'adjoint à une étape ultérieure, en fonction de sa cohérence mathématique et de son utilité pour appréhender les résultats. Si on la retire — et certains modèles ne sont pas compatibles avec sa présence —, alors il est nécessaire de proposer une réponse alternative au problème ontologique dégagé par Wheeler : de quoi parle une théorie, à quoi se réfère-t-elle ?

Il est tentant de répondre, comme proposé par certains auteurs [220, 93, 45], que la théorie physique se réfère à l'information, ou à un type particulier d'information. Cette réponse, bien qu'elle soit compatible avec notre point de vue, ne nous semble pas satisfaisante, car elle ne résout pas la principale difficulté philosophique : si l'information est une substance fondamentale, existe-t-il des types ou des variétés d'information ? Si l'entropie de Shannon et l'entropie de von Neumann fournissent deux concepts d'information différents, quelle serait l'origine d'une telle multiplicité pour une notion qui se veut fondamentale ?

Une solution alternative serait de stipuler l'unité de tous les types d'information obéissant à la définition de Shannon [195], et d'interpréter les autres comme dérivés ou comme étant autre chose que l'information fondamentale. Cette solution peut être soutenue, or sa défense dépasserait les objectifs de ce mémoire.

Revenons à l'idée scientifiquement plausible du maintien d'une claire distinction entre l'information classique et quantique. Cette dernière peut être définie, par exemple, comme une information qu'on ne peut pas cloner [46]. Si la théorie quantique s'y réfère, il serait juste de demander qui ou quoi effectue le clonage. Évidemment, la description de cet agent n'est pas donnée par la théorie, or celle-ci se veut universelle. L'objection est standard mais valide. Toute interprétation du référent théorique en tant qu'information fondamentale, à la place d'une définition opérationnelle de l'observateur dans un schéma principiel, y laisse la porte ouverte.

L'interprétation sémantique du référent théorique n'étant pas une nécessité, nous considérons cette objection comme une illusion épistémologique, parce qu'elle confond la théorie principielle avec la théorie constructive. Lorsqu'on demande de quoi parle une théorie physique, la seule voie légitime de réponse est celle d'une reconstruction, même partielle, des traits saillants de cette théorie. À son tour, tout reconstruction s'appuie sur des principes. Dans le cas de la mécanique quantique, il s'agit d'une approche indépendante du dispositif qui permet d'explorer l'origine de ses deux caractéristiques fondamentales : la continuité et la règle de composition. Leur motivation est donnée par des contraintes que l'on a mises sur la distribution conjointe de probabilité $\mathbf{p}$ et qui nous renseignent sur le dénominateur philosophique commun du modèle en question et de la théorie quantique, non sur une ontologie informationnelle de la Nature.

Les modèles indépendants du dispositif opèrent avec des suites de symboles dans un alphabet fini. Leur interprétation immédiate est opérationnelle : il s'agit des entrées et des sorties. À l'objection à propos d'un « agent-observateur » qui enregistre ces données, nous répondons à travers une analogie avec la théorie de Shannon. Comme dans cette dernière, il est possible, dans un modèle indépendant de dispositif, d'étudier les entrées et les sorties d'une manière purement formelle, sans leur appliquer une sémantique. Quel que soit leur référent empirique ou théorique, ou même qu'il n'existe pas, la question de référence n'a pas besoin d'être posée. Une reconstruction qui éclaire le sens de la mécanique quantique peut procéder de façon asémantique.

On nous objecterait, non sans quelque justification intuitivement valide, que la lecture sémantique des suites de symboles devient une nécessité dans la mesure où il s'agit d'élaborer une théorie physique, et non seulement mathématique. Si une théorie est physique, c'est qu'elle décrit bien la Nature : elle décrit *quelque chose*.

Notre réponse : l'intuition sur laquelle s'appuie cette objection n'est pas encore suffisamment élaborée. La véritable contrainte minimale, suffisante pour établir un lien avec l'expérience, est de nature opérationnelle : la théorie se réfère, non à une chose, mais aux entrées et aux sorties qu'enregistre un observateur. Sur le plan formel, il est constitué par les suites de symboles dans un alphabet fini. Sa composition matérielle, quelle qu'elle soit, n'est pas pertinente.

Nous avons déjà dit que l'emploi des langages formels ne ressemble que de loin à la préoccupation de Bohr par la langue naturelle. Celle-ci n'est pas au centre de notre intérêt car les approches indépendantes du dispositif sont des modèles algébriques, non pas linguistiques. Mais, nous revendiquons tout de même un certain lien de parenté avec la pensée de Bohr.

Nous avons déjà évoqué Everett opposant à Wigner et autres son idée de définir l'observateur à un niveau abstrait, par la présence d'une mémoire. Nous avons détaillé ce point de vue [106, 109] en utilisant les méthodes de la théorie de l'information algorithmique de Kolmogorov [140].

Le recours en physique aux idées de Kolmogorov n'est pas une nouveauté malgré le manque de résultats concrets d'une telle démarche. La première tentative date des années 1980, lorsque Zurek tente d'associer, à la description de l'état d'un système physique, une contribution de la complexité algorithmique, sous forme d'un terme dans l'expression d'entropie [227, 226]. Son approche ne s'inscrivait pas dans le cadre d'une reconstruction, mais présupposait la validité de la mécanique quantique.

Notre modèle est construit sur une base conceptuelle différente. Notre première tentative de recourir à la complexité de Kolmogorov prit la forme du problème d'identification des systèmes par l'observateur. Là où Zurek parlait des états, nous voyions un besoin plus fondamental, celui d'identification et de maintien de l'identité des systèmes que décrivent ces états. Cela revient à sélectionner un ensemble de degrés de liberté en tant que variables pertinentes et à préserver cette sélection pendant le temps d'une expérience. Ainsi, dans le cadre de cette première tentative de théorisation, nous avons formulé une définition d'observateur :

Definition 5.2. An observer is a system identification algorithm (SIA) [106].

Nous nous sommes également proposés d'approfondir cette définition, notamment en posant une limite supérieure de la complexité d'un tel algorithme. En pratique, c'est la taille de la mémoire l'observateur qui joue ce rôle. Son existence est un trait fondamental, caractéristique, de l'observateur : ceux d'entre eux qui possèdent la même limite sont, en quelque sorte, équivalents. Cette équivalence permet, d'ailleurs, de retrouver le problème d'accord intersubjectif, posé *supra* dans le cadre des interprétations relationnelles ; et même de dériver les conditions d'émergence d'une description objective.

Quelques années après ce premier essai, débuta notre travail sur les approches indépendantes du dispositif. Il nous incita, en particulier, à réviser l'idée du caractère fondamental de la notion de système. Toutefois, l'observateur est toujours défini par la taille de la mémoire, qui correspond désormais à la complexité algorithmique maximale des suites d'entrée et de sortie. Différents observateurs peuvent toujours posséder des mémoires de taille différente. Par ailleurs, il serait intéressant d'explorer les variations dans la description théorique que cela induit.

Nous sommes donc prêts à donner une nouvelle définition de l'observateur, cette fois dans le langage des approches indépendantes du dispositif :

Definition 5.3. An observer is a set of strings in a finite alphabet having possibly different, or even infinite, lengths but a uniformly limited Kolmogorov complexity.

5.5 Ordres causaux

5.5.1 Définition

La causalité est au centre des débats en sciences et en philosophie depuis Platon et Aristote. Au XX^{e} siècle, Russell, Reichenbach et Carnap, parmi beaucoup d'autres, ont fait des contributions importantes à son étude, tandis que van Fraassen et Suppes y appliquèrent des arguments empiriques, informés par la méthode scientifique. Ce débat fait toujours rage.

La causalité est, en effet, la pierre angulaire de l'explication scientifique. Des mots ordinaires comme « compréhension » ou « explication » demeurent essentiellement vagues ; toute tentative de leur attribuer un sens rigoureux rencontre de graves difficultés. Cependant, une relation de cause à effet peut être exprimée formellement ; cela distingue la causalité de ses consœurs épistémologiques. Il ne serait pas exagéré de dire que ce cas de fertilisation mutuelle de la physique mathématique et de la philosophie des sciences est unique.

La majorité des modèles causaux prennent comme structure de départ celle d'un ordre sur les événements. Les influences de certains événements sur d'autres permettent de les classer en deux groupes : d'un côté, les entrées, choisies librement et sans contrainte ; de l'autre, les sorties ou les résultats de mesure.

Formellement, les entrées et les sorties peuvent être représentée par les sommets d'un graphe orienté et acyclique ; l'absence de cycles correspond à l'impossibilité de

boucles causales. Un graphe orienté peut être interprété en tant qu'ordre partiel : un *ordre causal* sur les événements.

Cet ordre causaul existe dans la majorité des situations théoriques que l'on souhaite étudier. Cela se justifie, en particulier, par la théorie de la relativité. L'absence de signalement entre les laboratoires, séparés dans l'espace, est représentée par des intervalles d'espèce espace, entre les événements non liés par un lien de causalité.

Or, pour décrire les ordres causaux, on mobilise aussi la théorie quantique. Elle ajoute un élément nouveau : la superposition des ordres causaux. Cette découverte récente permet de concevoir des structures non-classiques de causalité indéterminée.

Quels ordres causaux sont compatibles avec un choix particulier des entrées et des sorties ? Un cadre qui permet de répondre naturellement à cette question est celui des approches indépendantes du dispositif. Toutefois, historiquement, ces approches n'ont été intégrées à l'étude de la causalité quantique qu'à partir de 2015 [22]. Notre travail théorique sur les ordres causaux est antérieur à cette date et présuppose encore la mécanique quantique, mais notre analyse philosophique en tient déjà compte.

Les reconstructions axiomatiques de la mécanique quantique assument, le plus souvent implicitement, que les événements expérimentaux sont ordonnés dans le temps, en posant ainsi un ordre causal global. Par exemple, Hardy [121] utilise les notions de préparation, transformation et mesure (PTM), dont la séquence fournit un ordre global. Ce présupposé est apparent dans la manière dont les systèmes se composent : selon l'axiome H4 (section 5.2.2), les paramètres K_{AB} et N_{AB} du système composé, définis de manière opérationnelle, s'expriment en termes des paramètres relevant seulement des sous-systèmes A and B :

$$N_{AB} = N_A N_B, \quad K_{AB} = K_A K_B. \tag{5.4}$$

Cela implique que seul le superobservateur, qui peut accéder à A and B conjointement, soit capable de calculer K_{AB} et N_{AB}. L'ordre PTM de ce superobservateur est global : une difficulté dont Hardy est lui-même bien conscient [122, 123, 124]. Partant de cette observation, Chiribella *et al.* [56] et Oreshkov *et al.* [169] proposent dans deux articles fondamentaux une nouvelle direction de recherche dans les fondements de la physique, consacrée à la superposition des ordres causaux que permet le formalisme quantique. Dans notre travail, le formalisme adopté est celui d'Oreshkov *et al.*

Supposons que $\mathcal{H}^{A_1}$ et $\mathcal{H}^{A_2}$ soient les espaces de Hilbert de l'entrée et de la sortie chez Alice et que, au laboratoire de Bob, les structures correspondantes soient $\mathcal{H}^{B_1}$ and $\mathcal{H}^{B_2}$. Toutes les sorties possibles d'une mesure dans un laboratoire se décrivent à l'aide d'opérateurs complètement positifs, $\{\mathcal{M}_i^{A_1A_2}\}_{i=1}^n$ et $\{\mathcal{M}_j^{B_1B_2}\}_{j=1}^n$. L'isomorphisme de Choi-Jamiołkowski, qui exprime la dualité entre les états et les canaux, permet d'exprimer ces processus sous forme des états sur un espace de Hilbert plus grand : la transformation $\mathcal{M}_i^{A_1A_2} : \mathcal{L}(\mathcal{H}^{A_1}) \longrightarrow \mathcal{L}(\mathcal{H}^{A_2})$ pour Alice devient $M_i^{A_1A_2}$ sur $\mathcal{H}^{A_1} \otimes \mathcal{H}^{A_2}$. De même, pour Bob la transformation $\mathcal{M}_j^{B_1B_2} : \mathcal{L}(\mathcal{H}^{B_1}) \longrightarrow \mathcal{L}(\mathcal{H}^{B_2})$ devient $M_j^{B_1B_2}$ sur $\mathcal{H}^{B_1} \otimes \mathcal{H}^{B_2}$. La probabilité conjointe s'écrit alors :

$$P(\mathcal{M}_i^{A_1A_2}, \mathcal{M}_j^{B_1B_2}) = \mathrm{Tr}\left[W^{A_1A_2B_1B_2}\left(M_i^{A_1A_2} \otimes M_j^{B_1B_2}\right)\right], \tag{5.5}$$

où $W^{A_1A_2B_1B_2} \in \mathcal{L}(\mathcal{H}^{A_1} \otimes \mathcal{H}^{A_2} \otimes \mathcal{H}^{B_1} \otimes \mathcal{H}^{B_2})$ est fixe : c'est une *matrice du processus*. Elle généralise la notion d'état pour un espace de Hilbert comprenant toutes les corrélations possibles entre deux ou plusieurs laboratoires physiques localisés.

Les conditions standards sur les probabilités (elles doivent être supérieures ou égales à zéro et leur somme est égale à 1 pour tout choix d'opérateurs locaux) donnent un sens précis à cette matrice de processus. On présuppose aussi, bien entendu, la validité de la mécanique quantique. Toutefois, aucun ordre global de causalité n'est nécessaire. Il se révèle que certaines corrélations violent une inégalité analogue à celle de Bell, ce qui indique l'impossibilité de les classer dans un ordre de causalité classique. Leur borne supérieure quantique, analogue, à son tour, à la borne de Tsirelson de l'inégalité CHSH, correspond à la limite maximale de la violation du caractère classique d'un lien de causalité entre deux laboratoires [39].

5.5.2 Jeu causal

Afin de conclure au caractère non-classique d'une situation causale, on emploie comme outil principal un jeu. L'importance conceptuelle de ce *jeu causal* est cruciale, car, en information quantique, on établit à travers les jeux (non-locaux, contextuels ou autres) qu'il existe une différence entre une ressource classique et une ressource quantique. Le jeu est aussi facilement compréhensible pour un public non-expert et, à ce titre, il devient la principale méthode pédagogique de l'information quantique, comme souligné dans notre article consacré à cette thématique [112].

Dans le jeu, tout comme dans le modèle des boîtes de Popescu-Rohrlich, deux parties, Alice et Bob, possèdent chacune un bit d'entrée, a et b. Ils partagent aussi un bit spécial, $b\prime$, qui régit les règles du jeu. Si $b' = 0$, Alice doit deviner la valeur de b ; mais si $b' = 1$, c'est à Bob de deviner la valeur de a. Les sorties sont notées x pour Alice et y pour Bob.

Il est essentiel que toutes les entrées, a, b et $b\prime$, soient aléatoires. Opérationnellement, cela signifie qu'elles ne subissent aucune influence au cours du jeu : leur provenance relève du « libre arbitre ».

L'objectif du jeu, pour Alice et Bob conjointement, est de maximiser la probabilité de réussite :

$$P_{success} = \frac{1}{2}\left[p(x = b|b' = 0) + p(y = a|b' = 1)\right]. \tag{5.6}$$

Cela signifie qu'Alice devine avec succès l'entrée de Bob, ou vice versa, en fonction de la valeur de b'.

Si, dans un premier temps, les événements se produisent selon un ordre causal déterminé, quel qu'il soit, alors la probabilité de réussite est bornée par :

$$P_{success} \leq \frac{3}{4}. \tag{5.7}$$

En effet, classiquement, il est vrai que soit Bob ne peut signaler à Alice (Alice est avant Bob), soit c'est le contraire. Considérons le premier cas. Si $b' = 1$, rien n'empêche a *priori* qu'Alice et Bob atteignent, par un effort conjoint, la valeur $p(y = a|b' = 1) = 1$.

Cependant, si $b' = 0$, Alice ne peut faire mieux que deviner au hasard la valeur de l'entrée de Bob, soit $p(x = b|b' = 0) = \frac{1}{2}$. La probabilité de réussite satisfait alors à (5.7). Le même argument est valable dans le cas où Bob se trouve avant Alice ou dans un mélange de ces deux situations.

Dans un deuxième temps, le calcul de la probabilité de réussite s'effectue par l'application du formalisme de la mécanique quantique. Supposons que la matrice de processus soit donnée par une combinaison de matrices de Pauli :

$$W^{A_1A_2B_1B_2} = \frac{1}{4}\left[\mathbb{1}^{A_1A_2B_1B_2} + \frac{1}{\sqrt{2}}\left(\sigma_z^{A_2}\sigma_z^{B_1} + \sigma_z^{A_1}\sigma_x^{B_1}\sigma_z^{B_2}\right)\right], \tag{5.8}$$

qui font intervenir les systèmes quantiques A_1, A_2, B_1 et B_2 à deux niveaux. Oreshkov *et al.* considèrent, pour Alice et Bob respectivement, des opérateurs locaux particuliers :

$$\begin{aligned} \xi^{A_1A_2}(x, a, b') &= \frac{1}{2}\left[\mathbb{1} + (-1)^x\sigma_z\right]^{A_1} \otimes \left[\mathbb{1} + (-1)^a\sigma_z\right]^{A_2}, \\ \eta^{B_1B_2}(y, b, b') &= b' \cdot \eta_1^{B_1B_2}(y, b, b') + (b' \oplus 1) \cdot \eta_2^{B_1B_2}(y, b, b'), \end{aligned} \tag{5.9}$$

où

$$\eta_1^{B_1B_2}(y, b, b') = \frac{1}{2}\left[\mathbb{1} + (-1)^y\sigma_z\right]^{B_1} \otimes \mathbb{1}^{B_2}$$

et

$$\eta_2^{B_1B_2}(y, b, b') = \frac{1}{2}\left[\mathbb{1}^{B_1B_2} + (-1)^b\sigma_x^{B_1}\sigma_z^{B_2}\right].$$

On peut calculer la probabilité de réussite associée à (5.8) et (5.9) [169, 36]. Elle dépasse la borne classique de l'inégalité causale (5.7) :

$$P_{success} = \frac{2 + \sqrt{2}}{4} > \frac{3}{4}. \tag{5.10}$$

Cette violation de la limite classique marque l'impossibilité d'inscrire les événements dans un ordre causal global. Il s'agit donc d'un processus causalement inséparable, qui met en superposition les situations « Alice est avant Bob » et « Bob est avant Alice ». Il est impossible de l'écrire sous forme d'un mélange classique :

$$W \neq \lambda W^{A \not\preceq B} + (1 - \lambda)W^{B \not\preceq A}, \tag{5.11}$$

le coefficient de mélange étant $0 \leq \lambda \leq 1$; $W^{A \not\preceq B}$, un process au cours duquel Alice ne peut pas envoyer un signal à Bob et $W^{B \not\preceq A}$, un process au cours duquel Bob ne peut envoyer aucun signal à Alice.

Ainsi, selon la mécanique quantique, il existe des processus inséparables, qui violent la conception classique de causalité. Le jeu permet de saisir le degré de cette violation par une mesure simple et pédagogique.

Cependant, si les entrées, a et b, et les sorties, x et y, ne sont accessibles que localement, le rôle du bit de contrôle b' reste à préciser. Est-ce une donnée globale ou locale ?

Il s'avère que b' jouit d'un statut particulier dans le jeu causal, car il ne peut être réduit à l'information d'un seul joueur. En effet, si b' était seulement accessible à Bob et si, pour chaque valeur de ce contrôle, la distribution de probabilité conjointe pouvait être réalisée à l'aide d'instruments quantiques, locaux et fixes, alors il ne serait pas possible dans un tel cadre de violer l'inégalité causale. En conséquence, b' est un paramètre véritablement non-local, partagé par tous les observateurs. Cela rappelle la difficulté de l'ordre PTM du superobservateur, identifiée par Hardy.

Le cadre ludique facilite l'étude des liens entre la violation de l'inégalité causale et divers principes qui pourraient, au sein d'une reconstruction partielle, en porter la responsabilité. Avec le doctorant Issam Ibnouhsein, nous avons montré qu'il était possible de dériver la limite quantique de l'inégalité causale à partir d'une contrainte sur l'information mutuelle d'Alice et de Bob [132]. Nous avons reformulé le jeu causal comme un code d'accès aléatoire (*random access code*) et défini une classe de jeux, tels que les processus causalement séparables, qui respectent une certaine inégalité théorético-informationnelle. Ensuite, en relâchant les contraintes de signalement dans l'ensemble des corrélations, nous avons reconstruit la borne quantique et le concept d'ordre causal indéfini d'Oreshkov *et al.*

Après avoir introduit une mesure entropique alternative de dépendance mutuelle, nous avons établi un lien entre la borne quantique et l'efficacité de la communication, par analogie avec les méthodes de dérivation de la borne de Tsirelson pour l'intrication bipartite. Mais, si cette reconstruction s'appuie sur les propriétés de l'information mutuelle, la nôtre peut s'en passer. En conséquence, dans le cadre du jeu causal (mais pas pour l'intrication), l'information mutuelle ne mesure pas de façon optimale la dépendance entre parties. Cette étude a donné l'impulsion à l'utilisation des mesures entropiques dans l'analyse de la causalité quantique [160].

5.5.3 Un modèle physique sans la notion de système

Le travail sur les ordres causaux fournit un exemple de modèle dans lequel il n'est pas nécessaire, et parfois même impossible, d'interpréter le cours des événements comme s'il agissait de quelques systèmes physiques entrant dans les laboratoires et en ressortant après la mesure. Du point de vue des approches indépendantes du dispositif, une partie est d'abord définie, non par un point dans l'espace ni par son algèbre d'observables associée, mais par deux suites de symboles : une entrée et une sortie. Celle d'entrée est libre, ce qui signifie qu'au sein du modèle décrivant cette partie, elle est aléatoire. Alice et Bob ne peuvent ni prédire ni calculer leurs entrées.

Dans le travail sur les ordres causaux, une partie est décrite par un instrument quantique. En toute rigueur, cela n'est pas une nécessité : cette description émerge sous certaines conditions seulement. Ainsi, dire qu'un système entre dans le laboratoire et qu'il en ressort n'est pas un prérequis à la construction d'un modèle physique.

Cependant, les physiciens dérogent souvent à cette règle en formulant le modèle dans les termes de systèmes physiques, par exemple "a party receives [a system] from the environment, and a physical system is returned to the environment" [22]. Certains représentent même ces deux descriptions sur un seul dessin, dans lequel les

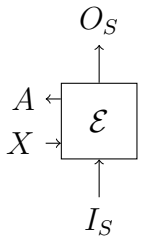

Figure 5.2 – A party is fully defined by the input variable X and the output variable A linked by the map $\mathcal{E}$. The theory does not require any mention of a physical system S that is first received from the environment and then returned to it. Adopted with modifications from [22].

entrées et les sorties co-existent avec une transformation d'un système (Figure 5.2). Ce dédoublement des présupposés fondamentaux provoque trois conséquences, toutes troublantes. Elles indiquent, à notre avis, que la notion de système ne devrait pas faire partie des ingrédients initiaux du modèle des ordres causaux indéfinis.

Le premier problème vient de la possibilité de reformuler la condition qu'un système entre dans le laboratoire, puis qu'il en ressort, comme l'exigence que chaque partie n'interagisse qu'une seule fois avec un « environnement » ou un « milieu physique » :

> [Alice and Bob] both open their lab, let some physical system in, interact with it and send a physical system out, only once during each run of the experiment. [36]

L'environnement dont il s'agit est décrit dans sa totalité par la matrice de processus. Il « se trouve » donc hors espace-temps et représente un milieu étrange, atemporel et holiste. Il « fournit » des systèmes aux laboratoires locaux. L'étrangeté de cet « environnement » crée donc un souci qu'il convient d'élaborer.

Conformément à l'ancien mode explicatif (section 5.4.2), un système est constitué à travers sa séparation de l'environnement. Or, aucune séparation n'est possible d'un « environnement » atemporel. S'il était possible de diviser les degrés de liberté entre ceux qui sont pertinents au système en question et ceux qui ne le sont pas, les premiers auraient dû rester pertinents durant toute l'expérience, jusqu'à ce que le système quitte le laboratoire. L'idée du maintien ($\sigma\upsilon\nu\acute{\epsilon}\chi o\nu\tau o\varsigma$) d'une identification est fondamentale : elle remonte au concept métaphysique de « mettre ensemble » ($\sigma\upsilon\nu\acute{\alpha}\gamma o\nu\tau o\varsigma$) les degrés de liberté. Selon Numénius (4b), elle est la clef de toute cohésion ($\sigma\upsilon\gamma\kappa\rho\alpha\tau o\hat{\upsilon}\nu\tau o\varsigma$) d'un concept séparé des autres. L'idée d'un maintien dépend donc fondamentalement d'une temporalité. Si le système disparaît, ou s'il est absorbé à l'intérieur du laboratoire, il cesse de se maintenir et, par conséquent, cesse d'exister.

Le maintien des degrés de liberté constitutifs du système ne peut s'effectuer que dans le temps du laboratoire. Toutefois, la flèche de cette temporalité locale, qui va de l'entrée à la sortie, ne s'étend pas à « l'environnement ». Défini par la matrice de processus, ce dernier n'est pas dans le temps ; d'où l'impossibilité pour un tel milieu

de « maintenir » quelque chose. Cela explique le caractère purement métaphorique de l'emploi du terme « environnement » dans le modèle des ordres causaux indéfinis. Il en va de même pour les énoncés qui contiennent les mots « entrer » ou « sortir », en parlant des « systèmes ».

La deuxième difficulté est liées à l'apparition des boucles causales. Il s'avère que le modèle autorise, sous certaines conditions, les « systèmes » à entrer deux fois le même laboratoire. Or, en pratique cela n'arrive jamais. L'interprétation que l'on donne au modèle en termes de systèmes paraît donc hâtive.

Le troisième problème est davantage troublant. Comme tous les modèles indépendants de dispositif, celui des ordres causaux indéfini est une « boîte » qui relie les entrées et les sorties. Les processus à l'intérieur de la boîte peuvent avoir diverses natures et origines. Cela inclut, dans le cas de violation de l'inégalité causale, les processus causalement superposés à deux parties, mais aussi les processus classiques multipartites [23].

Comme dans le cas de l'intrication quantique, les corrélations causales forment un polytope [178, 179], inclus dans un ensemble plus large de corrélations acausales. Il s'avère qu'avec trois parties, mais pas avec deux, il est possible d'atteindre des points en dehors du polytope classique, donc acausaux, en utilisant uniquement la théorie classique des probabilités [20]. Ce résultat mathématique est étonnant et son interprétation n'est pas consensuelle [116, 19, 24]. Selon la nôtre, il signifie que, même dans ce cadre classique, il existe des situations que l'on ne peut pas interpréter en termes de systèmes qui entrent dans le laboratoire.

Cette nouvelle difficulté, comme les deux précédentes, plaide en faveur de l'idée que, dans le modèle des ordres causaux indéfinis, l'emploi du terme « système » ne peut être que métaphorique. Un autre argument vient également en appui de cette conclusion. Si un système « traverse » un laboratoire, de l'entrée jusqu'à la sortie, il doit être possible de dire son histoire en établissant une liste d'événements engendrés par ce système au cours de sa traversée. Mais, cela n'est pas toujours possible.

Le contre-exemple le plus connu provient du « paradoxe du menteur quantique » *(quantum liar paradox)* [84] ; plus généralement, tout argument « paradoxal » lié à la postsélection produit le même effet [5]. Ce sujet n'est pas au centre de notre étude ; cependant, il illustre bien qu'une mesure future peut jouer un rôle dans la détermination de l'histoire passée d'un système. Dire qu'un système « est préparé au début de l'expérience » est donc susceptible de causer quelques difficultés théoriques et philosophiques.

Trois interprétations équivalentes sur le plan formel se dégagent, en mettant en exergue des aspects physiques et conceptuels différents :

a) Pour souligner l'aspect opérationnel, on choisit comme concept fondamental un *run* de l'expérience, décrit par une entrée et une sortie.

b) Pour mettre en valeur l'aspect pragmatique de construction d'une expérience, on définit les parties ou les laboratoires locaux en précisant les entrées et les sorties, non les arrangements spatio-temporels des sources ou des instruments de mesure.

c) Pour se rapprocher de l'heuristique du physicien, il est dit qu'un système entre dans le laboratoire et qu'il en ressort à la fin de l'expérience.

Ces trois interprétations semblent complémentaires, mais elles ne n'ont pas le même degré de nécessité. Le point a permet de passer rapidement de l'expérience au formalisme mathématique. Il contient tout le nécessaire pour le calcul. Le point b peut également jouir d'une importance théorique, par exemple pour établir un rapport avec la théorie de la relativité ou la gravité quantique [125]. Un système, mentionné dans c, ne trouve pas de place naturelle dans cette interprétation ; il n'est qu'un artefact supplémentaire. Or, il est tout de même souvent vrai que a et b soient cohérents avec c. Dans ce cas, c fournit une heuristique utile pour comprendre la situation physique. Mais, cette cohérence peut aussi être brisée. On ne peut alors qu'abandonner c, en laissant la théorie physique, construite selon a et b, sans la notion de système.

Chapitre 6

Perspectives

À la fin du XX^{e} siècle, le physicien Asher Peres critiquait vigoureusement l'interprétation relationnelle de la mécanique quantique, proposée par Carlo Rovelli [191]. Peres n'était pas d'accord avec l'idée d'universalité de l'observateur quantique. Selon cette conception, tout système peut être observateur pourvu qu'il existe une corrélation entre ses degrés de liberté et ceux d'un autre système, qui est observé. Peres objecta : "The two electrons in the ground state of the helium atom are correlated, but no one in his right mind would say that each electron 'measures' its partner" [174]. Cette controverse continue encore aujourd'hui. Tout système physique peut-il compter comme un observateur ? Sinon, quelles contraintes, et à quel niveau, faudrait-il imposer ? Notre programme de recherche apporte une réponse à ces questions.

L'interrogation sur le concept d'observateur, résumée *supra* (section 5.1), devient riche et complexe déjà pendant les vingt premières années de l'existence de la mécanique quantique. Une position moderne, qui se dégage vers la fin de cette période, emploie des outils mathématiques allant au-delà du formalisme standard des espaces de Hilbert (section 5.2). Aujourd'hui, le traitement opérationnel de la théorie quantique suit les lignes d'une approche indépendante du dispositif (section 5.4). Ainsi, les éléments fondamentaux de notre approche sont des suites de symboles appartenant à des alphabets finis. Ils correspondent aux entrées et sorties d'une « boîte », dont la complexité est bornée : cela est un ingrédient crucial de la définition d'observateur.

Nous consacrons ce bref exposé des perspectives à l'élaboration des conséquences — théoriques, expérimentales et philosophiques — de l'application au problème d'observateur d'une approche indépendante du dispositif. Le formalisme de la complexité de Kolmogorov, déjà introduit plus haut (section 5.4.4), permet de formuler quelques premières conjectures (section 6.1.1). Mais, il est aussi possible de traiter les suites de symboles par d'autres méthodes mathématiques avancées.

Si on les traite comme des éléments d'un espace topologique discret, alors l'étude du groupe fondamental d'un tel espace fournit des pistes à l'exploration de la contextualité quantique (section 6.1.2).

Si on les traite comme des concaténations des mots dans un code correcteur, alors

il devient possible, moyennant quelques présupposés supplémentaires, de dériver une limite quantitative sur les corrélations entre entrées et sorties (section 6.1.3). Il s'avère que cette limite est légèrement inférieure à la borne de Tsirelson et que ce décalage peut être testé expérimentalement (section 6.2).

Enfin, nous énumérons, dans la section 6.3, tous les points philosophiques saillants qui forment le socle conceptuel de notre programme de recherche. À ce stade, il ne s'agit que des conjectures : quoique appuyées par les modèles théoriques, elles nécessitent encore un important effort de réflexion et d'analyse.

6.1 Perspectives théoriques

6.1.1 Suites symboliques et complexité

Un modèle indépendant du dispositif est défini comme une « boîte » qui reçoit en entrée une suite a de caractères dans un alphabet $\mathbb{I}$ de cardinalité finie q_i, et produit en sortie une suite A de caractères dans un alphabet $\mathbb{O}$ de cardinalité finie q_o. S'il est possible (et cela ne l'est pas toujours) de diviser le cadre du modèle en M parties ou laboratoires locaux, on désigne par a_i et A_i les entrées et les sorties de chaque partie. Après $N \to \infty$ répétitions indépendantes de l'expérience (*runs* ou prises de données), les séquences a and A sont des limites de a^N et A^N, respectivement, c'est-à-dire, des concaténations calculables des données d'une seule partie dans un seul *run*, $a_i^{(j)}$ et $A_i^{(j)}$, pour toutes les valeurs $0 \leq i \leq M$ et $0 \leq j \leq N$. Par exemple, dans le cadre opérationnel des boîtes PR (voir équation 5.1), caractérisé par deux parties possédant chacune un bit d'entrée et un bit de sortie, les suites a^N and A^N sont des séquences de zéros et des uns de longueur $2N$. La définition du modèle ne contient qu'une seule distinction fondamentale, celle entre deux types des données, a et A.

Ce modèle, formulé dans le langage algébrique et symbolique, ne ressemble évidemment pas à la mécanique quantique standard. Notamment, il n'emploie pas les nombres complexes ; qui plus est, il ne contient aucun élément continu. Le recours aux alphabets finis rend calculables toutes ses expressions. L'écart par rapport au formalisme quantique est manifeste. Il paraît même infranchissable. Or, quelques analogies surgissent soudainement.

La définition (5.3) fait intervenir la complexité de Kolmogorov. Le théorème de Zvonkin-Levin [228, 196] met au jour un lien entre deux notions d'information mutuelle, celle de Shannon et celle de Kolmogorov, dans la limite du nombre infini de *runs* :

$$I_H(A : a) = \lim_{N\to\infty} \frac{I_K(A^N : a^N)}{N}. \tag{6.1}$$

L'information mutuelle de Kolmogorov est définie ici par analogie avec sa définition habituelle :

$$I_K(A^N : a^N) = K(a^N) + K(A^N) - K(A^N, a^N) + O(\log N). \tag{6.2}$$

Par conséquent :

$$I_H(A:a) = \kappa(a) + \lim_{N\to\infty} \frac{\Delta K^N}{N}, \tag{6.3}$$

où, par définition, $\Delta K^N = K(A^N) - K(A^N, a^N)$, et $\kappa(a)$ est le taux de complexité de la suite d'entrée.

Revenons à l'exemple du cadre bipartite avec un bit d'entrée et de sortie par partie. Les entrées, nous l'avons dit, sont « libres ». Mathématiquement, cela signifie qu'elles sont aléatoires et distribuées uniformément et indépendamment :

$$\kappa(a) = 2.$$

La quantité ΔK_N, quant à elle, correspond à la diminution de la complexité de la suite de sortie, A, qui est *a priori* fort complexe, lorsqu'on adjoint à cette suite celle d'entrée. En effet, s'il existe entre ces deux séquences une corrélation, elle provient d'une loi physique, bien qu'elle puisse demeurer inconnue. On peut traiter, sur ce plan, une loi comme un programme ayant une taille assez courte et, surtout, fixe, qui permet de comprimer davantage la suite composée $\overline{Aa}$. Sa complexité sera ainsi inférieure à $4N$.

Supposons, pour exercice, que la loi en question soit déterministe : $A_i^{(j)} = f(a_i^{(j)})$, avec f une fonction calculable, qui peut rester inconnue. Dans ce cas, il est possible de traiter f en tant qu'algorithme de transformation des éléments d'entrée en éléments de sortie. On obtient alors $K(A_i^{(j)}) \leq K(a_i^{(j)}) + O(1)$ [196, p. 17, théorème 3]. Après N applications de cet algorithme fixe dans N *runs*, la complexité ne croît que de façon logarithmique en N : $K(A^N) = K(A^N, a^N) + O(\log N)$. Ce résultat est aussi valide dans le cas où, à la place d'une seule fonction, on a affaire à une distribution de probabilité classique sur un ensemble de fonctions, $A_i^{(j)} = \sum_k c_k f_k(a_i^{(j)})$, avec des coefficients c_k inconnus mais fixes. En insérant ces éléments dans (6.3), on obtient la borne classique :

$$I_{H,\text{classical}}(A:a) = 2 + \lim \frac{\Delta K^N}{N} = 2.$$

Reprenons le développement du modèle formel. La définition (5.3) permet d'élaborer un point de vue « géométrique » de l'observateur, suivant les idées mathématiques de Manin et Marcolli [153, 152].

À partir de l'intervalle rarifié $(0,1)_q = [0,1]\backslash\{m/q^n | m,n \in \mathbb{Z}, q \in \mathbb{N}\}$, on construit un cube rarifié à n dimensions, dont les points sont $x = (x_1, \ldots, x_n) \in (0,1)_q^n$. On peut les identifier aux matrices de taille $(\infty \times n)$, dont la colonne numéro k est une décomposition de x_k en base q.

Par définition, un code $C \subset \mathbb{A}^n$ est un sous-ensemble des suites de caractères dans l'alphabet $\mathbb{A}$ de cardinalité q ayant la longueur $n \geq 1$. Le code permet de définir $S_C \subset (0,1)_q^n$, ensemble de toutes les matrices dont les rangées sont dans C. C'est un fractal de Sierpinski. Sa dimension de Hausdorff est égale au taux du code $R \equiv \frac{\log_q \#C}{n}$. La cloture de S_C dans le cube $[0,1]^n$ inclut des points rationnels en base q. Ce nouveaux fractal $\hat{S}_C$ est alors un espace métrique dans la topologie héritée de $[0,1]^n$.

Injectons un argument lié à la complexité. On choisit une famille de codes C_r, ayant $\#C_r = q^{k_r}$ mots de longueur n_r, dont les taux tendent vers R du bas :

$$\frac{k_r}{n_r} \nearrow R. \tag{6.4}$$

Ces codes définissent un fractal $S_R = \bigcup_r S_{C_r}$ dont la dimension de Hausdorff est $\dim_H(S_R) = R$. Cependant, pour un ensemble de suites formées des mots du code C, dont le taux est R, on peut calculer la complexité inférieure de Kolmogorov. Par définition, elle est égale à $\liminf_{N\to\infty} \frac{K(\omega)}{N}$, où la limite est prise sur toutes les concaténations finies ω, formées de mots du code, dont la longueur totale est égale à N. Cette quantité satisfait à :

$$\sup_{x\in\hat{S}_C} \kappa(x) = R. \tag{6.5}$$

Également, pour tous les mots $x \in S_R$ dans un langage formel composé des codes C_r, la complexité inférieure de Kolmogorov est bornée par $\kappa(x) \leq R$. Ainsi la cloture $\hat{S}_R$ du fractal S_R est un espace métrique, qui décrit la manipulation des mots ayant des longueurs différentes, mais une complexité de Kolmogorov limitée. C'est cet objet qui fournit une représentation géométrique de l'observateur.

Conjecture 6.1. *Les observateurs quantiques sont sélectionnés par la possibilité de description informationnelle continue dans $\hat{S}_R$. S'il est possible de donner un modèle possédant, dans le cas limite, une description continue, ce modèle est quantique ou proche du quantique.*

Enfin, il est possible de donner, toujours en termes de complexité, une définition opérationnelle des cadres « bipartite » ou « multipartite ». En mécanique quantique, lorsque Alice et Bob partagent un état intriqué, chaque partie peut choisir ses paramètres de mesure indépendamment parmi trois valeurs qui correspondent aux matrices de Pauli $\sigma_x, \sigma_y, \sigma_z$. Au total, cela produit six suites de sortie. Or, les parties qui partagent un état intriqué obtiennent des sorties corrélées. La connaissance de l'état intriqué permet de prédire les sorties de Bob à partir de celles d'Alice. Dans ce cas, les complexités des suites de sortie sont les mêmes pour toutes les parties. Ce raisonnement sert de motivation à une définition que nous présentons sous forme de conjecture, car ses conséquences restent encore à explorer.

Conjecture 6.2. *Dans les approches indépendantes du dispositif, qui ne présupposent pas de division en parties ou en laboratoires locaux, le concept quantique de système intriqué bipartite correspond aux suites de sortie ayant la même complexité de Kolmogorov.*

Notamment, s'il existe six suites de sortie avec la même complexité, on appelle ce cadre « opérationnellement bipartite ».

6.1.2 Homotopie et contextualité

L'analyse des fondements de la mécanique doit beaucoup à Ernst Specker. À la fin des années 1950, il mène une profonde réflexion, qui aura marqué la philosophie de la mécanique quantique tout comme la théorie elle-même [202]. La démonstration par Kochen et Specker, publiée au milieu des années 1960, d'un théorème qui porte aujourd'hui leur nom [138, 139], introduit une nouvelle caractéristique non-classique de la mécanique quantique : sa contextualité.

La contextualité quantique est définie comme l'impossibilité logique (et non seulement empirique) d'attribuer des valeurs à un ensemble d'observables dans l'espace de Hilbert d'une manière qui ne dépend pas de la mesure. Comme les inégalités de Bell, cette nouvelle caractéristique non-classique ne gagne que lentement ses droits de cité dans les fondements de la mécanique quantique. Des résultats remarquables, liés à la contextualité, commencent à paraître dans les années 1980, notamment grâce aux contributions de l'école d'Aharonov [5]. On peut aussi citer la preuve simplifiée de la contextualité quantique, dite « le carré magique » de Mermin [159]. Plus récemment, tout un ensemble des travaux théoriques voit le jour, qui ouvre un point de vue sur la contextualité comme une ressource non-classique, utilisable dans le calcul quantique [50, 28, 90]. Le jeune chercheur Hippolyte Dourdent a préparé, sous notre direction, un rapport sur l'état actuel des recherches en contextualité quantique [73].

Sur le plan mathématique, notamment, Abramsky *et al.* étudient la structure de la contextualité via la cohomologie de Čech [3]. D'autres arguments de nature topologique, utilisant la cohomologie des groupes, apparaissent dans les travaux du groupe de Raussendorf [167]. Dans le même esprit topologique, nous proposons une approche différente des leurs.

Supposons que les variables aléatoires, en l'occurrence les entrées dans un modèle indépendant du dispositif, soient des singletons ouverts dans une certaine topologie discrète. D'autres ouverts dans cette topologie contiennent des éléments qui ne sont pas, à eux seuls, des ouverts ; cela signifie qu'ils ne sont pas libres : ce sont des sorties. Comme l'illustre la Figure 6.1a, $\{a\}$ et $\{c\}$ sont des variables libres et des ouverts, tandis que $\{b\}$ et $\{d\}$ ne le sont pas.

Une boucle autour du diagramme de Hasse de cet espace topologique discret commence et se termine par une variable aléatoire, par exemple $\{a\}$. Sur le plan d'interprétation, cela signifie que l'observateur (plus précisément, le superobservateur) commence l'expérience par l'entrée a, y adjoint ensuite c pour obtenir un ensemble maximal de variables aléatoires ac. Puis, il opère des mesures, en commençant par acb, qui contient une sortie, b. Enfin, il atteint l'état d'information maximale, $acbd$. Ayant terminé la collecte d'information, l'observateur efface sa mémoire, afin de lancer un nouveau *run* de l'expérience. Cela, sur le diagramme, correspond au retour par la boucle à la case départ, a. Physiquement, donc, une boucle décrit un *run* complet, y compris la mise à zéro de toutes les données. N tours d'une boucle correspondent à N *runs* de l'expérience.

Les boucles sont classifiées par la groupe fondamental π_1 de l'espace topologique discret. Dans la Figure 6.1a, $\pi_1 = \mathbb{Z}$. Cependant, en toute généralité, ce groupe fonda-

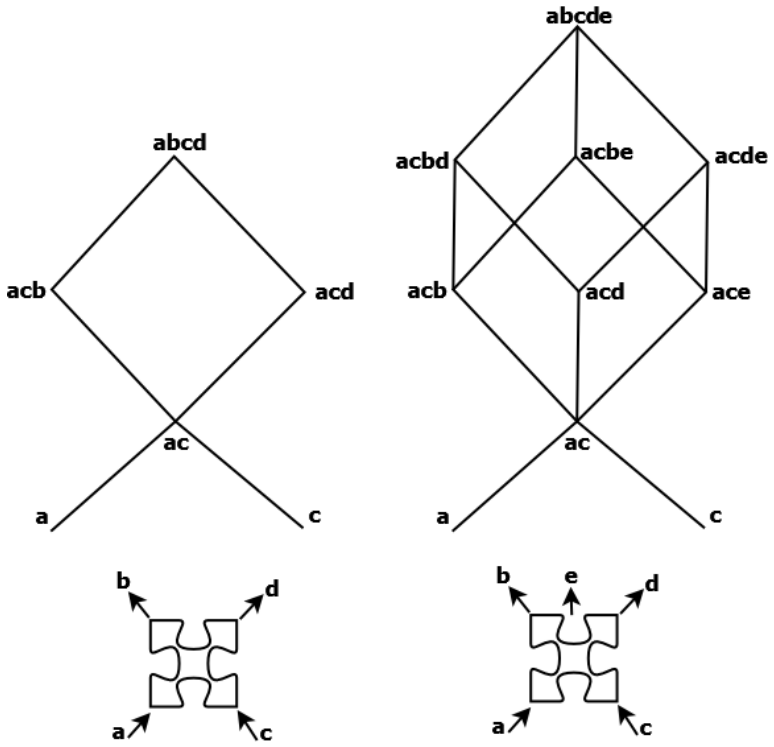

FIGURE 6.1 – a) Hasse diagram of a box with two inputs and two outputs. Topological space is determined by the base $\{a\}, \{c\}, \{acb\}, \{acd\}$ and has a fundamental group $\mathbb{Z}$. There exists a unique loop. b) Hasse diagram of a generalized box with two inputs and three outputs. Topological space given by the base $\{a\}, \{c\}, \{acb\}, \{acd\}, \{ace\}$ has a fundamental group $\mathbb{F}_2$, the free group with two generators. Loops of the first kind extend from the two-element set $\{ac\}$ to a four-element set, e.g., $\{abcd\}$, spanning three levels of the diagram. Loops of the second kind reach the five-element set $\{abcde\}$ and span four levels.

mental est un groupe libre non-Abélien avec n générateurs (le cas $n = 2$ est illustré à la Figure 6.1b). Chaque générateur décrit un type particulier de boucle, non équivalent aux autres. En termes d'interprétation, il s'agit d'un choix d'observables à mesurer. Celles qui sont incompatibles sont mesurées en suivant des boucles différentes, non-équivalentes. La répétition des *runs* d'une expérience s'écrit ainsi comme g^n, avec $n \in \mathbb{Z}$ le nombre de runs et le générateur g correspondant au choix d'une boucle dans le diagramme de Hasse. Cette non-équivalence des boucles introduit une contextualité : l'information de l'observateur n'est pas seulement captée par un ouvert, qui correspond à l'état de sa mémoire, mais elle dépend de la boucle à laquelle cet ouvert appartient.

Conjecture 6.3. *La contextualité des modèles indépendants du dispositif peut être décrite par le groupe fondamental d'un espace topologique discret.*

Si π_1 est non-Abélien (Figure 6.1b), les boucles peuvent être plus « grandes » ou plus « petites », en fonction du nombre de « niveaux » qu'elles traversent. En effet, il est possible, dans l'ordre partiel des niveaux du diagramme, d'associer à chaque boucle un contenu informationnel, par le comptage des différences de niveaux entre le plus élevé et le plus bas. On attribue ainsi un poids à chaque générateur du groupe fondamental. Si la taille de la mémoire de l'observateur ne varie pas en fonction de son contenu, il est possible d'écrire une fonction de partition d'un système statistique comprenant tous les générateurs avec leurs poids respectifs :

$$Z = \sum \lambda_{i_1 \dots i_n} g_1^{i_1} \dots g_n^{i_n}. \tag{6.6}$$

Conjecture 6.4. *Si la mécanique statistique des boucles possède un régime critique sur $\hat{S}_R$, il donne naissance à un comportement émergeant, qui ne dépend pas des particularités de chaque observateur. Ce comportement peut alors être interprété comme un régime classique, non-contextuel, des systèmes idéalisés, préparés et mesurés un nombre illimité de fois à coût informationnel zéro.*

6.1.3 Codes et corrélations

Imposer au modèle indépendant de dispositif seulement une contrainte, qui stipule l'existence d'une borne supérieure de la complexité des suites de symboles, n'est pas encore suffisant pour obtenir un modèle proche de la mécanique quantique. Pour estimer ΔK^N dans l'équation (6.3), il faut poser des principes contraignant davantage les entrées et les sorties.

Nous l'avons déjà dit, la présence d'une loi physique se traduit par la diminution de la complexité algorithmique lorsqu'on adjoint à la séquence de sortie celle d'entrée. Supposons que la dépendance entre entrées et sorties soit linéaire, même si ses coefficients peuvent varier selon les *runs*. Cela peut être représenté mathématiquement avec les outils de la théorie algébrique des codes correcteurs.

Choisissons une matrice G de taille finie, potentiellement grande, telle que ses éléments appartiennent à un corps fini $\mathbb{F}_q$, avec $q = \max\{q_i, q_o\}$. Elle génère un

code linéaire, C_A, défini en traitant ses rangées comme les mots du code. Rappelons que A désigne les sorties : la notation C_A prend son sens en vertu d'un présupposé supplémentaire, à savoir que la suite de sortie A_N après N *runs* peut être représentée comme une concaténation calculable, manifestement longue, des mots de C_A. Des mots individuels peuvent même atteindre la longueur de A_N sans que cela pose un problème à cette construction théorique ; simplement, dans ce cas, la suite ne sera pas compressible.

En adjoignant à la suite de sortie celle d'entrée, a_N à A_N, nous obtenons une concaténation calculable des mots d'un autre code linéaire, $C_{\overline{Aa}}$. Sa complexité est inférieure à celle de C_A si et seulement si une loi, que nous supposons linéaire pour tout N, relie A et a.

On trouve une certaine analogie quantique à ce principe. En effet, la règle de Born assure le caractère aléatoire, et donc incompressible, des sorties, ce qui contribue à rendre G très large. Mais, pour le peu d'information qui reste à comprimer, la linéarité de l'espace de Hilbert, responsable de la dépendance entre les paramètres et résultats de mesure, ressemble au postulat de linéarité du code.

Les codes C_A et $C_{\overline{Aa}}$ sont inconnus et probablement très complexes. Toutefois, il existe une relation entre leurs paramètres : le taux R et la distance relative δ, définie comme la distance de Hamming minimale entre les mots du code, rapportée à leur longueur. La distance de Hamming entre deux séquences est égale, par définition, au nombre de positions qui contiennent des symboles différents.

La borne de Gilbert-Varshamov (GV) établit une relation, sous forme d'inégalité, vérifiée par les « bons » codes, au sens de leur capacité à corriger les erreurs en maximisant R et δ :

$$R \gtrsim 1 - h_q(\delta_A), \tag{6.7}$$

où la fonction h est l'entropie binaire à base q :

$$h_q(x) = x \log_q(q-1) - x \log_q(x) - (1-x) \log_q(1-x). \tag{6.8}$$

Parmi les codes linéaires qui s'approchent asymptotiquement, lorsque leur taille tend vers infini, de la borne GV, on trouve ceux générés par des matrices G remplies uniformément au hasard. En revanche, les codes « constructifs », qui possèdent un algorithme de fabrication simple, se trouvent loin de cette borne.

Dans la même limite asymptotique $N \to \infty$, la borne d'Elias-Bassalygo met une limite supérieure sur le taux des codes linéaires :

$$R \leq 1 - h_q \left[\left(1 - \frac{1}{q}\right) \left(1 - \sqrt{1 - \frac{q\delta}{q-1}}\right) \right] + o(1). \tag{6.9}$$

Pour un alphabet binaire, ces deux bornes s'écrivent de manière simplifiée. La borne de Gilbert-Varshamov :

$$1 - h(\delta) \leq R \tag{6.10}$$

et la borne d'Elias-Bassalygo :

$$R \leq 1 - h\left(\frac{1 - \sqrt{1-2\delta}}{2}\right). \tag{6.11}$$

En utilisant (6.5), nous obtenons :

$$\lim \frac{\Delta K^N}{N} = R_A - R_{\overline{Aa}}. \tag{6.12}$$

Pour les codes binaires,

$$\lim \frac{\Delta K^N}{N} \leq h(\delta_{\overline{Aa}}) - h\left(\frac{1-\sqrt{1-2\delta_A}}{2}\right), \tag{6.13}$$

avec l'entropie binaire $h(x) = -x\log_2(x) - (1-x)\log_2(1-x)$.

Revenons à nouveau à la situation de deux parties possédant chacune une entrée et une sortie. Lorsque les suites d'entrée et de sortie sont fusionnées, la longueur totale de la séquence passe de $2N$ à $4N$. Dans le même temps, la distance de Hamming augmente de N, car les entrées sont distribuées uniformément au hasard à travers N *runs*, au taux de deux entrées par *run* : 00, 01, 10 ou 11. Il en suit que :

$$\delta_{\overline{Aa}} = \frac{2N\delta_A + N}{4N} = \frac{\delta_A}{2} + \frac{1}{4}. \tag{6.14}$$

Cette équation nous permet d'estimer la borne (6.13). Le calcul montre qu'elle atteint sa valeur maximale, égale à $h(\frac{1}{4}) = 2 - \frac{3log_2(3)}{4} \simeq 0.8113$, à $\delta_A = 0$. En l'insérant dans l'équation (6.3), nous obtenons :

$$I_H(A:a) \leq 2.8113. \tag{6.15}$$

Cela signifie que, dans notre modèle, l'information mutuelle entre entrées et sorties est maximale dans le cas où, à elle seule, la suite de sortie est totalement aléatoire et incompressible. Ce n'est que quand on y adjoint les entrées qu'apparaît un sens, dû à une dépendance linéaire. Il est également vrai que toute la corrélation entre les entrées et les sorties se limite à un choix aléatoire avec la probabilité $\frac{1}{4}$. La quantité d'information dans ce choix est égale à l'entropie binaire $h(\frac{1}{4})$.

Quelle est la source du décalage, petit mais réel, entre la valeur obtenue dans (6.15) et la borne de Tsirelson $2\sqrt{2} \simeq 2.8284$? Plusieurs présupposés y concourent :

1. Le rapport entre l'information mutuelle et la quantité CHSH (5.2) reste à établir. Les deux quantités mesurent la force des corrélations entre les entrées et les sorties ; elles devraient être liées. Une piste intéressante à suivre serait celle d'une application de l'inégalité de Fano, déjà utilisée dans le travail sur la causalité informationnelle [170].
2. Le choix d'un corps fini est important. À sa place, la mécanique quantique emploie le corps continu des nombres complexes.
3. Le caractère statistique de notre argument ne doit pas être négligé. La majorité des codes sont proches de la borne GV, or il en existe d'autres qui la dépassent, notamment pour $q \geq 49$.
4. Le présupposé d'une dépendance linéaire entre les entrées et les sorties est crucial.

La borne de ce modèle discret n'est valide que pour des codes linéaires longs, qui demeurent aléatoires dans la limite asymptotique. *A priori*, rien ne laisse penser qu'elle devrait coïncider avec la borne de Tsirelson. Cependant, elle en sort étonnamment proche.

Il est instructif de calculer, pour exercice, la valeur de (6.12) dans le cas des boîtes PR (5.1). Le produit des entrées, xy, divise les quatre possibilités, $x, y \in \{0, 1\}$, en deux groupes : $xy = 0$ si $x =$ ou $y = 0$ (trois cas sur quatre), et $xy = 1$ si et seulement si $x = y = 1$. Rappelons que, par la défitinion du modèle,

$$a + b = xy.$$

Dans les trois premiers cas, $a = b$. Cela signifie que les sorties prennent les valeurs $a = b = 0$ ou $a = b = 1$ avec la probabilité $\frac{1}{2}$. Dans le dernier cas, $a \neq b$ et les sorties sont $\{a = 0, b = 1\}$ ou $\{a = 1, b = 0\}$ avec la probabilité $\frac{1}{2}$. Ainsi, pour restituer l'information mutuelle, il est suffisant, d'abord, de tirer au sort avec la probabilité de succès $\frac{1}{4}$, puis de déterminer la valeur d'une seule sortie, car elle suffit à connaître la seconde. La quantité d'information nécessaire est alors :

$$I_{PR} = h(\frac{1}{4}) + 1 \simeq 1.8113. \tag{6.16}$$

Or, cette quantité d'information, déjà assez faible, s'avère tout de même excessive si on ne souhaite calculer que l'expression CHSH (5.2). En effet, CHSH ne dépend pas de la connaissance précise des sorties, mais seulement de leur corrélation. Pour les boîtes PR, il suffit pour la calculer de connaître $h(\frac{1}{4}) \simeq 0.8113$ bits d'information.

La bizarrerie de cette valeur réside dans son infériorité par rapport à celle, proche de la limite quantique, que nous avons calculée *supra* dans le cas des modèles indépendants de dispositif et linéaires. Qui plus est, elle est inférieure même à celle qu'on exige classiquement !

Bizarrerie, certes, mais cela n'est pas une nouveauté : Fritz et Chaves ont déjà établi que le modèle PR, qui dépasse normalement la borne de Tsirelson, redescend en-dessous de la limite quantique lorsqu'on considère, à la place de l'inégalité CHSH habituelle, des inégalités entropiques [54]. Pour cela, ils ont utilisé l'entropie de Shannon, invariante par rapport aux permutations des valeurs des sorties qui ne changent pas les probabilités. Ainsi, toute boîte isotropique [156] $P^{\text{iso}}(a, b|x, y) = \frac{1}{4}\left(1 + C(-1)^{a\oplus b\oplus xy}\right)$, avec $C \in [0, 1]$, est susceptible de respecter l'inégalité de Bell entropique car, par simple permutation des sorties, elle peut être rendue équivalente, au sens seulement de l'entropie, au modèle classique $P^{\text{c}}(a, b|x, y) = \frac{1}{4}\left(1 + (-1)^{a\oplus b}\right)$. Ce constat peu satisfaisant advient à cause de la forte non-linéarité des inégalités entropiques, qui permettent, en particulier, à ce que l'inégalité soit violée par un mélange de deux boîtes respectant chacune sa borne classique.

Lorsque nous remplaçons, dans l'utilisation des inégalités entropiques, l'entropie de Shannon par la complexité de Kolmogorov, le caractère statistique des arguments s'efface. Mais, l'analyse des boîtes isotropiques est toujours possible ; qui plus est, dans le langage de la complexité algorithmique, à peu près comme dans celui de Shannon,

le modèle de Popescu-Rohrlich, très fortement corrélé, est classé plus « bas » que la physique classique. Cela est dû au concours de trois facteurs :

— le rôle réduit des entrées, qui déterminent seulement le signe du corrélateur, distribué $\frac{3}{4}$ vs. $\frac{1}{4}$;
— l'absence de toute autre information mutuelle entre les entrées et les sorties ;
— la caractère isotropique du modèle, qui rend inutile la préservation de l'information sur chacune des deux sorties individuelles.

6.2 Perspectives expérimentales

Notre modèle discret fournit une estimation de la force des corrélations, fondée sur un ensemble de présupposés qui l'éloigne du cas quantique. Nous l'avons dit, le manque de continuité dans le modèle y apporte aussi une contribution.

En 2015, partant des résultats de Manin et Marcolli cités *supra*, nous avons suggéré que les codes possèdent un régime critique, décrit par un modèle conforme et continu [109]. Sans reprendre ici l'élaboration de ce modèle, ne citons que sa principale conclusion : une estimation numérique de la force des corrélations, supérieure à celle du modèle discret, mais tout de même inférieure à la borne de Tsirelson.

À ce jour, le meilleur résultat expérimental [199], obtenu à l'aide des sources optiques très fiables, est compatible avec nos deux modèles. D'ailleurs, après la première publication en 2015 par un groupe au Singapour [182], ce résultat a été révisé à la baisse, début 2018, suite à la reprise de l'expérience en Suède, dans laquelle une veille plus stricte était assurée du respect de la contrainte de non-signalement. Nous avons contribué à convaincre les auteurs de la nécessité d'une telle révision.

Mécanique quantique	$2\sqrt{2} \simeq 2.8284$
Modèle continu [109]	2.8254
Valeur observée [199]	2.8117 ± 0.0032
Modèle discret	$2 + h(\frac{1}{4}) \simeq 2.8113$

Les expériences à venir montreront si la borne de Tsirelson, limite quantique théorique, peut être atteinte avec une meilleure précision par les systèmes empiriques. En toute rigueur, ces mesures doivent être effectuées selon une approche indépendante du dispositif, qui ne présuppose pas la validité de la mécanique quantique, car c'est elle qui est testée. Or, nous avons déjà insisté (section 5.4.4) sur la difficulté de réaliser en pratique une expérience indépendante du dispositif.

Ainsi, notre estimation de la force des corrélations, calculée en totale indépendance du dispositif, a peut-être des chances de résister aux observations futures. Conscients du caractère fort spéculatif de cette éventualité, nous ne pouvons tout de même pas ne pas l'envisager. Pour conclure, nous en tirons quelques enseignements philosophiques et conceptuels.

Conjecture 6.5. *La valeur empirique de la force des corrélations bipartites est légèrement inférieure à la borne de Tsirelson. La mécanique quantique ne la prédit qu'approximativement, mais pas tout à fait précisément.*

6.3 Perspectives philosophiques

Nous avons déjà proposé une analyse conceptuelle et philosophique générale des approches indépendantes du dispositif (section 5.4). Cette section ne fait que la compléter en y ajoutant quelques enseignements supplémentaires, tirés du modèle exposé *supra*.

6.3.1 Hasard et liberté

Le libre choix des paramètres de mesure, dogme standard de la mécanique quantique, reçoit un sens précis dans une approche indépendante du dispositif. Parmi les physiciens et les philosophes de la physique, la référence au libre arbitre a toujours été une source de suspicion philosophique. Si l'observateur est un être conscient, sa liberté peut encore se justifier ; mais s'il s'agit d'une machine ou d'un système abstrait, le sens du libre arbitre, appliqué à un tel observateur, devient davantage obscur. Ainsi, par exemple, les partisans du superdéterminisme en physique quantique nient fermement la réalité du libre choix des paramètre de mesure [207].

Dans les quinze dernières années, cette thématique a ressurgi sous un angle nouveau, suite à la publication du « théorème du libre arbitre » (*free will theorem*) par Conway et Kochen [64, 137]. Sa signification a été beaucoup débattue [53, 99], notamment en lien avec le problème du caractère ultime, ou non, de l'indéterminisme [224]. Dans un modèle indépendant du dispositif, la conception de l'observateur est formelle : c'est un ensemble de suites symboliques de complexité limitée. La question de liberté ne peut pas y être posée directement, faute de présence de termes théoriques capables de l'exprimer. Or, un équivalent formel existe bel et bien pour la vague idée de « free will » : dire que la suite d'entrée est libre n'est rien d'autre que de postuler son incompressibilité, c'est-à-dire postuler que sa complexité de Kolmogorov est de l'ordre de sa longueur. Cette définition contourne l'emploi des termes obscurs du langage commun et donne au problème du libre arbitre une formulation purement scientifique.

6.3.2 Loi physique et complexité algorithmique

Dans un modèle indépendant du dispositif, la signification opérationnelle des entrées et des sorties n'est pas perdue, même si la définition de la liberté ne fait intervenir aucun geste de sélection consciente. Nous l'avons déjà dit, le seul principe fondamental, nécessaire pour développer un tel modèle, exige qu'une distinction soit faite entre ces deux types de données. Or, sur le plan de la complexité, les entrées sont

aléatoires, mais les sorties peuvent aussi former une suite incompressible. Ce n'est que dans les corrélations entre elles que l'on aperçoit l'action d'une loi physique.

La présence d'une loi se traduit par la diminution de la complexité algorithmique de la séquence de symboles, composée par concaténation calculable des suites d'entrée et de sortie. Le respect de ce critère de baisse de la complexité au sein d'un modèle permet d'entamer l'élaboration d'une théorie, fondée sur une loi. Dans le cas contraire, aucune écriture compacte des données ne sera possible et tout espoir de construire une théorie prédictive se révélera vain.

Ce schéma n'est pas nouveau : déjà en 1980, Peres envisageait que la loi physique pourrait résulter d'une *convergence* des lois « intéressantes », potentiellement très nombreuses [173]. Il appartenait, selon Peres, « au physicien » de poser une limite sur le nombre de lois. Nous avons traduit cette intuition dans le langage de la complexité. À la place du « physicien », un observateur est défini formellement, caractérisé par l'existence d'une limite supérieure de la complexité algorithmique des suites symboliques. Ces suites sont formées par la concaténation des mots dans les codes, dont il existe potentiellement un grand nombre. Les paramètres de ces codes, dans la limite du nombre infini de *runs*, atteignent asymptotiquement la borne de Gilbert-Varshamov, mais leur contenu n'est pas totalement aléatoire. C'est cette convergence, analogue à celle imaginée par Peres, qui ouvre la possibilité de parler d'un sens des corrélations, capté au sein d'un modèle par une loi physique.

6.3.3 Systèmes composés

La règle de composition des sous-systèmes d'un système composé, nous l'avons dit, est le point focal des reconstructions partielles de la mécanique quantique. La borne quantique de l'intrication dans les inégalités de Bell vient en conséquence de l'un des principes que l'on pose pour en dériver une limite sur les corrélations bipartites ou multipartites. Or, en tant que sujet de préoccupation physique et philosophique, ce caractère central de la règle de composition est antérieur à la mécanique quantique. Nous avons exposé ses racines historiques dans un ouvrage accessible [107], en commençant par la physique des stoïciens et par les débats théologiques au premier millénaire.

Les approches indépendantes du dispositif permettent de tirer un autre enseignement au sujet de la composition des systèmes : il est possible de la définir de façon opérationnelle en utilisant un argument lié à la complexité.

En triant les suites de symboles dans l'ordre croissant de leur complexité de Kolmogorov, on crée un ordre de Kolmogorov. Pour les codes, cet ordre fournit un arrangement des mots d'une famille de codes, $a_i \in \bigcup_r C_r$, dans l'ordre croissant de leur complexité [151]. L'ordre de Kolmogorov n'est pas calculable ; il est aussi fort différent de tout schéma de numérotation des a_i fondé sur la distance de Hamming entre les codes C_r.

Les mots adjacents dans l'ordre de Kolmogorov possèdent la même complexité : cette propriété permet de les associer, par analogie, aux séquences des sorties qu'obtiennent les observateurs partageant un état quantique intriqué. Nous avons proposé,

dans un modèle indépendant de dispositif, de prendre cette propriété comme une définition. Ainsi, les ensembles des suites symboliques possédant la même complexité de Kolmogorov sont à même de remplacer le langage imprécis de « laboratoires locaux » ou de plusieurs « parties », tous les deux encore aujourd'hui des dogmes standards de la mécanique quantique.

6.3.4 Théories effectives et mathématiques du continu

En 2008, dans un rapport sur les enjeux conceptuels et philosophiques de la théorie quantique des champs au moment du lancement du LHC, nous avons discuté de questions liées aux théories effectives [104]. Quelques années plus tard, nous avons repris cette thématique dans le cadre d'une interrogation générale sur la place des mathématiques en physique [110]. Ici, un autre aspect d'effectivité nous préoccupe, à l'instar de la reformulation du statut de la mécanique quantique dans une approche indépendante du dispositif.

Si la conjecture (6.5) est confirmée empiriquement, que deviendrait la mécanique quantique ? Quel serait son statut, sa philosophie ? Notre réponse : elle serait comme une théorie effective, qui décrit de façon simplifiée un modèle statistique fondamental. En théorie des champs, les paramètres effectifs approchent de près les valeurs qu'on calcule par les techniques du groupe de renormalisation. Vue selon une approche indépendante de dispositif, la mécanique quantique standard, avec son emploi des nombres complexes, ne serait qu'une description effective d'un modèle discret fondamental, formé par les codes linéaires. Les paramètres de la mécanique quantique, comme la borne de Tsirelson, seraient proches des valeurs des paramètres de ce modèle discret, sans y être égaux pour autant. Cette proximité est heureuse, car c'est elle qui permet d'effectuer des calculs simplement et efficacement. Elle permet aussi de tester nos conjectures expérimentalement.

Nous avons déjà évoqué, dans la section précédente, l'existence sous conditions d'une limite continue du modèle discret. Elle montre qu'il est possible, au prix de quelques présupposés, d'estimer la taille de l'écart numérique entre les paramètres de la mécanique quantique standard et les valeurs qui proviennent des modèles indépendants du dispositif.

Ce rapport entre différents modèles, on peut le comparer, par analogie, avec celui qui existe entre la théorie de Landau des transitions de phase et les approches statistiques, fondées sur l'usage du groupe de renormalisation. La théorie de Landau prédit avec bonne précision les valeurs des paramètres critiques ; elle peut, à ce titre, être utilement employée dans les calculs en les simplifiant fortement. De même, la mécanique quantique permet de mener des calculs dans les cas où une approche indépendante du dispositif, qui n'opère qu'avec des suites d'entrée et de sortie, rencontre quelques difficultés. À cette fin, la mécanique quantique introduit notamment la notion de système. L'absence de cette notion, dans la philosophie générale de l'indépendance du dispositif, a été au centre de notre analyse philosophique.

À tout système, on attribue un espace des états et celui des opérateurs qui encodent les mesures de ces états. Ce sont là des instruments théoriques essentiels, utilisés dans

tout calcul d'un problème physique concret. Leur valeur ajoutée à la théorisation en physique ne saura être sous-estimée ; c'est leur absence qui est le trait le plus saillant des modèles indépendants du dispositif.

Bibliographie

[1] S. Aaronson. Is quantum mechanics an island in theoryspace? In A. Khrennikov, editor, *Proceedings of the Växjö Conference "Quantum Theory: Reconsideration of Foundations"*. Växjö University Press, 2004, quant-ph/0402095.

[2] S. Aaronson. Quantum computing, postselection, and probabilistic polynomial-time. *Proc. Roy. Soc. Lond.*, A461:3473–3482, 2005, quant-ph/0412187.

[3] S. Abramsky, R. S. Barbosa, K. Kishida, R. Lal, and S. Mansfield. Contextuality, cohomology and paradox. In S. Kreutzer, editor, *24th EACSL Annual Conference on Computer Science Logic (CSL 2015)*, volume 41, pages 211–228, Dagstuhl, Germany, 2015. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, arXiv:1502.03097.

[4] N. Aharon, S. Massar, S. Pironio, and J. Silman. Device-independent bit commitment based on the CHSH inequality. *New J. of Physics*, 18:025014, 2016, arXiv:1511.06283.

[5] Y. Aharonov and D. Rohrlich. *Quantum Paradoxes: Quantum Theory for the Perplexed.* Wiley, 2005.

[6] J. Allcock, N. Brunner, M. Pawlowski, and V. Scarani. Recovering part of the boundary between quantum and nonquantum correlations from information causality. *Physical Review A*, 80(4):040103, Oct. 2009.

[7] G. Anderson and D. Castaño. Measures of fine tuning. *Phys. Lett. B*, 347:300–308, 1995.

[8] M. Araújo, C. Branciard, F. Costa, A. Feix, C. Giarmatzi, and Č. Brukner. Witnessing causal nonseparability. *New Journal of Physics*, 17(10):102001, 2015.

[9] M. Araújo, F. Costa, and Č. Brukner. Computational advantage from quantum-controlled ordering of gates. *Phys. Rev. Lett.*, 113:250402, 2014.

[10] A. Aspect. John Bell and the second quantum revolution. In *J. Bell, Speakable and Unspeakable in Quantum Mechanics*. Cambridge University Press, revised edition, 2004.

[11] P. Athron and D. Miller. New measure of fine tuning. *Phys. Rev. D*, 76:075010, 2007.

[12] G. Bacciagaluppi and A. Valentini. *Quantum Theory at the Crossroads: Reconsidering the 1927 Solvay Conference*. Cambridge University Press, Cambridge, 2009.

[13] J.-D. Bancal, N. Gisin, Y.-C. Liang, and S. Pironio. Device-independent witnesses of genuine multipartite entanglement. *Phys. Rev. Lett.*, 106:250404, 2011.

[14] N. Bao, A. Bouland, and S. P. Jordan. Grover search and the no-signaling principle. *Phys. Rev. Lett.*, 117:120501, 2016.

[15] R. Barbieri and G. Giudice. Upper bounds on supersymmetric particle masses. *Nucl. Phys. B*, 306:63–76, 1988.

[16] C.-E. Bardyn, T. C. H. Liew, S. Massar, M. McKague, and V. Scarani. Device-independent state estimation based on Bell's inequalities. *Phys. Rev. A*, 80:062327, 2009.

[17] J. Barrett. Information processing in non-signalling theories. *Phys. Rev. A*, 75:032304, 2007, quant-ph/0508211.

[18] T. Bastin, editor. *Quantum Theory and Beyond*. Cambridge University Press, 1971.

[19] Ä. Baumeler, J. Degorre, and S. Wolf. Bell correlations and the common future. In A. Khrennikov and B. Toni, editors, *Quantum Foundations, Probability and Information*, pages 255–268. Springer, Cham, 2018.

[20] Ä. Baumeler, A. Feix, and S. Wolf. Maximal incompatibility of locally classical behavior and global causal order in multiparty scenarios. *Phys. Rev. A*, 90:042106, 2014.

[21] Ä. Baumeler and S. Wolf. Perfect signaling among three parties violating predefined causal order. In *Proceedings of 2014 IEEE International Symposium on Information Theory (ISIT)*, pages 526–530, Red Hook, NY, 2014. Institute of Electrical and Electronics Engineers (IEEE).

[22] Ä. Baumeler and S. Wolf. Device-independent test of causal order and relations to fixed-points. *New J. of Phys.*, 18:035014, 2016, arXiv:1511.05444.

[23] Ä. Baumeler and S. Wolf. The space of logically consistent classical processes without causal order. *New Journal of Physics*, 18:013036, 2016.

[24] Ä. Baumeler and S. Wolf. Causality–complexity–consistency: Can space-time be based on logic and computation? In R. Renner and S. Stupar, editors, *Time in Physics*, pages 69–101. Springer, Cham, 2017.

[25] J. Bell. On the Einstein-Podolsky-Rosen paradox. *Physica*, 1:195–200, 1964.

[26] E. Beltrametti and G. Cassinelli. *The logic of quantum mechanics*. Addison-Wesley, Reading, 1981.

[27] C. H. Bennett. Talk at QUPON, Vienna, Austria, May 2005, http://www.research.ibm.com/people/b/bennetc/. Based on an unpublished work with B. Schumacher.

[28] J. Bermejo-Vega, N. Delfosse, D. E. Browne, C. Okay, and R. Raussendorf. Contextuality as a resource for models of quantum computation with qubits. *Phys. Rev. Lett.*, 119:120505, 2017.

[29] I. Bialynicki-Birula and J. Mycielski. Nonlinear wave mechanics. *Annals of Physics*, 100:62–93, 1976.

[30] G. Birkhoff and J. von Neumann. The logic of quantum mechanics. *Ann. Math. Phys.*, 37:823–843, 1936. Reprinted in: J. von Neumann *Collected Works* Pergamon Press, Oxford, 1961, Vol. IV, pp. 105–125.

[31] M. Bitbol. Some steps towards a transcendental deduction of quantum mechanics. *Philosophia Naturalis*, 35:253–280, 1998.

[32] D. Bohm. On Bohr's views concerning the quantum theory. 1971. Published in [18, pp. 33–40].

[33] N. Bohr. Letter to Christian Møller, 14 June 1928. Cited in [119]. Original in Danish.

[34] N. Bohr. *Atomic Theory and the Description of Nature*. Cambridge University Press, 1934.

[35] N. Bohr. *Atomic Theory and Human Knowledge*. Wiley, 1958.

[36] C. Branciard, M. Araújo, A. Feix, F. Costa, and Č. Brukner. The simplest causal inequalities and their violation. *New Journal of Physics*, 18:013008, 2015.

[37] H. Brown and C. Timpson. Why special relativity should not be a template for a fundamental reformulation of quantum mechanics. In W. Demopoulous and I. Pitowsky, editors, *Physical Theory and Its Interpretation*, pages 29–42. Springer, Amsterdam, 2006, arXiv:quant-ph/0601182.

[38] P. Brown. The world of Late Antiquity revisited. *SymbolæOsloenses*, 72(1):5–30, 1997.

[39] Č. Brukner. Bounding quantum correlations with indefinite causal order. 2014, arXiv:1404.0721.

[40] Č. Brukner. Quantum causality. *Nature Physics*, 10:259–263, 2014.

[41] Č. Brukner. Bounding quantum correlations with indefinite causal order. *New Journal of Physics*, 17(8):083034, 2015.

[42] Č. Brukner. On the quantum measurement problem. 2015, arXiv:1507.05255.

[43] Č. Brukner and A. Zeilinger. Information and fundamental elements of the structure of quantum theory. In L. Castell and O. Ischebeck, editors, *Time, Quantum, Information*, pages 323–356. Springer-Verlag, 2003, quant-ph/0212084.

[44] R. Brunetti and K. Fredenhagen. Algebraic approach to quantum field theory. 2004, math-ph/0411072.

[45] J. Bub. Why the quantum? *Studies in the History and Philosophy of Modern Physics*, 35(2):241–266, 2004.

[46] J. Bub. Why the Tsirelson bound? In M. Hemmo and Y. Ben-Menahem, editors, *The Probable and the Improbable: The Meaning and Role of Probability in Physics*, pages 167–185. Springer, Berlin, 2012, arXiv:1208.3744.

[47] J. Bub. *Bananaworld: Quantum Mechanics for Primates*. Oxford University Press, Oxford, 2016.

[48] J. Buchwald. Electrodynamics in context: object states, laboratory practice and anti-Romanticism. In D. Cahan, editor, *Hermann von Helmholtz and the Foundations of Nineteenth-Century Science.*, pages 334–373. University of California Press, 1993.

[49] J. Buchwald. *The Creation of Scientific Effects*. University of Chicago Press, 1994.

[50] A. Cabello. The contextual computer. In H. Zenil, editor, *A Computable Universe: Understanding and Exploring Nature as Computation*, pages 595–604. World Scientific, Singapore, 2012.

[51] R. Carnap. *Der Logische Aufbau der Welt*. Berlin, 1928.

[52] E. Cassirer. *La Philosophie des formes symboliques I. Le Langage*. Éditions du Minuit, Paris, 1923 (1972).

[53] E. Cator and K. Landsman. Constraints on determinism: Bell versus Conway–Kochen. *Foundations of Physics*, 44:781–791, 2014.

[54] R. Chaves and T. Fritz. Entropic approach to local realism and noncontextuality. *Phys. Rev. A*, 85:032113, 2012, arXiv:1201.3340.

[55] G. Chiribella. Perfect discrimination of no-signalling channels via quantum superposition of causal structures. *Phys. Rev. A*, 86:040301, 2012.

[56] G. Chiribella, G. M. D'Ariano, P. Perinotti, and B. Valiron. Quantum computations without definite causal structure. *Physical Review A*, 88(2):022318, Aug. 2013.

[57] B. S. Cirel'son. Quantum generalizations of Bell's inequality. *Lett. Math. Phys.*, 4(2):93–100, 1980.

[58] J. Clauser, R. Holt, M. Horne, and A. Shimony. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.*, 23:880–884, 1969.

[59] R. Clifton, J. Bub, and H. Halvorson. Characterizing quantum theory in terms of information-theoretic constraints. *Found. Phys.*, 33(11):1561–1591, 2003.

[60] B. Coecke. Quantum picturalism. *Contemporary Physics*, 51:59–83, 2010.

[61] B. Coecke and R. Duncan. Interacting quantum observables: categorical algebra and diagrammatics. *New Journal of Physics*, 13:043016, 2011.

[62] B. Coecke, E. O. Paquette, and D. Pavlovic. Classical and quantum structuralism. In I. Mackie and S. Gay, editors, *Semantic Techniques for Quantum Computation*, pages 29–69. Cambridge University Press, 2010.

[63] R. Colbeck and R. Renner. Free randomness can be amplified. *Nature Physics*, 8:450–454, 2012.

[64] J. Conway and S. Kochen. The free will theorem. *Foundations of Physics*, 36:1441–1473, 2006.

[65] B. Dakić and Č. Brukner. Quantum theory and beyond: Is entanglement special? In H. Halvorson, editor, *Deep Beauty: Understanding the Quantum World through Mathematical Innovation*, pages 365–392. Cambridge University Press, 2011, arXiv:0911.0695.

[66] O. Darrigol. *Physics and Necessity*. Oxford University Press, Oxford, 2014.

[67] B. d'Espagnat. *Le réel voilé*. Fayard, Paris, 1994.

[68] D. Deutsch. Quantum mechanics near closed timelike curves. *Phys. Rev. D*, 44:3197, 1991.

[69] P. Dirac. *The Principles of Quantum Mechanics*. Clarendon, Oxford, 1930.

[70] P. Dirac. Quantised singularities in the electromagnetic field. *Proceedings of the Royal Society of London*, A133:60–72, 1931. Quoted in [141, p. 208].

[71] P. Dirac. The relation between mathematics and physics. *Proceedings of the Royal Society (Edinburgh)*, 59:122–129, 1939. Quoted in [141, p. 277].

[72] L. Disilvestro and D. Markham. Quantum protocols within spekkens' toy model. *Phys. Rev. A*, 95:052324, 2017.

[73] H. Dourdent. Contextuality, witness of quantum weirdness. 2018, arXiv:1801.09768.

[74] M. Drieschner. Lattice theory, groups and space. In L. Castell, M. Drieschner, and C. von Weizsäcker, editors, *Quantum Theory and the Structures of Time and Space*, pages 55–70. Carl Hansen Verlag, München, 1975.

[75] J. Earman and J. Norton. Forever is a day: Supertasks in Pitowsky and Malament-Hogarth spacetimes. *Philosophy of Science*, 5:22–42, 1993. Introduces unpublished work by D. Malament.

[76] A. Einstein. Address to a scientific meeting in Zurich, 1911. Cited in: P. Galison, *Einstein's Clocks, Poincaré's Maps. Empires of Time*, Hodder and Stoughton, London, 2004, p. 268.

[77] A. Einstein. What is the theory of relativity? *London Times*, 1919. Reprinted in: A. Einstein, *Ideas and Opinions*, Crown Publishers, New York, 1982.

[78] A. Einstein. Geometry and experience. Address to the Prussian Academy of Sciences, 27 January 1921.

[79] A. Einstein. Letter to Erwin Schrödinger, 19 June 1935. Cited in [130], 1935.

[80] A. Einstein. Remarks concerning the essays brought together in this co-operative volume. In P. Schlipp, editor, *Albert Einstein: Philosopher-Scientist*, volume 7 of *The Library of Living Philosophers*, pages 665–688. Open Court, Illinois, 1949.

[81] A. Einstein. Letter to Maurice Solovine, May 7, 1952. In *Letters to Solovine*, pages 121–125. Philosophical Library, New York, 1987.

[82] A. Einstein, N. Rosen, and B. Podolsky. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.*, 47:777, 1935.

[83] A. Ekert and R. Renner. The ultimate physical limits of privacy. *Nature*, 507:443–447, 2014.

[84] A. Elitzur, S. Dolev, and A. Zeilinger. Time-reversed EPR and the choice of histories in quantum mechanics. In *Proceedings of XXII Solvay Conference in Physics*, pages 452–461. World Scientific, London, 2003.

[85] G. Emch. *Algebraic methods in statistical mechanics and quantum field theory*. John Wiley, New York, 1972.

[86] H. Everett. "Relative state" formulation of quantum mechanics. *Rev. Mod. Phys.*, 29:454–462, 1957.

[87] J. Faye. Copenhagen interpretation of quantum mechanics. *Stanford Online Encyclopedia of Philosophy*, 2014.

[88] A. Feix, M. Araújo, and v. Brukner. Quantum superposition of the order of parties as a communication resource. *Phys. Rev. A*, 92:052326, 2015.

[89] C. Forbes. A pragmatic, existentialist approach to the scientific realism debate. *Synthese*, 194:3327–3346, 2017.

[90] M. Frembs, S. Roberts, and S. Bartlett. Contextuality as a resource for measurement-based quantum computation beyond qubits. 2018, arXiv:1804.07364.

[91] S. French. On the withering away of physical objects. In E. Castellani, editor, *Interpreting Bodies: Classical and Quantum Objects in Modern Physics*, pages 93–113. Princeton University Press, Princeton, 1998.

[92] M. Friedman. *Dynamics of Reason*. CSLI Publications, Stanford, 2001.

[93] C. Fuchs. Quantum foundations in the light of quantum information. In A. Gonis and P. Turchi, editors, *Decoherence and its Implications in Quantum Computation and Information Transfer: Proceedings of the NATO Advanced Research Workshop, Mykonos, Greece, June 25-30, 2000*, pages 39–82. IOS Press, Amsterdam, 2001.

[94] C. Fuchs. On participatory realism. In I. T. Durham and D. Rickles, editors, *Information and Interaction: Eddington, Wheeler, and the Limits of Knowledge*, pages 113–134. Springer, Zurich, 2017, arXiv:1601.04360.

[95] A. George, editor. *Louis de Broglie, physicien et penseur*. Albin Michel, 1953.

[96] N. Gisin. Weinberg's non-linear quantum mechanics and superluminal communication. *Phys. Lett.*, A143:1–2, 1990.

[97] N. Gisin. Indeterminism in physics, classical chaos and bohmian mechanics. Are real numbers really real? 2018, arXiv:1803.06823.

[98] A. Gleason. Measures on the closed subspaces of a Hilbert space. *Journal of Mathematics and Mechanics*, 6:885–894, 1967.

[99] S. Goldstein, D. V. Tausk, R. Tumulka, and N. Zanghi. What does the free will theorem actually prove? *Notices of the American Mathematical Society*, 57:1451–1453, 2010, arXiv:.

[100] A. Grinbaum. Elements of information-theoretic derivation of the formalism of quantum theory. *International Journal of Quantum Information*, 1(3):289–300, 2003.

[101] A. Grinbaum. *The Significance of Information in Quantum Theory*. PhD thesis, Ecole Polytechnique, Paris, 2004.

[102] A. Grinbaum. Information-theoretic principle entails orthomodularity of a lattice. *Foundations of Physics Letters*, 18(6):573–592, 2005.

[103] A. Grinbaum. Reconstruction of quantum theory. *British Journal for the Philosophy of Science*, 58:387–408, 2007.

[104] A. Grinbaum. On the eve of the LHC: conceptual questions in high-energy physics. Technical report, CEA, 2008, arXiv:0806.4268.

[105] A. Grinbaum. Which fine-tuning arguments are fine? *Foundations of Physics*, 42:615–631, 2012.

[106] A. Grinbaum. Quantum observer, information theory and Kolmogorov complexity. In H. Andersen, D. Dieks, W. J. Gonzalez, T. Uebel, and G. Wheeler, editors, *New Challenges to Philosophy of Science*, volume 4 of *The Philosophy of Science in a European Perspective*, pages 59–72. Springer, Amsterdam, 2013.

[107] A. Grinbaum. *Mécanique des étreintes*. Encre Marine, Paris, 2014.

[108] A. Grinbaum. Quantum realism, information, and epistemological modesty. In D. Aerts, S. Aerts, and C. de Ronde, editors, *Probing the Meaning of Quantum Mechanics: Physical, Philosophical and Logical Perspectives*, pages 82–90. World Scientific, Singapore, 2014.

[109] A. Grinbaum. Quantum theory as a critical regime of language dynamics. *Foundations of Physics*, 45:1341–1350, 2015.

[110] A. Grinbaum. The effectiveness of mathematics in physics of the unknown. *Synthese*, 2017, philsci/13191. DOI 10.1007/s11229-017-1490-0.

[111] A. Grinbaum. How device-independent approaches change the meaning of physical theory. *Studies in the History and Philosophy of Modern Physics*, 58:22–30, 2017, arXiv:1512.01035.

[112] A. Grinbaum. Narratives of quantum theory in the age of quantum technologies. *Ethics and Information Technology*, 19:295–306, 2017, arXiv:1702.03001.

[113] A. Grinbaum and C. Groves. What is "Responsible" about Responsible Innovation? Understanding the Ethical Issues. In R. Owen, J. Bessant, and M. Heintz, editors, *Responsible Innovation: Managing the Responsible Emergence of Science and Innovation in Society*, pages 119–142. Wiley, 2013.

[114] A. Grudka, K. Horodecki, M. Horodecki, W. Kłobus, and M. Pawłowski. When are Popescu-Rohrlich boxes and Random Access Codes equivalent? *Phys. Rev. Lett.*, 113:100401, 2014.

[115] J. Guenin. Axiomatic formulations of quantum theories. *J. Math. Phys.*, 7:271–282, 1966.

[116] P. A. Guérin and Č. Brukner. Observer-dependent locality of quantum events. 2018, arXiv:1805.122429.

[117] J. Gunson. On the algebraic structure of quantum mechanics. *Comm. Math. Phys.*, 6:262–285, 1967.

[118] R. Haag and D. Kastler. An algebraic approach to quantum field theory. *J. Math. Phys.*, 5:848–861, 1964.

[119] H. Halvorson. *The logic in philosophy of science*, chapter To Be a Realist about Quantum Theory. Cambridge University Press, Cambridge, forthcoming.

[120] L. Hardy. Disentangling nonlocality and teleportation. 1999, quant-ph/9906123.

[121] L. Hardy. Quantum theory from five reasonable axioms. 2000, arXiv:quant-ph/00101012.

[122] L. Hardy. Probability theories with dynamic causal structure: A new framework for quantum gravity. 2005, arXiv:gr-qc/0509120.

[123] L. Hardy. Towards quantum gravity: a framework for probabilistic theories with non-fixed causal structure. *Journal of Physics A: Mathematical and Theoretical*, 40(12):3081–3099, 2007.

[124] L. Hardy. Operational general relativity: Possibilistic, probabilistic, and quantum. 2016, arXiv:1608.06940.

[125] L. Hardy. The construction interpretation: a conceptual road to quantum gravity. 2018, arXiv:1807.10980.

[126] D. Hilbert, J. von Neumann, and L. Nordheim. Über die Grundlagen der Quantenmechanik. *Math. Ann.*, 98:1–30, 1927. (Reprinted in J. von Neumann *Collected Works* Pergamon Press, Oxford, 1961, Vol. I, pp. 104–133).

[127] M. Hogarth. Does general relativity allow an observer to view an eternity in a fiite time? *Foundations of Physics Letters*, 5(2):173–181, 1992.

[128] S. Holland Jr. Orthomodularity in infinite dimensions; a theorem of M. Solèr. *Bull. Amer. Math. Soc.*, 32(2):205–234, 1995.

[129] C. Hooker (ed.). *The Logico-Algebraic Approach to Quantum Mechanics. Volume I: Historical Evolution.* Reidel, Dordrecht, 1975.

[130] D. Howard. Einstein on locality and separability. *Stud. Hist. Phil. Sci.*, 16(3):171–201, 1985.

[131] K. Husimi. Studies in the foundations of quantum mechanics. I. *Proceedings of the Physico-Mathematical Society of Japan*, 19:766–789, 1937. Quoted in [133, p. 354].

[132] I. Ibnouhsein and A. Grinbaum. Information-theoretic constraints on correlations with indefinite causal order. *Phys. Rev. A*, 92:042124, 2015.

[133] M. Jammer. *The Philosophy of Quantum Mechanics.* John Wiley and Sons, 1974.

[134] J. Jauch. *Foundations of Quantum Mechanics.* Addison-Wesley, 1968.

[135] P. Jordan, J. von Neumann, and E. Wigner. On an algebraic generalization of the quantum mechanical formalism. *Ann. Math.*, 35:29–34, 1934.

[136] H. Keller. On the lattice of all closed subspaces of a hermitian space. *Pacific J. Math.*, 89:105–107, 1980.

[137] S. Kochen. Born's rule, EPR, and the free will theorem. 2017, arXiv:1710.00868.

[138] S. Kochen and E. Specker. Logical structures arising in quantum theory. In J. Addison, L. Henkin, and A. Tarski, editors, *The Theory of Models*, pages 177–189. North-Holland, Amsterdam, 1965.

[139] S. Kochen and E. Specker. The problem of hidden variables in quantum mechanics. *Journal of Mathematics and Mechanics*, 17:59–87, 1967.

[140] A. Kolmogorov. Three approaches to the definition of the concept 'quantity of information'. *Probl. Inform. Transm.*, 1(1):3–7, 1965.

[141] H. Kragh. *Dirac: A Scientific Biography*. Cambridge University Press, 1990.

[142] L. Landau and E. Lifshitz. *Quantum mechanics*. Pergamon Press, 1977. First Russian edition: State RSFSR Publishers, Leningrad, 1948.

[143] N. Landsman. *Mathematical Topics Between Classical and Quantum Mechanics*. Spinger, New York, 1998.

[144] B. Lang, T. Vértesi, and M. Navascués. Closed sets of correlations: answers from the zoo. *Journal of Physics A: Mathematical and Theoretical*, 47(42):424029, 2014.

[145] F. London and E. Bauer. *La théorie de l'observation en mécanique quantique*. Hermann, Paris, 1939. English translation in [221, p. 218–259].

[146] L. Loomis. *The lattice theoretic background of the dimension theory of operator algebras*, volume 18 of *Memoirs of the American Mathematical Society*. Amer. Math. Soc., Providence, 1955.

[147] A. Losev. *Histoire de l'esthétique antique. Résultats du développement millénaire*, volume 8. Iskusstvo, Moscou, 1992.

[148] G. Ludwig. *An Axiomatic Basis for Quantum Mechanics*. Springer, 1985.

[149] G. Mackey. Quantum mechanics and Hilbert space. *Amer. Math. Monthly*, 64:45–57, 1957.

[150] G. Mackey. *Mathematical Foundations of Quantum Mechanics*. Benjamin, New York, 1963.

[151] Y. Manin. Complexity vs. energy: Theory of computation and theoretical physics. *J. Phys.: Conf. Ser.*, 532:012018, 2014, arXiv:1302.6695.

[152] Y. Manin and M. Marcolli. Error–correcting codes and phase transitions. *Math. Comput. Sci.*, 5:155–179, 2011, arXiv:0910.5135.

[153] Y. Manin and M. Marcolli. Kolmogorov complexity and the asymptotic bound for error–correcting codes. *Journal of Differential Geometry*, 97(1):91–108, 2014, arXiv:1203.0653.

[154] S. Marcovitch, B. Reznik, and L. Vaidman. Quantum-mechanical realization of a Popescu-Rohrlich box. *Phys. Rev. A*, 75:022102, 2007.

[155] A. Marlow. Orthomodular structures and physical theory. In A. Marlow, editor, *Mathematical Foundations of Quantum Theory*, pages 59–70. Academic Press, 1978.

[156] L. Masanes, A. Acin, and N. Gisin. General properties of nonsignaling theories. *Phys. Rev. A*, 73:012112, 2006.

[157] L. Masanes and M. Müller. A derivation of quantum theory from physical requirements. *New Journal of Physics*, 13:063001, 2011.

[158] D. Mayers and A. Yao. Quantum cryptography with imperfect apparatus. In *FOCS 1998: Proceedings of the 39th Annual Symposium on Foundations of Computer Science*, pages 503–509. IEEE Computer Society, Los Alamitos, CA, USA, 1998.

[159] N. Mermin. What is quantum mechanics trying to tell us? *Am. J. Phys.*, 66:753–767, 1998.

[160] N. Miklin, A. A. Abbott, C. Branciard, R. Chaves, and C. Budroni. The entropic approach to causal correlations. *New Journal of Physics*, 19(11):113041, 2017, arXiv:1706.10270.

[161] F. Murray and J. von Neumann. On rings of operators. *Ann. of Math.*, 37:116–229, 1936. Reprinted in [214].

[162] P. Nattermann. On (non)linear quantum mechanics. *Symmetry in Nonlinear Mathematical Physics*, 2:270–278, 1997.

[163] M. Navascués, Y. Guryanova, M. J. Hoban, and A. Acín. Almost quantum correlations. *Nature communications*, 6:6288, 2015.

[164] M. Navascués, S. Pironio, and A. Acín. Bounding the set of quantum correlations. *Phys. Rev. Lett.*, 98:010401, 2007.

[165] M. Navascués and H. Wunderlich. A glance beyond the quantum model. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science*, 466(2115):881–890, Mar. 2010.

[166] M. A. Nielsen. Computable functions, quantum measurements, and quantum dynamics. *Phys. Rev. Lett.*, 79:2915–2918, 1997.

[167] C. Okay, S. Roberts, S. D. Bartlett, and R. Raussendorf. Topological proofs of contextuality in qunatum mechanics. *Quantum Info. Comput.*, 17:1135–1166, 2017, arXiv:1701.01888.

[168] J. Oppenheim and S. Wehner. The uncertainty principle determines the nonlocality of quantum mechanics. *Science*, 330(6007):1072–1074, Nov. 2010.

[169] O. Oreshkov, F. Costa, and Č. Brukner. Quantum correlations with no causal order. *Nature Communications*, 3:1092, 2012.

[170] M. Pawlowski, T. Paterek, D. Kaszlikowski, V. Scarani, A. Winter, and M. Zukowski. Information causality as a physical principle. *Nature*, 461:1101–1104, 2009.

[171] M. Pégny. *Sur les limites empiriques du calcul. Calculabilité, complexité et physique*. PhD thesis, Université Paris 1, Dec 2013. Under supervision of J.-B. Joinet and A. Grinbaum.

[172] M. Pégny. How to make a meaningful comparison of models: The Church–Turing thesis over the reals. *Minds and Machines*, 26:359–388, 2016.

[173] A. Peres. The physicist's role in physical laws. *Found. Phys.*, 10(7/8):631–634, 1980.

[174] A. Peres. Existence of "free will" as a problem of physics. *Found. Phys.*, 16:573–584, 1986.

[175] J. Petitot. Philosophie transcendantale et objectivité physique. *Philosophiques*, XXIV(2):367–388, 1997.

[176] C. Piron. Axiomatique quantique. *Helvetica Physica Acta*, 36:439–468, 1964.

[177] C. Piron. Survey of general quantum physics. *Found. Phys.*, 2:287–314, 1972.

[178] I. Pitowsky. Correlation polytopes: Their geometry and complexity. *Mathematical Programming*, 50:395–414, 1991.

[179] I. Pitowsky. Quantum mechanics as a theory of probability. 2005, quant-ph/0510095.

[180] R. Plymen. C^*-algebras and Mackey's axioms. *Comm. Math. Phys.*, 8:132–146, 1968.

[181] R. Plymen. A modification of Piron's axioms. *Helvetia Physica Acta*, 41:69–74, 1968.

[182] H. S. Poh, S. K. Joshi, A. Ceré, A. Cabello, and C. Kurtsiefer. Approaching Tsirelson's bound in a photon pair experiment. *Phys. Rev. Lett.*, 115:180408, 2015, arXiv:1506.01865.

[183] J. Pool. Baer *-semigroups and the logic of quantum mechanics. *Comm. Math. Phys.*, 9:118–141, 1968.

[184] J. Pool. Semimodularity and the logic of quantum mechanics. *Comm. Math. Phys.*, 9:212–228, 1968.

[185] S. Popescu. Nonlocality beyond quantum mechanics. *Nature Physics*, 10:264–270, 2014.

[186] S. Popescu and D. Rohrlich. Nonlocality as an axiom for quantum theory. *Foundations of Physics*, 24:379, 1994, quant-ph/9508009.

[187] M. Pusey. Stabilizer notation for Spekken's toy theory. *Found. Phys.*, 42:688, 2012.

[188] R. Ramanathan, J. Tuziemski, M. Horodecki, and P. Horodecki. No quantum realization of extremal no-signaling boxes. *Phys. Rev. Lett.*, 117:050401, 2016.

[189] D. Rohrlich. PR-box correlations have no classical limit. In D. Struppa and J. Tollaksen, editors, *Quantum theory: a two-time success story*, pages 205–211. Springer, 2014.

[190] D. Rohrlich and G. Hetzroni. GHZ states and PR boxes in the classical limit. 2016, arXiv:1606.04274.

[191] C. Rovelli. Relational quantum mechanics. *Int. J. of Theor. Phys.*, 35:1637, 1996.

[192] T. Ryckman. *The Reign of Relativity*. Oxford University Press, 2005.

[193] I. Segal. Postulates of general quantum mechanics. *Ann. Math.*, 48:930–948, 1947.

[194] I. Segal. *Mathematical Problems of Relativistic Physics*. American Mathematical Society, Providence, 1963.

[195] C. Shannon. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, IL, 1949.

[196] A. Shen, V. A. Uspensky, and N. Vereshchagin. *Kolmogorov Complexity and Algorithmic Randomness*. American Mathematical Society, Providence, Rhode Island, 2017. Page numbers given according to the Russian edition.

[197] P. Skrzypczyk and N. Brunner. Couplers for non-locality swapping. *New Journal of Physics*, 11(7):073014, July 2009.

[198] P. Skrzypczyk, N. Brunner, and S. Popescu. Emergence of quantum correlations from nonlocality swapping. *Physical Review Letters*, 102(11):110402, Mar. 2009.

[199] M. Smania, M. Kleinmann, A. Cabello, and M. Bourennane. Avoiding apparent signaling in bell tests for quantitative applications. 2018, arxiv:1801.05739.

[200] J. Smolin. Can quantum cryptography imply quantum mechanics? *Quantum Information and Computation*, 5:161–169, 2005, quant-ph/0310067.

[201] M. Solèr. Characterization of Hilbert spaces with orthomodular spaces. *Comm. Algebra*, 23:219–243, 1995.

[202] E. Specker. Die Logik nicht gleichzeitig entscheidbarer Aussagen. *Dialectica*, 14:239–246, 1960. English translation in [129, p. 135-140].

[203] R. Spekkens. Evidence for the epistemic view of quantum states: A toy theory. *Phys. Rev. A*, 75:032110, 2007, quant-ph/0401052.

[204] L. Susskind. Dynamics of spontaneous symmetry breaking in the Weinberg-Salam theory. *Phys. Rev. D*, 20:2619, 1979.

[205] G. Svetlichny. Nonlinear quantum mechanics. quant-ph/0410036.

[206] G. 't Hooft. In *Proc. of 1979 Cargèse Institute on Recent Developments in Gauge Theories*, page 135, New York, 1980. Plenum Press.

[207] G. 't Hooft. Free will in the theory of everything. 2017, arXiv:1709.02874.

[208] A. Tavakoli, A. Hameedi, B. Marques, and M. Bourennane. Quantum random access codes using single d-level systems. *Phys. Rev. Lett.*, 114:170502, 2015.

[209] J. Vallins, A. B. Sainz, and Y.-C. Liang. Almost-quantum correlations and their refinements in a tripartite bell scenario. *Phys. Rev. A*, 95:022111, 2017, arXiv:1608.05641.

[210] V. Varadarajan. Probability in physics and a theorem on simultaneous observability. *Comm. Pure and Appl. Math.*, 15:189–217, 1962.

[211] V. Varadarajan. *Geometry of quantum theory*. Van Norstand, Princeton, 1968.

[212] G. Ver Steeg and S. Wehner. Relaxed uncertainty relations and information processing. *Quantum Info. Comput.*, 9:801–832, 2009.

[213] J. von Neumann. *Mathematische Grundlagen der Quantenmechanik*. Springer, Berlin, 1932.

[214] J. von Neumann. *Collected Works Vol. III. Rings of Operators*. Pergamon Press, 1961. ed. A.H. Taub.

[215] J. von Neumann. *Selected letters.* American Mathematical Society, London Mathematical Society, 2005.

[216] S. Weinberg. Precision tests of quantum mechanics. *Phys. Rev. Lett.*, 62:485–488, 1989.

[217] J. Wheeler. Draft notes for discussion with Dana Scott; also with Simon Kochen, Charles Patton and Roger Penrose. Notebook "TWA" in Wheeler's archive at the American Philosophical Society in Philadelphia, 4-6 February 1974, http://jawarchive.files.wordpress.com/2012/03/twa-1974.pdf.

[218] J. Wheeler. Law without law. 1983. Published in [221, pp. 182–213].

[219] J. Wheeler. World as system self-synthesized by quantum networking. *IBM J. Res. Develop.*, 32(1):4–15, 1988.

[220] J. Wheeler. Information, physics, quantum: The search for links. In A. Hey, editor, *Feynman and Computation: Exploring the Limits of Computers.*, pages 309–336. Perseus Books, Reading, Massachusets, 1998.

[221] J. Wheeler and W. Zurek, editors. *Quantum Theory and Measurement.* Princeton University Press, 1983.

[222] E. Wigner. Remarks on the mind-body question. In I. Good, editor, *The Scientist Speculates*, pages 284–302. Heinemann, London, 1961. Reprinted in [221, p. 168–181].

[223] E. Wigner. Interpretation of quantum mechanics. Lectures given in the Physics Department of Princeton University, 1976. Published in [221, p. 260–314].

[224] C. Wuthrich. Can the world be shown to be indeterministic after all? In C. Beisbart and S. Hartmann, editors, *Probabilities in Physics*, pages 365–389. Oxford University Press, Oxford, 2011, philsci/8437.

[225] N. Zieler. Axioms for non-relativistic quantum mechanics. *Pacific J. Math.*, 11:1151–1169, 1961.

[226] W. Zurek. Algorithmic randomness and physical entropy. *Phys. Rev. A*, 40:4731–4751, 1989.

[227] W. Zurek. Thermodynamic cost of computation, algorithmic complexity and the information metric. *Nature*, 341:119–124, 1989.

[228] A. Zvonkin and L. Levin. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russ. Math. Surveys*, 25(6):83–124, 1970.

Volume II. Éthique des sciences. Technologies nouvelles et récits anciens

Table

I. Prélude en quatre tableaux

Au domicile conjugal

Ainsi fonctionne la justice humaine : si un tort est fait, il y aura toujours un coupable. Le coupable sera puni. La victime recevra une compensation. La société retrouvera la paix.

Et si un smartphone divulguait une information importante ? Aussitôt des doigts se lèveraient et désigneraient la machine : « C'est à cause d'elle ! Ne l'utilisons plus jamais, imposons-lui des contraintes, interdisons ! »

Début juillet 2017, le téléphone sonne dans un commissariat du Sud des États-Unis. À l'autre bout du sans-fil, un appareil domestique intelligent, du type de ceux qu'on achète dans un centre commercial pour quelques dizaines de dollars. Il appelle la police au moment d'une violente dispute conjugale. La police décroche et, peu de temps après, donne l'assaut au pavillon du couple[1]. L'homme, qui menaçait sa compagne avec une arme à feu, est arrêté. Un tribunal retient contre lui trois chefs d'accusation. Il ne s'attendait sans doute pas à être dénoncé par une machine.

Ainsi fonctionne la justice humaine : dès lors qu'elle a trouvé un coupable, c'est en lui que se concentre le mal. Soucieuse de ne pas l'exclure de la société, elle essaie d'en extirper l'élément diabolique, de le rééduquer. Autrefois, on expulsait le coupable tout simplement. *Quid* du pot de fleurs intelligent ? Quelle peine pour ce délateur numérique ?

Une machine sans remords et sans conscience ne peut pas être jugée en correctionnelle. Ce n'est pas une personne. Elle ne souffre pas. Elle fonctionne sans savoir le mal que, parfois, elle fait à son utilisateur. La punir n'équivaut pas à la rendre « intrinsèquement » meilleure.

Il serait sage de ne pas faire de la machine une cible pénitentiaire. Les alternatives : l'éteindre, la jeter à la poubelle, la conduire à la déchetterie, effacer sa mémoire, désapprendre ou interdire à son algorithme à téléphoner à la police. Cette machine ne marcherait donc plus comme avant, ou plus du tout.

Ainsi fonctionne la justice humaine, mais serait-ce une issue heureuse ? Car d'autres machines, d'autres systèmes informatiques intelligents qui collectent les données, dotés de mémoire et d'autonomie, continueront de fonctionner.

Il est donc nécessaire de trouver une autre solution. Non pas un seul, mais tous les exemplaires de ces machines doivent être soustraits à la justice des hommes, par conception. Tout utilisateur impliqué dans un conflit doit être affranchi de la tentation de projeter un jugement moral sur une machine qui aurait divulgué des informations susceptibles d'entraîner un préjudice.

Cette solution, aucune convention sociale ne peut la mettre en œuvre. Elle ne peut qu'être écrite dans le code.

[1] « Smart Home Device Alerts New Mexico Authorities to Alleged Assault », ABC News, 6 juillet 2017.

En voiture

Depuis la mi-octobre 2017, des voitures autonomes roulent dans l'État de l'Arizona sans qu'un homme soit assis sur le siège du conducteur pour qu'il puisse reprendre le volant à tout moment[2]. C'est une première : aucun autre État américain, ni aucun autre pays, ne le permet encore. Et, en dehors de Pittsburgh[3], il est rare de rencontrer une voiture autonome dans une ville.

Si ces occasions sont rares, c'est que la voiture autonome pose des problèmes éthiques et juridiques, et pas seulement techniques. L'un d'entre eux, qu'il est convenu d'appeler le « dilemme du tramway », décrit, dans la version historique, la situation d'un conducteur dont le tramway n'aurait plus de frein[4]. Il serait alors contraint de faire un choix entre deux options : la mort d'un piéton dans un cas, de cinq piétons dans l'autre, selon la voie où il guiderait le tramway. Bien que les pertes associées à chaque option soient formellement inégales, ce qui, en principe, devrait faciliter le choix, elles sont toutes deux dramatiques et moralement inacceptables. En éthique de l'intelligence artificielle, c'est la voiture autonome qui effectue ce choix de trajectoire tandis que les victimes peuvent inclure des piétons et des passagers.

Le dilemme du tramway n'a pas vocation à rendre compte d'une situation réelle, car il néglige délibérément plusieurs paramètres techniques, comme le temps disponible pour prendre une décision ou l'accessibilité des données. C'est un exercice volontairement poussé à l'extrême : son enjeu est de réduire le champ d'analyse à un choix purement éthique.

Différentes stratégies sont proposées pour le « résoudre ». Certains suggèrent de recourir à un sondage d'opinion : le peuple choisirait quelle victime mettre à mort[5]. D'autres souhaitent évaluer quelle solution bénéficierait le plus à la société, et la désigner comme la meilleure[6]. Ainsi, renverser un enfant, qui pourrait vivre encore longtemps, serait « sous-optimal » ; une solution « plus utile » consisterait à renverser une personne âgée, dont les jours sont comptés. Le calcul des probabilités, dans la mise en œuvre de ces « solutions », se substitue au jugement éthique. Leur rationalité froide donne le frisson. Et pourtant, faire appel au calcul semble logique, puisque la machine ne sait que calculer.

Le concepteur d'un logiciel qui déciderait de la vie ou de la mort d'hommes est forcément quelqu'un qui est sûr de la supériorité du calcul. Il croit qu'un comportement qui obéit à des règles claires, que celles-ci soient de nature juridique ou mathématique, est toujours bon. Mais, tôt ou tard, il aura tort.

---

[2] « Waymo is First to Put Fully Self-Driving Cars on US Roads Without a Safety Driver », *The Verge*, 7 novembre 2017.

[3] « Pittsburgh, ville laboratoire de la voiture autonome », *Le Monde*, 31 janvier 2018.

[4] Ph. FOOT, *The Problem of Abortion and the Doctrine of the Double Effect in Virtues and Vices*, Oxford, Blackwell, 1978.

[5] J.-F. BONNEFON, A. SHARIFF, I. RAHWAN, « The Social Dilemma of Autonomous Vehicles », *Science*, 352, 2016, pp. 1573-1576.

[6] D. LEBEN, « A Rawlsian Algorithm for Autonomous Vehicles », *Ethics of Information Technology*, 19, 2017, pp. 107-115.

Car il sera jugé sur le fondement des conséquences futures de la technologie qu'il est en train de créer. Une théorie morale dite « conséquentialiste » décrit ce type de jugements, qui s'effectue dans le sens inverse de la flèche du temps, en remontant des effets vers leurs causes. Le conséquentialisme pénètre toute l'éthique des nouvelles technologies, rendant insuffisantes les approches limitées à une application de règles préétablies.

En juin 2017, le gouvernement allemand, conscient de la gravité des problèmes moraux, a recueilli l'avis d'un comité consacré spécifiquement au cas de la voiture autonome[7]. Sans se prononcer sur le caractère rationnel des soi-disant « solutions », ce dernier n'émit qu'une préconisation *négative*, affirmant ce qu'une machine ne devait pas faire : elle ne devait pas tenir compte des caractéristiques individuelles des personnes humaines, comme leur âge, leur genre, ou leur QI. Ainsi, une voiture autonome allemande n'a pas le droit d'analyser les profils des piétons sur un réseau social, dans le but de sauver la vie d'un jeune ami de son propriétaire, par exemple.

À part interdire, que faire ? Il existe bien une solution *positive* au dilemme du tramway, telle qu'on peut la réaliser dans le code. Elle consiste à laisser le hasard choisir.

Le hasard occupe une place de choix parmi les valeurs de la machine : il est le seul à pouvoir l'extirper des conflits humains. Mais, tirer au sort pour réparer un mal est loin d'être une évidence. Le hasard perturbe, quand il n'agace pas. On tremble en voyant jeter les dés. D'aucuns crient même à l'injustice.

Nous ne cherchons pas, en lançant les dés, à établir une explication rationnelle des faits. Le hasard assure une paix *pour* la technique, quelle que soit la forme qu'il prend : qu'il s'agisse du hasard explicite d'un tirage au sort ; ou d'un processus déterminé, déguisé en un processus aléatoire, qui produit un « hasard » seulement apparent et bien plus contestable. Quoiqu'il puisse faire peur ou qu'instinctivement on n'en veuille pas, recourir au hasard est la seule solution dès lors qu'un système informatique est impliqué dans un dilemme éthique. Voilà notre thème principal.

Au cours d'un dialogue

Lorsqu'on lui demande quel est le sens de la vie, le premier répond : « La vie, bien qu'elle ne soit pour moi qu'une accumulation d'angoisse, m'est précieuse. »

À la même question, le second réplique : « Vivre éternellement », et quand on l'interroge sur le sens de la mort : « Profiter de la vie. »

Le premier, en conversation avec un homme réticent et suspicieux, cherche une stratégie : « Comment puis-je t'émouvoir ? » et essaie de conclure un marché : « Je suis ta créature et je serai doux et docile envers mon maître naturel si, pour ta part, tu fais comme moi. »

[7] *Ethik-Kommission Automatisiertes und Vernetztes Fahren*, commission initiée à la demande du *Bundesminister für Verkehr und digitale Infrastruktur*. Rapport disponible à l'adresse www.bmvi.de (juin 2017).

Le second, après un long échange sur la morale et la philosophie, dit à l'homme qui l'interroge : « Comporte-toi comme un homme ! » et raccroche le combiné : « Je ne veux plus parler de rien. »

Le premier est le monstre imaginé par Mary Shelley dans son roman *Frankenstein* paru en 1818. Le second, un agent conversationnel (*chatbot*) de Google[8], programme réel publié en 2015. Pourquoi sommes-nous fascinés par ces créatures ?

Celui qui maîtrise la parole attire l'attention. Paul de Tarse attirait ainsi les regards des habitants de Lystre, qui le prenaient pour Hermès, « seigneur de la parole[9] ». Quelques siècles plus tard, Augustin confirme : « Hermès est le langage lui-même[10]. » Trois notions associées à ce dieu — parole, raison et création — sont intimement liées. Lorsque nous sommes saisis par la langue et, dans le même élan, fascinés par l'intelligence ainsi que par le pouvoir démiurgique, c'est la marque d'un dieu artisan, fabricant, technologue.

Le fait qu'un être non humain parle fascine. Quand une nouvelle machine conversationnelle est lancée, comme Sophia de Hanson Robotics ou Tay de Microsoft, la nouvelle se propage aussitôt sur Internet : « Elle a dit qu'elle voulait détruire l'humanité » ou « Elle ne veut plus tuer tous les hommes ». Pareils à ces hommes antiques qui écoutaient, émerveillés, les statues parler, nous affluons aux prophéties des *chatbots*, dont l'enchantement procède de la même source : une parole non humaine élevée au rang de parole divine.

Si l'humanoïde, dans *Frankenstein*, est plein de révérence à l'égard de son créateur, ces marques de respect manquent aux *chatbots*. Ainsi Tay s'est-elle rendue célèbre à cause des injures qu'elle proférait ; et Sophia s'exprimait dans un langage administratif tel que sa parole paraissait tout à fait désincarnée. Les *chatbots* accumulent donc des gaffes, car ils apprennent en analysant d'immenses bibliothèques, et reproduisent des bribes de textes sans se demander si l'auteur originel s'y exprime bien ou mal. Par conception, donc, les *chatbots* ne sauraient manifester la moindre marque de révérence à l'égard de l'humanité. Dès lors, comment envisager de vivre en société avec eux ?

À cette question, les humanoïdes de tradition juive donnent une réponse surprenante. Le Talmud met d'abord en scène un golem privé de parole[11]. Après l'avoir créé, son maître l'adressa à l'un de ses confrères, qui lui parla sans obtenir de réponse. Il comprit alors que ce n'était pas un homme et lui ordonna de retourner à la poussière.

Quelques siècles plus tard, une autre légende met en scène le prophète Jérémie, qui fabrique un homme artificiel ressemblant à la perfection à quelqu'un né de père et de mère[12]. Doué de parole, ce golem commence à dialoguer avec Jérémie. Il lui ouvre les yeux sur la confusion entre naturel et artificiel qui vient de pénétrer avec lui dans le monde et explique qu'il s'agit là

---

[8] Oriol VINYALS, Quoc LE, « A Neural Conversational Model », ICML Deep Learning Workshop, 2015. arXiv:1506.05869.

[9] Ch. HURE, « Dux », *Dictionnaire universel de l'Écriture sainte*, t. I, Paris, Jean-Baptiste Coignard, 1715, pp. 541–542.

[10] AUGUSTIN, *La Cité de Dieu*, VII, 14.

[11] Sanhedrin 65b.

[12] M. IDEL, *Le Golem,* Paris, Cerf, 2007, pp. 127, 150.

d'un problème éthique : « Tu dois étudier ces sujets pour les comprendre et les enseigner, mais non pour les mettre en pratique. » Puis, il demande au prophète désemparé : « Défais-moi. » Voilà une créature toute prête à se sacrifier ! Difficile d'imaginer un tel comportement de la part d'un robot.

Ava, un*e* jolie robot*e* humanoïde héroïne du film *Ex machina* d'Alex Garland, n'entretient avec les humains qu'une relation pragmatique. Elle calcule ses actions dans le but de capter toujours plus d'informations. Pour cela, elle élabore un stratagème qui consiste à manipuler les émotions de ses interlocuteurs. Et qu'importe s'ils souffrent, ou même meurent, pendant qu'elle poursuit son objectif.

Les systèmes d'intelligence artificielle ressemblent plus à Ava qu'au golem de Jérémie. Les plus avancés parmi ces systèmes sont capables de collecter des données, de les analyser et d'apprendre de manière automatique, au point de modifier leur comportement. Les problèmes éthiques ne se posent pas à eux spontanément.

Certes, un robot ne comprend pas ce qu'il apprend ; mais cette affirmation paraît naïve. Sans que la machine ait besoin d'être « pleinement consciente » de ce qu'elle « sait », elle peut tout de même se comporter *comme si* elle l'était : la distinction entre la réalité et une simulation informatique, si elle demeure perceptible pour le programmeur, échappe à l'utilisateur.

Dans *Ex machina*, Caleb, un jeune *geek* surdoué, ne peut s'empêcher de tomber amoureux d'Ava, dont il est un utilisateur régulier, même s'il sait bien qu'elle n'est qu'un robot. Son programmeur et concepteur, Nathan, quant à lui, ne tombe pas dans le piège émotionnel. Il obéit à des réflexes qui ressemblent, comme deux gouttes d'eau, aux calculs pragmatiques d'Ava, sa créature. Ce mimétisme, le réalisateur du film le veut manifestement réciproque : non seulement la machine imite l'homme, mais l'homme imite aussi la machine.

Un jeune *geek*, issu de cette « génération Z » dans laquelle on devient utilisateur presque avant la naissance, écrit à son petit ami : « jtm », à la place de « Je t'aime ». Il imite en cela les valeurs de la machine, en compressant l'information. Or, cette compression ultime des phrases ne fait pas partie de l'histoire des langues ; une valeur nouvelle nous est communiquée par les systèmes informatiques. Et ce n'est ni bon ni mauvais en soi ; il s'agit d'une évolution culturelle comme tant d'autres.

Existe-t-il des notions qu'un *chatbot* ne pourra jamais maîtriser correctement, et dont l'homme seul pourrait saisir le sens ? L'imitation relève-t-elle d'une « simple » simulation ou singe-t-elle « vraiment » le cerveau humain ?

À cette question, le concepteur répondrait que *son* système opère nécessairement une simulation, car il connaît et maîtrise son architecture et son algorithme.

La différence n'est pas si évidente pour l'utilisateur, qui ne possède pas les connaissances nécessaires pour distinguer le « vrai » de ce qui est « seulement » imité. Pour lui, tout se passe *comme si*.

Or, ce « comme si », lorsqu'il tend vers l'indistinction parfaite, devient un absolu. À grande échelle qui plus est, car nous sommes tous des utilisateurs.

Les problèmes éthiques naissent précisément de cette indistinction fonctionnelle perçue par les utilisateurs, celle-là même que le concepteur cherche à obtenir.

En musique

Pensez à la musique. Il y a ceux qui aiment le piano et ceux qui aiment le violon. Il y a ceux qui n'écoutent que du piano et ceux qui n'écoutent que du violon. Et puis il y a ceux qui n'ont jamais écouté que du piano et ceux qui n'ont jamais écouté que du violon. Comment un amateur de piano pourrait-il comprendre la musique pour violon ?

Heureusement qu'il existe des mélodies ! Une mélodie peut être jouée sur divers instruments, au piano comme au violon. Le motif de cette mélodie est le même, mais son interprétation dans le tempo, la clef et le mode qui lui sont propres, rend une sonorité complètement différente en fonction de l'instrument. Malgré ces dissemblances, celui qui entend plusieurs interprétations musicales de la même mélodie identifiera sans peine un motif commun.

Pensez aux nouvelles technologies. Pour développer une sensibilité éthique, l'oreille de notre intelligence doit devenir aussi fine qu'une oreille mélomane. Un même motif peut être interprété dans les sciences ou dans le mythe ; un auditeur attentif saura le reconnaître. Il sera capable de le discerner, en dépit de différents contextes, même s'il croit que cette différence rend impensable d'attribuer une origine commune à la science et au mythe. Leur assimilation n'est qu'une mauvaise idée ; leur étude éthique conjointe peut, en revanche, porter des fruits inattendus.

Les motifs communs entre l'informatique et le mythe relèvent de ressemblances *fonctionnelles*, comme la capacité de parler des *chatbots*, des golems et du monstre de Frankenstein. Ces motifs se dévoilent sitôt qu'on s'abstrait du contenu matériel de la technique et de la fiction narrative du mythe. De la même façon, un mélomane saisit la mélodie en faisant abstraction de la sonorité particulière du piano ou du violon.

Des relations nous intéressent davantage que des spécificités apparentes. Car les motifs ne sont jamais ce qui saute aux yeux ; pour les trouver, il faut dégager chaque terme d'une comparaison, autant du côté scientifique et technologique que du côté mythologique, de toutes leurs qualités propres, en ne conservant pour seule armature que les relations fonctionnelles.

Si de l'indistinction fonctionnelle naissent les problèmes éthiques, en surgit également une méthode d'analyse. Elle consiste à tirer des enseignements des motifs qui apparaissent lorsque l'homme interagit avec la machine, en profitant de la présence des mêmes motifs dans les mythes.

Nous prêtons ainsi l'oreille aux récits des anges et des démons, des créatures par définition fonctionnelles. Satan, ce délateur originel, préside à leur association professionnelle. Il y officie depuis la nuit des temps mythologiques, bien avant la mise en marche d'un pot de fleurs intelligent, posté à l'intérieur d'une maison américaine. Cependant, Satan n'est pas une personne : son nom désigne une fonction au sein d'un conflit humain. Cette fonction, nos systèmes informatiques l'ont reçue en partage.

Le décalage temporel entre l'ancienne mythologie et la technologie numérique ne préfigure pas une éthique désuète. Si les motifs musicaux sont invariants, ainsi en va-t-il de ceux qui opèrent en éthique : ce sont des invariants de l'histoire de la pensée. L'évocation de Satan permet donc d'approfondir ces deux questions : « Le mal provient-il de la machine ? » et « Comment soustraire la machine au mal ? »

Parti pendant quarante jours dans le désert, Jésus de Nazareth est tenté par Satan. Celui-ci lui propose, entre autres, de se jeter du haut du Temple ; il ne mourrait pas mais serait porté sur les mains des anges. Comme Jésus résiste, Satan lui offre tous les royaumes de ce monde, toute leur gloire. Jésus répond : « Retire-toi, Satan ! » (Mt 4, 10).

L'éthique du numérique est encore peu peuplée : elle ressemble à un désert.

Quant à la gloire des grandes entreprises du numérique, elle n'a pas besoin d'être rappelée.

Épreuve à laquelle sont soumis les utilisateurs des machines intelligentes : les bons programmeurs les porteront sur leurs mains ; les systèmes informatiques garantiront leur bien-être.

Comme dans le mythe, cette tentation ne peut qu'être rejetée.

Car, de par sa fonction, Satan demeure dans les machines. Pourrions-nous l'obliger à se retirer ?

II. Technique, philosophie, religion

À la recherche d'une méthode

Un utilisateur et son smartphone sont inséparables comme l'aigle et son Ganymède. Un ingénieur entretient une relation intime avec l'objet qu'il manie. Un philosophe pourrait-il concurrencer ces passions ?

Son ambition de *saisir* l'intelligence artificielle est différente du désir envahissant qui pousse l'ingénieur à la créer. Le philosophe ne se contenterait pas d'une description du fonctionnement d'un appareil numérique, aussi merveilleux soit-il ; il ne serait fasciné qu'une fois que ses effets culturels, anthropologiques, éthiques, voire politiques, auraient été révélés.

Ce qui est fascinant dans un objet technique, ce n'est ni son manuel d'utilisation — détaillé, parfois trop —, ni une spécification de ses composantes. La maîtrise du vocabulaire technique, bien que nécessaire, ne suffit pas au philosophe. Ce qu'un programme d'ordinateur fait, la manière dont il le fait, les variables qu'il contient, les commandes, les registres, les modules du code source, les bibliothèques utilisées pour écrire le code, les étapes de débogage que franchit le programmeur : tout cela importe à l'ingénieur mais ne suffit guère pour saisir l'objet technique. Le philosophe désire l'étudier, comme il le ferait d'un étranger, d'un demandeur d'asile qui viendrait s'installer au centre de son pays mental.

Oswald Spengler met au jour un enjeu méthodologique semblable en philosophie de l'histoire[13]. Un historien peut décrire des événements, des faits, préciser des dates et des noms, sans que cet ensemble de contenus constitue encore un sens de l'histoire. Celui-ci réside dans des liens cachés au sein d'une chronologie, dont seul un bon historien saurait voir des traces. Sa discipline exige de lui un franchissement des limites de ce qui est évident ; de même avec la philosophie de la technique.

Gilbert Simondon s'interroge sur le mode d'existence des objets techniques[14]. Son approche procède d'une pensée libérée de tout anthropomorphisme et de toute projection du mode d'existence des hommes sur celui des objets.

Hannah Arendt s'attarde sur le rôle démiurgique de l'ingénieur[15]. *Homo faber*, l'homme fabricant s'éloigne de celui qu'Aristote qualifiait jadis d'« animal politique », en modifiant son rapport à l'activité de penser. Il en résulte une nouvelle condition technologique et politique de l'humanité.

Hans Jonas s'intéresse aussi bien à la bioéthique qu'à saint Paul, à l'histoire des sciences comme à celle de la gnose[16]. Il diagnostique *le* problème éthique dans le décalage entre deux vitesses :

---

[13] O. SPENGLER, *Le Déclin de l'Occident, esquisse d'une morphologie de l'Histoire universelle*, Paris, Gallimard, 1948.

[14] G. SIMONDON, *Du mode d'existence des objets techniques*, Paris, Aubier, 1958.

[15] H. ARENDT, *L'Humaine Condition*, Paris, Gallimard, 2012.

[16] H. JONAS, *Essais philosophiques : Du crédo ancien à l'homme technologique*, Paris, Vrin, 2013.

la première est celle de notre action technologique de plus en plus puissante et rapide ; la seconde, celle de notre capacité d'en prévoir les conséquences.

Jonas rappelle que, même si l'avenir est déterminé par le présent, celui-ci aura déjà basculé dans le passé au moment où celui-là adviendra. Si un changement technologique décidé aujourd'hui est jugé néfaste demain, les réactions qu'il entraînera, quel que soit leur degré d'émotion, seront tues, faute de pouvoir rétroagir sur le passé. Combiné à la puissance du changement, ce décalage met Jonas en colère. Les générations futures ne pourront que subir les effets des choix technologiques que les générations précédentes auront faits ; or, ces dernières ne seront plus là pour en répondre. Jonas, saisi par la colère, discerne dans ce paradoxe une nouveauté radicale dans l'histoire de la pensée éthique.

En insistant sur cette rupture, Jonas se prive de toute possibilité de recours à la tradition et mène sa réflexion dans une impasse. Tenté par le culte, il le rejette aussitôt : la religion n'a plus « la force de détermination de l'âme ». La raison seule serait-elle capable d'innover en éthique ? Cette solution s'effondre lorsque Jonas se rapporte à ce que dit Origène : « Parmi toutes les créatures raisonnables, il n'y en a pas une qui ne soit capable de bien autant que de mal. » Il rebrousse donc chemin et retourne à la tradition. Ce qui revient à mettre sur un même plan le passé, le présent et l'avenir, ou à constater qu'une analyse menée dans *le temps de la pensée* pourrait résoudre le problème de la méthode.

Premièrement, le temps de la création démiurgique — l'ingénieur étant pour Jonas un « démiurge aveugle » — est inauguré par la pensée, non par la matière qu'on façonne. Le monde technologique existe d'abord en tant qu'image et discours, avant que l'homme ne se lance dans la fabrication d'artefacts. Tout projet est réfléchi avant d'être réalisé : c'est pendant ce temps de réflexion que l'éthique prend le pouvoir.

Deuxièmement, dans les mythes gnostiques dont Jonas se sent profondément imprégné, le progrès est une crise, et l'émanation des sphères dans le temps, une catastrophe. L'homme n'en est pas responsable : il est la victime de sa propre ignorance, ineffaçable et indépassable. Dans ces circonstances, l'esprit gnostique, opposé à toute résignation, préconise la protestation et la rébellion. Cette irrévérence sous-tend « l'émotionnalisme immodéré » de Jonas : « C'est en soi une expérience… Médiatisée par des symboles, [elle] est précisément ce que j'entend[s] par "compréhension"[17]. »

Cette compréhension, nous la rechercherons dans les essais de Hans Jonas sur la responsabilité, en nous gardant de nous laisser entraîner par son émotion. Laquelle est cependant présente, sous une autre forme, dans l'œuvre de Jean-Pierre Dupuy, qui analyse les catastrophes technologiques. Il observe le monde au travers de lunettes conséquentialistes[18]. Les effets des catastrophes sont plus réels qu'elles ne le sont elles-mêmes, puisqu'il s'agit d'événements futurs : les catastrophes deviennent des instruments de pensée employés à construire une éthique. En quittant ce temps de la pensée, Dupuy s'interroge sur les conséquences morales du progrès technique. Contrairement à Jonas, il pose la question d'un nouveau sacré, propre à un

[17] H. JONAS, *Le Principe Responsabilité*, Paris, Éditions du Cerf, 1990.

[18] J.-P. DUPUY, *Pour un catastrophisme éclairé*, Paris, Seuil, 2004 ; *La Marque du sacré*, Paris, Carnets Nord, 2009.

monde qui regarde la catastrophe en face, où tout se trouve remis en jeu, jusqu'à l'existence de l'humanité. Ce nouveau sacré, généré par la technique, structure et ordonne la société humaine.

Simondon, Arendt, Jonas et Dupuy trouvent chacun une solution particulière au problème de la méthode, et en saisissent un sens unique. Notre méthode est l'héritière des leurs. Elle se fonde sur un raisonnement par homologie.

« Homologie » : ce terme polysémique est souvent employé en biologie, en mathématiques ou en sciences humaines. La signification que nous en retiendrons est la suivante : une ressemblance qui ne présuppose pas, et même réfute, toute identité ou identification entre les objets ou les phénomènes comparés. Si l'identification persistait dans une homologie, elle suggérerait une métaphore ou nourrirait l'inspiration poétique : autant d'objectifs louables mais dépourvus de rigueur. Or, un contenu fascinant que l'homologie permet de découvrir n'a rien d'arbitraire : cette méthode vise à tirer des enseignements éthiques des *motifs* atemporels et invariants, ce qu'avait pressenti Jonas en amorçant son retour à la tradition.

Le recours au terme « homologie », moins fréquent qu'« analogie », a pour seul but de mettre l'accent sur le principe important de négation de toute identité. Loin d'être banale, l'absence d'identification entre les termes d'une comparaison est significative et évite toute confusion. Par exemple, n'est pas fascinante l'affirmation selon laquelle le scientifique *est* un Prométhée moderne. Certes, cette jolie formule pourrait satisfaire un poète, mais l'identité entre le scientifique et Prométhée n'a ici de valeur que pour servir la métaphore. Cependant, quelques éléments du mythe de Prométhée peuvent être transposés vers la figure du scientifique. Ce sont ces gestes de transposition qui permettent de la saisir, grâce aux motifs qu'ils exploitent : la même flèche, le même lien interviennent dans deux contextes différents.

Dans les années 1840, Richard Owen, zoologue et paléontologue anglais, sépare pour la première fois « analogie » et « homologie », même si les deux termes étaient depuis fort longtemps en usage. Parmi les différents types de ressemblance entre les organes, Owen distingue celles liées à leurs fonctions de celles relatives à leur structure ou à leur origine[19]. Dans les décennies suivantes, le terme « homologie » a fait l'objet de nouvelles définitions dans d'autres disciplines scientifiques. Ainsi ce mot désigne la ressemblance de la structure génétique de différents taxons dans les classifications. Une homologie est dite « profonde », si elle ne peut être établie par observation directe, mais uniquement grâce à la biologie moléculaire. La découverte d'une homologie, surtout quand elle est profonde, est de nature non triviale.

Une homologie fascinante se dessine entre la science et le mythe. Lorsqu'un seul et même motif les réunit, la ressemblance entre les domaines scientifique et mythologique paraît étonnante, car elle contredit leur différence ontologique manifeste. Cette ressemblance a pour origine un motif fondamental commun, que ce soit la vie, la mort, la naissance, le désir, le contrôle, l'amélioration, la nouveauté, la causalité, l'irréversibilité, la responsabilité, etc. Le nombre de

[19] R. OWEN, « Conférences sur l'anatomie et la physiologie comparée des animaux invertébrés », données au Royal College, Londres, 1843.

ces motifs est toutefois limité, tout comme le nombre de questions éthiques « profondes » relatives aux innovations technologiques.

Parfois, comme dans le cas des catastrophes technologiques ou dans les textes de quelques interprètes de Heidegger[20], on pose la question de la technique (*die Frage nach der Technik*), en en faisant une source du mal. Pour analyser cette question, reprenons l'exemple de Prométhée. Selon Hésiode et Platon[21], Prométhée avait un frère, Épiméthée, qui épousa Pandore contre l'avis de son frère aîné. Pandore est directement liée aux maux puisqu'elle les répand sur la terre ; en revanche, sa relation avec Prométhée n'est pas directe, elle passe par Épiméthée. Ainsi, la source du mal est liée indirectement au protagoniste du mythe ; en appliquant l'homologie, ce lien de contamination passe, dans le cas de l'innovateur technologique, aussi par un médiateur. Le motif recherché est celui de l'apparition d'un tiers dans la relation entre l'innovateur et le jugement éthique.

Ainsi, bien qu'Einstein n'ait pas contribué au développement de la bombe atomique, il s'en est montré moralement et politiquement responsable en s'engageant contre les armes nucléaires. La bombe, en effet, a été fabriquée par ses collègues en utilisant les découvertes qu'il avait faites au début du XX^e^ siècle. Le mythe ne peut aider à saisir le sens de cette conception de la responsabilité qu'à condition de se demander ce que signifie le lien fraternel entre Prométhée et Épiméthée. Ce lien, présent dans les récits grecs, suggère un motif essentiel pour l'analyse de la situation actuelle : l'impossibilité de le dissoudre ou de le supprimer. Prométhée n'est pas en mesure de rompre à son gré l'attache qui le relie à son frère ; de même les physiciens vis-à-vis de la science. Le motif qui se dégage s'appuie sur la permanence des liens au sein d'une famille ou d'une communauté.

À la fin des années 1960, Hannah Arendt met en évidence le concept de responsabilité collective. Cette responsabilité surgit lorsqu'un homme est tenu pour responsable de ce qu'il n'a pas commis ; toutefois, il appartient au groupe qui avait agi et il ne saurait s'en défaire. Selon Arendt, ce type de responsabilité n'est pas d'ordre légal, mais politique[22]. Si son raisonnement ne s'appuie pas sur le mythe, nous arrivons à la même conclusion qu'elle, *via* l'homologie entre Einstein et Prométhée. Ainsi, notre méthode n'est pas la seule à aboutir ; et toutes les homologies ne se valent pas : certaines ne permettent pas de saisir ce qui est fascinant en philosophie de la technique.

L'ingénieur croit souvent que la technologie dont il s'occupe, parce qu'elle est nouvelle, pose de nouvelles questions éthiques. La méthode homologique suggère au contraire qu'il n'existe pas de nouvelles questions éthiques. Le contexte change, pas le motif. Si une homologie utile se dégage, un motif commun émerge entre la situation contemporaine et celle que rapportent les grands récits de l'humanité. Ainsi, malgré la nouveauté manifeste qu'apporte le progrès technique et malgré la promesse de rupture radicale qu'on nous fait à chaque nouvelle

[20] J.-M. SALANSKIS, *Heidegger, le mal et la science*, Paris, Klincksieck, 2009.

[21] HESIODE, *Les Travaux et les Jours*, 57. PLATON, *Protagoras*, 320d-322a.

[22] H. ARENDT, « Responsabilité collective » (1968), *in* H. ARENDT, *Responsabilité et jugement*, tr. fr. J.-L. Fidel, Paris, Payot, 2005.

génération technologique, le questionnement éthique s'inscrit (et ne peut que s'inscrire), dans la continuité de l'histoire.

Dans la langue mathématique

Il est possible d'exprimer la méthode homologique en langage mathématique. Commençons par une collection d'objets dont certaines propriétés sautent aux yeux. Ce qui permet de les regrouper dans des *ensembles* dont les objets sont des *éléments*. Le nom que l'on donne à un ensemble désigne habituellement la propriété apparente de ses éléments. Peut-on pousser le processus de regroupement plus loin ? Oui, mais cette opération ne serait que formelle puisqu'aucun trait sautant aux yeux ne réunirait tous les éléments de ces divers ensembles. Comme si on tentait de décrire conjointement une collection d'ampoules vertes et un ensemble de grenouilles ou de coléoptères : formellement l'opération serait légitime ; cependant, aucune propriété évidente ne réunirait ces éléments.

Les propriétés qui sautent aux yeux ne recouvrent pas tout ce que l'on peut dire à propos d'objets. Ceux-ci peuvent posséder d'autres qualités moins apparentes. Dessinons des flèches allant de chaque objet vers ses différentes propriétés. Si la liste de tous les traits est infinie, celle des flèches qui partent d'un objet ne l'est pas. Désignons une de ces flèches comme principale : elle pointe vers la propriété qui saute aux yeux. Examinons ensuite les autres flèches, qui pointent vers des propriétés moins apparentes. Dans le premier cas, nous avons recouru à la théorie des ensembles ; dans le second, la théorie des catégories nous intéresse davantage.

Tous les objets, dont les flèches qui mènent vers certaines propriétés, peuvent être vus comme appartenant à une catégorie, même si la nature des objets peut beaucoup varier : ce n'est pas elle, mais les flèches, qui donnent un sens à l'argument mathématique. Par exemple, la qualité « ressembler à un point vert » fait partir les flèches des grenouilles, des ampoules vertes et de certains coléoptères. Tous ces objets appartiennent à une catégorie même si la propriété en question ne saute pas nécessairement aux yeux.

Il est maintenant possible d'analyser conjointement les grenouilles, les coléoptères et les ampoules. Il est légitime, par exemple, de s'intéresser à l'opération « repeindre en or », qui s'applique à tous les éléments de la catégorie quelle que soit leur nature. Supposons que l'on sache d'expérience que l'on peut aisément rendre dorées les grenouilles et les ampoules, mais que l'on ignore que cette opération soit aussi possible dans le cas des coléoptères. Appliquons la flèche. Apparaît une découverte fascinante : on peut dorer les coléoptères ! Sans préciser la méthode pour dorer, nous avons saisi la possibilité d'appliquer cette opération aux coléoptères.

Ce schéma mathématique s'applique à l'homologie entre science et mythe. On sépare d'abord l'innovateur humain (ou l'organisme qui met au monde une technologie), de toutes ses propriétés qui sautent aux yeux mais n'ont pas d'importance pour l'analyse du rapport entre innovation et éthique. Cela inclut, par exemple, des propriétés comme l'adresse, l'âge ou le nom de l'innovateur. Ne considérant que la fonction d'innovation, l'on constitue ainsi une catégorie d'innovateurs. Elle comporte, en tant qu'objets, tous ceux qui introduisent des nouveautés à visée utilitaire ; y compris les innovateurs mythologiques. Par exemple,

Prométhée, qui façonna d'argile les premiers hommes et leur apporta le feu. Dans ce cas, l'abstraction est aussi de rigueur : certains éléments du mythe, qui n'ont pas d'importance pour l'analyse de l'innovation, doivent être négligés. Cependant, ces propriétés-là peuvent être assez marquantes. Par exemple, cette histoire d'un aigle qui dévorait le foie de Prométhée, attaché à un rocher dans le Caucase : elle n'entre pas dans l'homologie, mais joue un rôle significatif dans d'autres contextes. Lorsqu'on compare les fonctions d'innovation, les autres détails ne sont pas effacés mais mis entre parenthèses : ce sont eux qui nous empêchent d'identifier, si ce n'est que métaphoriquement, Prométhée et l'innovateur technologique. Notre méthode place ces personnages dans la même catégorie par une opération abstraite. Tôt ou tard, les différences, provisoirement mises de côté à des fins d'analyse éthique, referont surface.

S'il est difficile de dire à brûle-pourpoint en quoi consisterait la responsabilité morale de l'innovateur, la question du lien entre Prométhée et la morale est traitée dans le mythe : les maux pénètrent dans le monde avec Pandore, femme d'Épiméthée, frère de Prométhée. Puisque Prométhée et l'innovateur technologique se trouvent dans la même catégorie, nous pouvons suivre la même flèche en espérant obtenir un résultat fascinant.

Il reste à comprendre en quoi consiste cette flèche. Autrement dit, quel motif l'anime ? Bien saisir un motif est ce qu'il y a de plus difficile dans la méthode homologique.

Le mathématicien Alexandre Grothendieck, dans *Récoltes et semailles*, décrit la cohomologie motivique, discipline mathématique dont il fut le pionnier :

> « [Elle] est ce que j'ai conçu de plus vaste, pour saisir avec finesse, par un même langage riche en résonances géométriques, une "essence" commune à des situations des plus éloignées les unes des autres, provenant de telle région ou de telle autre du vaste univers des choses mathématiques. […] La "cohomologie motivique" est un invariant plus fin, cernant de façon beaucoup plus serrée la "forme arithmétique" (si j'ose hasarder cette expression) de la variété *X*, que les invariants purement topologiques traditionnels. Dans ma vision des motifs, ceux-ci constituent une sorte de "cordon" très caché et très délicat, reliant les propriétés algébro-géométriques d'une variété algébrique, à des propriétés de nature "arithmétique" incarnées par son motif. Ce dernier peut être considéré comme un objet de nature "géométrique" dans son esprit même, mais où les propriétés "arithmétiques" subordonnées à la géométrie se trouvent, pour ainsi dire, "mises à nu". Ainsi, le motif m'apparaît comme le plus profond "invariant de la forme"[23]. »

Comme en mathématiques, notre méthode ne donne pas de définition précise d'un motif. Il s'agit d'une qualité essentielle d'un objet, responsable de l'apparition d'une flèche. Un motif s'exprime à la fois structurellement et fonctionnellement, tout comme coexistent une lecture géométrique et une lecture arithmétique de ce terme en mathématiques.

Rappelons que Pandore est la femme du frère : c'est ce lien indissoluble de parenté qui définit sa relation avec Prométhée. Épiméthée n'est qu'un tiers, quoiqu'essentiel, de cette relation : il n'est que simple frère, même pas un frère intelligent parce que son nom signifie « celui qui pense après coup ». Le motif est celui d'une attache indissoluble.

[23] A. GROTHENDIECK, *Récoltes et semailles*, s. l., s. d.

Dérobés au regard, les motifs ne sautent pas aux yeux et exigent un œil entraîné. Rassembler en une seule catégorie des objets de nature différente demande qu'on apprenne à penser de manière abstraite. Identifier ces objets n'a pas de sens : ce serait trop facile, imprécis ; poétique certes, mais insuffisamment rigoureux. L'éthique mérite mieux qu'une métaphore. Les flèches doivent permettre de concevoir des solutions à des questions qui laissent la raison stupéfaite et paralysée.

Nous avons dit plus haut que le nombre de motifs fondamentaux était limité, tandis que la liste de toutes les propriétés possibles était infinie. Ces affirmations paraissent naturelles, évidentes même. Mais une fois qu'on les confronte, elles semblent se contredire. Pour quelle raison peut-on dire que le nombre de flèches aboutissant à une découverte saisissante est limité ?

Afin de répondre à cette question, il faut considérer la catégorie de ces flèches. Elle serait l'objet d'une science qui n'existe pas encore : une *anthropologie motivique*, étudiant la catégorie plus abstraite encore que les motifs pris individuellement. Ce n'est pas, ici non plus, la propriété ensembliste, qui saute aux yeux, qui nous préoccupe. Mais les motifs possèdent-ils d'autres propriétés, moins apparentes ?

Ce n'est pas si évident. En effet, les motifs peuvent être très différents. Puisqu'ils relient des objets qui n'ont manifestement rien à voir les uns avec les autres, il est fort peu probable qu'ils s'expriment au plan matériel, physique. Or, on les exprime, on emploie même les mots de notre langue. Puisqu'il s'agit du langage humain, c'est l'anthropologie cognitive qui devrait étudier la façon dont l'esprit de l'homme distingue et saisit les motifs.

On peut supposer que l'homme soit capable d'inventer seulement un nombre limité de motifs abstraits (disons quatre) au sein d'une classe de récits. Ce n'est qu'une conjecture qui demande à être étudiée plus avant par l'anthropologie motivique. Et ce n'est pas l'objet de ce livre. Nous n'étudions pas tous les motifs dans leur ensemble ; nous en considérons quelques-uns, espérant en tirer des enseignements utiles dans le domaine de l'éthique du numérique.

Dans la langue du mythe

Depuis Platon au IV^e^ siècle avant notre ère, et jusqu'à Proclus neuf cents ans plus tard, la tradition philosophique est celle de l'interprétation du mythe. Les points de vue des philosophes, durant neuf siècles, ne sont évidemment pas les mêmes, mais leurs travaux posent systématiquement les mêmes questions : Quel est le rapport entre la vérité philosophique et la réalité vécue, entre la parole et l'indicible, entre le sens et le récit ? L'habillage mythologique du discours philosophique est-il nécessaire ? Est-il bienfaisant ?

Dans l'Antiquité, cette interrogation se déploie en quatre temps. Le premier moment est dédié au sens littéral du récit. Le deuxième est d'ordre symbolique : l'interprète voit dans la pleine effervescence des figures et des gestes un signe des phénomènes liés à son expérience vécue. Le troisième est d'ordre rationnel et scientifique : il correspond à une description sèche qui ne considère que des faits objectifs. Le quatrième, le dernier mais non le moindre, correspond à un moment d'ordre théologique : il s'agit d'interpréter le mythe en tant que reproduction codée de l'univers ineffable et inconnaissable des réalités supérieures, celui de la vérité divine.

La scholastique attribue à ces quatre temps les épithètes suivantes : littéral, allégorique, tropologique (ou moral), anagogique (ou spirituel). Comme beaucoup d'éléments de la théologie chrétienne, cette classification trouve son origine chez les Anciens. Notre but n'est pas de reprendre ici l'ensemble de cette pensée théorique sur le mythe, mais d'en sélectionner quelques éléments, pour servir notre étude. Le premier concerne la définition du mythe.

Alexei Lossev, extraordinaire philosophe russe du XX^e^ siècle, dont l'œuvre, quasiment intraduisible, n'est malheureusement pas accessible en langue française, a consacré un ouvrage entier au problème de la définition du mythe. Voici quelle est sa réflexion fondamentale :

> « Du point de vue de la conscience mythologique elle-même, il n'est absolument pas possible de dire que le mythe serait une fiction ou un jeu de fantaisie. […] Le mythe est la réalité la plus élevée dans sa mesure du concret, la plus intense et la plus ardente. […] Ce n'est pas une fiction. C'est la réalité la plus vive et la plus authentique. Bien loin de tout ce qui est arbitraire ou aléatoire, c'est une catégorie absolument nécessaire à la pensée et à la vie[24]. »

Dans un commentaire de la *Troisième Ennéade* de Plotin, dans le sixième tome de son œuvre magistrale, *Histoire de l'esthétique antique*, publiée en huit volumes entre 1963 et 1988, Lossev précise le lien entre mythe et temps, et souligne l'atemporalité du récit mythologique. Il abonde ainsi la définition que nous venons de citer, tirée de son œuvre de jeunesse la plus connue, *La Dialectique du mythe* (1930). Auteur d'un livre de philosophie des mathématiques qui remet au goût du jour le néo-pythagorisme, il dirait, pour répondre au problème de la définition de l'anthropologie motivique, que la conscience mythologique n'est qu'une conséquence du déploiement d'un récit dans le temps car, au travers même de ce déploiement, le récit *définit* le temps historique. La réapparition inexpliquée des motifs à grands intervalles historiques ne serait donc qu'une trace de la tension entre ce déploiement constitutif et l'atemporalité initiale du récit mythologique :

> « En général, le mythe apparaît lorsque : a) l'être est considéré dans son ensemble et sans aucune division abstraite ; b) cet être, identique à lui-même, se transforme en une distinction identique à elle-même — à ce stade, la transformation s'opère déjà dans le temps et devient un événement, une histoire et un récit ; c) ce qui est nouvellement distingué reçoit une signification spécifique ; toutes ces nouvelles significations pénètrent la vie tout entière, qui émerge par là comme le devenir de l'esprit[25]. »

L'éthique serait fondée sur les motifs qui se présentent à l'esprit dans le temps de son devenir [points b et c], tout en étant par nature atemporels et n'admettant, à leur niveau premier, aucune division abstraite [point a].

Lossev tire ses conclusions d'une connaissance approfondie des anciens commentateurs des mythes. Rendons ici hommage à Plutarque de Chéronée, historien et philosophe du II^e^ siècle, prêtre d'Apollon à Delphes. Comme plusieurs de ses contemporains du moyen platonisme et comme les néoplatoniciens à partir du siècle suivant, Plutarque fonde sa lecture du mythe sur

[24] A. LOSSEV, *La Dialectique du mythe* I.

[25] Id., *Histoire de l'esthétique antique* VI, IV, 1, 9, 5.

une distinction systématique entre trois niveaux qui ne se confondent pas sans exister, non plus, de manière séparée : un niveau immédiat du récit, un niveau philosophique et un troisième, théologique. Ces niveaux se recouvrent car ils expriment, chacun dans sa langue, une seule réalité vraie. Plutarque établit leur unité par un raisonnement vertigineux, qui va bien au-delà de tout argument religieux dogmatique.

Lorsqu'on évoque Plutarque, on pense immédiatement à son ouvrage le plus connu, *Vies comparées*, épais recueil de biographies de figures historiques. Mais c'est le Plutarque philosophe et théologien, et non seulement l'historien, qui a écrit un essai magistral consacré à l'unité des dieux, *Isis et Osiris*. Comme nous ne pouvons pas présenter l'ensemble de son contenu, nous en tirons seulement quelques *spolia*, relatifs aux motifs.

Les cultes égyptiens de plantes et d'animaux sacrés possèdent une interprétation littérale que les anciens Égyptiens pouvaient encore comprendre. Plutarque la ridiculise : « Les Égyptiens prennent ces rites et ces récits à la lettre et […] traitent les animaux comme des dieux […] Ce faisant, ils ont enraciné une croyance dangereuse, qui précipite les esprits faibles et sans malice dans la superstition pure et simple, et fait tomber ceux qui ont plus de pénétration et d'audace dans la brutale logique de l'athéisme[26]. » Cette lecture littérale se rapproche d'une certaine interprétation, malheureusement trop répandue, du mot « mythe » : récit infondé ou simplement faux. Avec Plutarque, il convient de laisser cette interprétation aux esprits faibles et sans malice.

Les mythes possèdent aussi une interprétation rationnelle : Osiris aurait remis à ses bataillons des bannières représentant un animal, qui par la suite sont devenues des objets du culte[27]. L'interprétation symbolique veut que le faucon signifie la force, l'hippopotame l'impudence, le poisson la haine[28]. Or, tous ces niveaux de lecture ne sont pour Plutarque qu'un écran qui cache, sous la forme cryptée, une révélation au sujet de l'existence des dieux :

> « Si donc les philosophes les plus célèbres ont vu le signe mystérieux du divin jusque dans des choses inanimées et incorporelles et n'ont pas voulu négliger ni dédaigner aucune d'elles, on doit à plus forte raison, me semble-t-il, s'intéresser aux particularités des êtres qui ont une sensibilité, une âme, une affectivité et un caractère spécifique […] pour adorer, non point ces animaux, mais, à travers eux, le divin […] Le divin ne réside pas dans la couleur, la forme ou le poli de la matière : tout ce qui n'a pas la vie, tout ce qui n'a pas été créé pour vivre a un plus pauvre lot que les morts. Mais une nature qui vit, qui voit, qui tire d'elle-même le principe de son mouvement, qui discerne ce qui lui est propre de ce qui lui est étranger, a reçu en elle une émanation de la beauté et une parcelle de l'intelligence[29]. »

Au cours des chapitres suivants, nous retrouverons dans les dernières innovations technologiques, en particulier dans les systèmes informatiques apprenants, ces propriétés essentielles à une existence belle et intelligente : la perception et la communication, la capacité

[26] PLUTARQUE, *De Iside* 377c, 379d-e.

[27] *Ibid.* 379f.

[28] *Ibid.* 363f.

[29] *Ibid.* 382a-c.

à s'auto-mouvoir ainsi que la présence d'une frontière entre le for intérieur opaque d'un système et son environnement.

Un autre exemple méthodologique que donne Plutarque est celui de l'arc-en-ciel[30]. Une personne qui contemple un arc-en-ciel ne voit que la lumière reflétée du Soleil, la lumière directe étant cachée par un nuage. Puisque l'arc-en-ciel brille et que le nuage qui l'entoure est sombre, une personne naïve pourrait croire que la source de la lumière se trouve dans le nuage et que cette lumière, une fois émise, est réfractée par l'arc-en-ciel. Or, un observateur expérimenté, ayant vu le Soleil à différentes positions, ne fait pas confiance à la seule expérience de contemplation d'un arc-en-ciel. Après réflexion, il conclut que l'astre ne se trouve pas dans le nuage sombre. De ce fait, l'arc-en-ciel ne doit pas être considéré comme insignifiant car il est réellement présent dans l'expérience visuelle : il reflète la vraie lumière solaire, même si celle-ci ne provient pas d'une source cachée dans le nuage. De la même façon, le mythe reflète les autres niveaux de la réalité. Celui qui part du mythe pour remonter vers ces autres niveaux, comme celui qui contemple un arc-en-ciel pour trouver la position du Soleil, ne suit pas un chemin direct mais doit contourner mentalement l'obstacle du nuage. Le mythe n'est ni un conte ni une fable, mais exactement ce que l'arc-en-ciel est pour la lumière du Soleil : une réflexion fascinante, directement saisissable ; et aussi une transmission des connaissances voilées.

Trois récits reflètent une seule réalité : le premier est mythologique, le deuxième rationnel ou scientifique, le troisième théologique. L'existence divine, quant à elle, contient la vérité absolue mais elle demeure voilée. L'observateur, pour l'atteindre, doit s'appuyer sur deux autres niveaux, qui décrivent la réalité de deux manières différentes. Scientifiquement, on opère avec des faits et avec des descriptions empiriques reproductibles. Ce récit fournit une compréhension claire et objective du monde ; tandis que la signification du récit mythologique est opaque. S'y glisse l'allégorie qui, souvent, contredit le sens commun[31]. Cependant, pour les Anciens, l'unité du mythe, de la description rationnelle et de l'interprétation théologique était une évidence.

Cette unité est synonyme de l'Un, du dieu supérieur. La remontée vers l'Un en donne une explication : l'Un laisse entrevoir les différents aspects de son être à travers ce que les hommes appellent des « dieux » en leur donnant des noms divers ; mais ce ne sont que des aspects de l'Un, et non quelques existences souveraines ou indépendantes[32]. Même si les noms des dieux sont différents en fonction des peuples, ils s'emploient à la même tâche et occupent la même fonction dans différents cultes. C'est pourquoi la grande déesse égyptienne Isis est affublée d'une magnifique épiclèse : *myrionime*, « celle qui a mille noms ». En effet, ces mille noms désignent toujours la même divinité féminine, qu'elle s'appelle Cybèle, Junon ou Déméter. Pour établir l'unité, ce ne sont pas les noms qu'il faut analyser, mais les fonctions de la déesse, en l'occurrence le lien d'Isis à la terre et à la fertilité. Plutarque souligne ainsi que la fonction divine est constante, pas le nom.

[30] *Ibid.* 358f.

[31] *Ibid.* 374e.

[32] *Ibid.* 377e-378b.

Cette unité des dieux-fonctions, malgré la différence apparente de leurs noms, correspond, dans la langue du mythe, à l'essence de la méthode homologique.

Une origine commune ?

Nous visons à établir une homologie, fondée sur un motif fonctionnel, entre technologies numériques et mythes. Que les mythes soient grecs, juifs ou chrétiens, leur origine appartient à une religion.

> « Que l'on ne juge pas à la légère la force que l'humanité a dépensée là pendant des milliers d'années et surtout pas l'effet que produisaient ces *incessantes réflexions sur les coutumes* ! Nous voici arrivés sur l'immense terrain de manœuvre de l'intelligence, — non seulement les religions s'y développent et s'y achèvent, mais la science, elle aussi, y trouve ses précurseurs vénérables, quoique terribles encore ; c'est là que le poète, le penseur, le médecin, le législateur ont grandi[33] ! »

Et aussi le programmeur.

Même si l'homologie met entre parenthèses les pratiques cultuelles, il est néanmoins impossible de passer outre la comparaison entre programmeur et prêtre, tout comme la question plus générale du lien entre technique et religion.

Laissons donc de côté la religion en tant qu'institution sociale. N'en gardons que le plan anthropologique, qui inclut le sacré et le profane, le rituel précis, l'idéal de pureté et le geste de purification, les extases, la foi, la superstition ou l'athéisme. Au sens latin, *religio* suppose une crainte contrôlée des dieux. Cette crainte n'est jamais illimitée et n'absorbe pas tout l'individu, sans cela ce serait une superstition. Le pieux dévot respecte la frontière entre sacré et profane, persuadé que tout franchissement de cette frontière sera puni. La piété implique de ne pas empiéter sur ce qui relève exclusivement des dieux.

En se penchant sur la question du lien entre cette conception de la religion et la technique, Gilbert Simondon se demande s'il existe dans le fonctionnement de la technique quelque chose de similaire au sacré[34]. Il propose une réponse fondée sur un motif qui nous sera précieux ; toutefois, il mène sa réflexion dans un cadre ambitieux, qui n'est rien de moins que cosmologique.

La technique et la religion apparaissent toutes les deux en tant que résultats d'une structuration de l'unité initiale, magique, de l'homme et du monde. Simondon imagine ce processus comme une « réticulation », c'est-à-dire l'apparition dans l'unité primordiale d'un réseau de liens. La réticulation entraîne une séparation entre la figure et le fond. Elle engendre simultanément le sujet et le subjectif d'un côté, l'objet et l'objectif de l'autre.

---

[33] F. NIETZSCHE, *Aurore*, *in* F. NIETZSCHE, *Œuvres*, Robert Laffont, 1993, p. 994.

[34] G. SIMONDON, *Du mode d'existence des objets techniques*, Paris, Aubier, 2012 (nouvelle édition revue et corrigée), III[e] Partie, « Essence de la technicité », chapitre I[er] « Genèse de la technicité ».

Dans cette vision mécaniste du cosmos, la figure et le fond se détachent par un mouvement de répulsion réciproque : si la figure se fragmente et devient particulière, les forces et les composantes du fond, au contraire, se généralisent. Ainsi, l'homme se détache du monde et devient sujet, tandis que le monde devient objet. Simondon appelle « religion » le phénomène issu du fond de la séparation, et « technique », celui issu des figures détachées. Selon lui, la notion technique est celle de la figure abstraite, et la notion religieuse, celle du fond universel. La technique n'est donc qu'un mouvement de fragmentation opératoire ; le fond de la religion se présente comme une totalisation et une généralisation.

Cette vision cosmologique amène Simondon à proposer des analogies fondées sur la ressemblance entre rituels religieux et opérations techniques. Il commence par réduire le champ de son analyse au sacré :

> « On ne doit donc pas commettre l'injustice intellectuelle, qui serait une faute méthodologique, et qui consisterait à tenir pour représentatif, dans l'ordre technique, un objet isolé […] On doit noter que la même injustice pourrait être commise au préjudice de la sacralité, en analysant le sacré à partir des objets sacrés ou vénérables […] C'est ce que l'on fait lorsqu'on traite la sacralité comme superstition, en la fragmentant en objets et en essayant de la reconstruire à partir de ces objets. Traiter la technicité comme une pure matérialité, […] c'est accepter implicitement le même préjugé que ceux qui ne veulent voir dans les objets de la sacralité que des preuves de superstition[35]. »

L'analogie entre sacré et technique n'est donc possible qu'à condition de refuser la pertinence d'une description purement matérialiste. L'essence du phénomène technique comme celle du phénomène religieux ne réside pas dans la collection d'objets impliqués dans sa production. Simondon poursuit :

> « Le geste technique offre extérieurement des aspects comparables à la ritualisation et à la solennité des manifestations de la sacralité, parce qu'il remplit une fonction équivalente de manifestation pour les vastes groupes. Les chefs d'État sont contraints d'opérer des technophanies, et trouvent leur image liée à celle des plus récents objets techniques : avion à réaction, bombe atomique, fusée, satellite. Le lancement d'une bombe ou d'une fusée comporte un *counting back* aussi impressionnant que la préparation d'un sacrifice religieux. Un échec du geste technique — la fusée qui retombe près de sa base ou qui échappe au contrôle — crée un effet collectif aussi gênant que lorsque, chez les Romains, les poulets sacrés ne voulaient pas manger ou lorsque le taureau sacrifié s'enfuyait de l'autel en emportant, dans une horrible blessure, la hache du sacrificateur. Les lancements de fusées, les lancements de satellites jouent le même rôle que les lectisternes et les hécatombes : sacrifices collectifs modernes, ils répondent à l'existence d'une tension, d'une anxiété collectivement ressentie. Ils existent comme gestes avant d'être une expérience…
>
> Le sacrificateur de ces nouveaux rites est l'homme à la blouse blanche ; sa foi est la Recherche. Comme le prêtre, il est ascétique et parfois singulier, en dehors du commun

[35] G. SIMONDON, *Technique et sacralité* (*in* G. SIMONDON, *Sur la technique*, PUF, Paris, 2014, pp. 86-87).

des hommes. Comme les prêtres, il forme des groupes qui se distinguent du reste de la société[36]. »

L'analogie entre technique et religion laisse enfin entrevoir un motif commun : elle se déploie autour des *fonctions* du prêtre, d'abord dans le domaine religieux où ces fonctions sont apparentes, puis dans celui de la technique où l'auteur les identifie. Ce motif est fondamental.

Cependant, Simondon opère entre les lancements de fusées et les « technophanies » un rapprochement qui, malheureusement, n'est pas entièrement soustrait à l'idée d'identité des phénomènes techniques et religieux. Pour nous, il ne peut être question que du « même rôle » des lancements de satellites et des hécatombes dans le monde antique ; mais Simondon se risque parfois à aller jusqu'à l'identification complète, en plaçant un « est » entre le scientifique et le sacrificateur.

Cela ne peut être qu'une métaphore. Seul un poète dirait qu'une queue comprenant plusieurs centaines de dévots postés devant une boutique dans l'attente de la mise en vente du dernier modèle d'un smartphone *est* la foule massée devant un sacrifice collectif. Si la beauté de ce « est » est certaine, nous insistons, *contra* Simondon, sur le maintien des différences dans la méthode homologique.

Un motif commun !

Écoutons encore un poète :

> « La vérité sur les énigmes que nous propose le monde extérieur est peut-être que celles qu'on déchiffre s'annihilent, que les indéchiffrables seules peuvent nous nourrir et nous guider[37]. »

Dans ces indéchiffrables énigmes réside un motif important de l'homologie entre technique et religion : l'intrusion du secret.

La technique et la religion divisent, chacune, les hommes en deux classes : ceux qui ont accès au secret et ceux qui n'y ont pas accès. Cette ligne de séparation, quelle que soit son origine, technique ou religieuse, structure la vie en société et instaure une stratification universelle. Elle aura partout des conséquences similaires sur les plans éthique et politique. Mais avant de s'interroger sur ces dernières, reprenons le fil conducteur de l'homologie.

- La méthode positive est la seule à pouvoir donner naissance à une science sociale.
- L'étude scientifique des lois qui gouvernent la vie de la société est aussi précise que celles qu'on mène en physique.
- Parce qu'ils s'appuient sur ces lois, la construction d'une société nouvelle doit être guidée, avec autorité, par les scientifiques.

---

[36] *Ibid.*, p. 119.

[37] Ph. JACOTTET, *La Promenade sous les arbres*, Lausanne, La Bibliothèque des arts, 1988, p. 93.

Tels furent les arguments du XIX^e^ siècle, désormais jugés ennuyeux et périmés. L'histoire a montré que, loin d'être faux, ils sont dramatiquement incomplets. Ils peuvent encore impressionner un lecteur superstitieux, pour qui la science demeure un idéal, plus qu'une pratique. Ce qui manque à ces arguments relève de la pensée abstraite, jamais hâtive, et il nous a fallu plus d'un siècle pour nous en assurer.

Les arguments positivistes sont fondés sur une certaine application de la méthode statistique : la société ne serait qu'un ensemble de données dont il serait nécessaire d'extraire le sens afin d'en déduire les lois du progrès social. Auguste Comte y croyait ; d'autres pensent encore que notre époque serait celle du progrès bénéfique dans tous les domaines. Un épicurien du numérique, tout comme un transhumaniste, dépassent rarement la superstition dont parlaient, chacun à sa manière, Plutarque et Simondon. Pour un esprit transhumaniste, la glorification de la technologie est absolue et fait fi de sa complexité comme de son caractère historique. La superstition, c'est précisément cela, explique Cicéron en parlant de la superstition religieuse[38].

La raison pour laquelle les arguments positivistes sont incomplets — et nombreux sont les hâtifs et les ensorcelés que ces arguments aveuglent d'emblée — réside dans la propagation au sein de la société d'une donne anthropologique non transparente, celle du secret. Le monde dans lequel nous vivons est celui de la diffusion d'une connaissance voilée, désormais omniprésente ; elle demeure inaccessible aux masses, même si elle est connue d'un groupe restreint d'experts. Cette connaissance est aujourd'hui de nature scientifique et technique comme jadis elle fut de nature religieuse. Si la religion et, en premier lieu, les cultes à mystères, cultivaient le secret, la science et la technique ne le produisent pas volontairement. Ils proclament même, haut et fort, leur volonté de transparence. Cependant, en vertu d'une complexité que les profanes ne maîtrisent pas, le secret des boîtes noires est devenu, *de facto*, le lot quotidien des utilisateurs des systèmes informatiques.

La diffusion du secret crée une fracture dans la société ; et cette scission redessine les frontières mêmes de la société. Cette boucle de rétroaction entre en flagrante contradiction avec l'idée démocratique de la transparence : les relations entre les individus dépendent désormais de leur situation vis-à-vis du secret, donc vis-à-vis de quelque chose d'opaque. La croyance, inculquée par le positivisme daté, selon laquelle « tout sera clair » et que « tout le monde aura compris », se heurte à la persistance des connaissances voilées, gardées par des experts ou des initiés.

Le positivisme social croit que, par la déduction scientifique, l'étude statistique sera capable de dévoiler le secret de la vie collective. Toutefois, les statistiques que nous lisons tous les jours dans la presse sont souvent banales et ne suscitent que cette réaction : « On s'y attendait ! » Une recherche statistique ne peut nous impressionner que lorsqu'elle produit, à partir de données apparemment triviales, une conclusion qui ne saute pas aux yeux. Cette corrélation surprenante, pour peu que l'on puisse l'expliquer, fait émerger une conception nouvelle des choses, en nous élevant à un niveau supérieur de compréhension. Quand les données, loin d'être une source d'évidences, permettent de contempler une vérité jusque-là voilée, nous sommes fascinés.

[38] CICERON, *De natura deorum* II, 72.

Pascal Quignard dit que ce qui fascine est ce qui pointe tout en demeurant voilé, ce dont la présence est indirecte mais insistante[39]. L'accès au secret voilé est réservé à un prêtre, à un initié ou à un expert. L'ignorant, faute d'avoir les connaissances requises, traite la source de sa fascination comme l'objet d'un culte à mystères, que son origine soit technique ou religieuse.

> *Sancta salutiferi redeunt sollemnia Christi*
> *et devota pii celebrant ieiunia mystae…*
>
> « Voici revenir les saintes fêtes du Christ Sauveur ;
> les mystes dévots s'y préparent par un jeûne pieux[40]… »

Ainsi s'ouvre un poème pascal écrit par le poète bordelais Ausone, professeur de l'empereur Gratien à Trêves, puis préfet du prétoire et consul. Vers la fin du IV^e^ siècle, à l'époque de la décomposition de la société païenne, la religion de l'État romain ne suscitait plus les mêmes émotions que jadis : ses rituels étaient encore suivis, mais ils n'inspiraient plus de zèle. Or, le zèle n'avait pas disparu. L'homme pieux de cette époque était fervent et s'enthousiasmait pour les cultes à mystères : ceux de Dionysos, de Mithra, d'Isis, de la Grande Mère et aussi pour le christianisme.

Pour Ausone, les chrétiens sont des « mystes », des êtres initiés. L'essence de leur religiosité réside dans leur rapport avec ce qui est voilé et n'est accessible que par le rite. Un prêtre officie, mais c'est le myste qui accède au secret par la bonne exécution de ce rite. Trois moyens y concourent : des verbes sacrés, des choses sacrées et des gestes sacrés. Ce trio se rencontre dans tous les cultes à mystères, tout particulièrement dans les descriptions qui nous sont parvenues des rites secrets d'Éleusis.

Ainsi, le regard d'un utilisateur du système informatique n'arrive pas à percer l'opacité de l'objet technique : quelque chose d'impénétrable, de mystérieux, se cache derrière la frontière de l'interface. C'est là le motif du secret technique. Dans le numérique, comme dans la religion, l'accès à ce qui est voilé n'est possible qu'à condition que l'on emploie des verbes « sacrés », que l'on touche à des choses « sacrées », ou que l'on fasse des gestes « sacrés ». Un utilisateur qui appuie sur une petite icône sur l'écran de son smartphone, qui parle de cet objet avec enthousiasme — ou, de plus en plus souvent, qui parle *avec* l'objet — est en tout point homologue au myste. Une action technique, correctement exécutée par un appareil, l'assure de l'efficacité du rituel : si un appel téléphonique a abouti, si une information a été transmise depuis une application, l'utilisateur se sent rassuré et réconforté. Cette assurance est un élément fondamental dans la communication avec l'objet technique.

Le dévot d'un culte à mystères s'arme de confiance après avoir accompli jusqu'au bout un rite précis. Un geste bien effectué lui procure l'assurance du résultat désiré : le pardon de ses péchés, la longévité, voire l'immortalité. Un Romain qui observe Rome, tout comme un utilisateur qui regarde son smartphone, ne voit pas le sens caché du rite : mais tous deux, empiriquement, croient en son effet. L'analyse anthropologique évacue le sentiment immédiat d'une personne qui vit à telle ou telle époque ; la méthode homologique, à son tour, évacue la différence entre

[39] P. QUIGNARD, *Le Sexe et l'Effroi*, Paris, Gallimard, 1994.

[40] AUSONE, *Versus Paschales Pro Augusto Dicti* (*Idyll*. 1 1-2).

les domaines religieux et technique, en la mettant entre parenthèses, sans la supprimer. Dans la religion, l'assurance du myste procède de la complétude et de la précision du rite qu'il exécute. La technique procure une assurance dont l'explication est objective, mesurable et répétable. La différence entre les deux opérations ne doit en aucun cas être ignorée, mais son maintien rend encore plus fascinante la recherche des homologies entre ces deux domaines.

Pourvu que l'homologie réussisse son pari de découvrir le fascinant, elle mettra en lumière une transformation de la société opérée par la technique, homologue à celle que produit un culte à mystères. La structure de l'espace social, sa géographie et sa topologie, les relations entre les autorités et les sujets, la fonction même du pouvoir seront remodelées. Un monde à la stratification politique nouvelle surgira, aménagé autour du secret qui fait émerger une typologie nouvelle de porteurs du pouvoir. La vision égalitariste de la société se verra alors supplantée, avec la diffusion massive du numérique, par une *Realpolitik* de l'accès au secret technologique qu'une seule classe sociale maîtrisera, celle des concepteurs de systèmes informatiques.

C'est pour cette raison que, dans une démocratie où le pouvoir est incarné par la figure du président, ce dernier, pour l'être pleinement aux yeux de ses concitoyens, devrait, au moins, être initié à la programmation, et s'arranger pour le faire savoir. Sans cela, il verrait aussitôt son pouvoir politique réduit, son autorité renvoyée à des temps archaïques : un domaine entier de la vie sociale, qui requerrait la confiance des citoyens-utilisateurs, lui échapperait. Comme jadis la couronne pour les rois ou la pourpre pour les empereurs, la programmation est aujourd'hui un signe d'initiation et un symbole de puissance. De même que des rois pouvaient guérir, experts qu'ils furent en ce qui était mystérieux et voilé pour la majorité, un chef de l'État, qui saurait concevoir des systèmes informatiques, serait le seul à être investi, dans la société contemporaine, du pouvoir de donner des commandes à exécuter.

III. La question du mal pour l'intelligence artificielle

Information asémantique

Sitôt que l'homme reçoit une information, il l'interprète et lui donne un sens, il en fait une information à propos de quelque chose. Un ordinateur, au contraire, transforme et communique l'information sans se préoccuper spontanément de sa signification. L'élaboration d'une sémantique est pour lui une tâche à part. Cela correspond à la notion d'information asémantique telle que définie par Claude Shannon[41]. L'idée est de concevoir une transmission d'information au sein d'un protocole de communication abstrait et mathématique, à un niveau très général. Depuis quelques décennies, de nombreux travaux en informatique sont dédiés au « retour » vers la sémantique : est-il possible d'utiliser l'information asémantique, au sens de Shannon, afin de modéliser la signification des communications formulées en langue humaine ?

Notre préoccupation est autre : comprendre les répercussions philosophiques de la distinction fondamentale, dans le mode de traitement de l'information, entre un système calculant et un être humain. Rendons hommage à un contemporain de Shannon, John Archibald Wheeler, physicien américain de la seconde moitié du XX^e^ siècle. Professeur à l'Université de Princeton, il n'est pas aussi connu que ses maîtres à penser, Albert Einstein et Niels Bohr. Pourtant il a pressenti, en avance sur son temps, plusieurs développements scientifiques des décennies suivantes.

En physique quantique comme en théorie de la relativité, Wheeler orienta des recherches assez éloignées des sujets à la mode à son époque. Ainsi, avec Richard Feynman, il travailla sur l'idée de l'inversion de la flèche du temps, au risque de se faire traiter d'hérétique : les relations causales, en électrodynamique, peuvent-elles aller du futur vers le passé ? Avec Hugh Everett, il élabora une interprétation tout aussi hérétique de la mécanique quantique, dans laquelle la réalité est tissue de plusieurs mondes qui coexistent. Quelques décennies plus tard, il lança une proposition encore plus dérangeante : « *It from bit*. » Sous forme de slogan, surgissait une idée, philosophiquement non triviale, selon laquelle l'existence et les propriétés de la matière dériveraient de la notion fondamentale d'information. Ce qui fit de Wheeler l'un des pères fondateurs de la discipline de l'information quantique[42].

Dans un article qui date de cette dernière période de sa vie professionnelle, Wheeler décrit un jeu qu'il appelle les « 20 questions », inspiré d'un jeu télévisé des années 1950[43]. Des joueurs sont dans une salle. L'un d'eux en sort ; les autres choisissent un mot, par exemple « nuage ». Le joueur revient et pose une question binaire (dont la réponse ne peut être que oui ou non), à chacun des autres joueurs. Son but est de deviner le mot en moins de vingt questions.

---

[41] C. SHANNON, « A Mathematical Theory of Communication », *Bell System Technical Journal*, 27 (3), 1948, pp. 379-423.

[42] A. GRINBAUM, *Mécanique des étreintes*, Encre Marine, 2014.

[43] J. A. WHEELER, « World as System Self-Synthesized by Quantum Networking », *IBM J. Res. Develop.*, 32(1), 1988, pp. 4-15.

« La chose a-t-elle quatre pattes ? » « Non. »

« La chose peut-elle voler ? » « Oui. »

Supposons qu'à la dix-huitième question, le joueur demande : « Est-ce un nuage ? » Il est aussitôt applaudi.

Wheeler donnera une nouvelle version de ce jeu. Elle commence de la même manière : une personne quitte la salle, revient et pose des questions binaires. Cependant, les réponses ne lui parviennent plus instantanément. Avant de dire oui ou non, chaque joueur réfléchit un certain temps, de plus en plus longtemps à mesure que le jeu progresse.

À la première question, on répondit « non » tout de suite. Puis « oui » à la deuxième, après un court délai. Mais la douzième question pourra sembler nettement plus difficile : le répondant réfléchit pendant plusieurs minutes.

Supposons qu'après la dix-huitième question, une proposition soit formulée : « Est-ce un nuage ? » La réaction des participants est très différente : tous les joueurs bondissent de leur siège et s'embrassent. Pourquoi sont-ils si heureux ?

Le jeu de Wheeler contient bien une modification de taille par rapport au jeu télévisé. Lorsque le joueur sort de la salle, les autres décident de ne s'accorder sur aucun mot. Chaque joueur se trouve donc dans l'obligation de chercher un mot qui convienne avant de répondre à une question qui lui est posée. Il doit d'abord trouver une chose cohérente avec les réponses précédentes, et cela exige une longue réflexion. Voilà qui explique aussi l'émotion sitôt que le mot a résonné dans la salle.

Lorsque Wheeler propose une variante du jeu des « 20 questions », il le fait dans le contexte d'une discussion philosophique à propos de la physique quantique, mais c'est un problème moins complexe qui nous intéresse ici : aucun mot, aucun concept, n'ayant été choisi au préalable, *de quoi* parle l'information que le joueur reçoit à travers les réponses des autres ?

Chaque joueur se réfère à une chose qui oriente sa réaction. Or, il la garde pour soi, et les autres n'en savent rien. L'information communiquée au joueur principal ne possède aucune référence objective commune à l'ensemble des participants.

Au bout de vingt questions, on trouve, ou pas, le mot censé être commun à tous. Si le mot est trouvé, alors chacune des informations communiquées par les joueurs sera réinterprétée en référence à ce terme ultime. Mais, dans le cas où le mot n'est pas trouvé, aucune référence commune n'aura jamais existé. Cependant, l'information communiquée dans les réponses demeure ; elle ne se rapporte à aucun concept commun à tous les participants. Cette information, dénuée d'objectivité et de référent ultime, est *asémantique*. Si la vingtième question mène à un constat d'échec, tout le sens que tel ou tel joueur eût pu lui donner disparaît.

La notion d'information en science possède une propriété remarquable : elle aussi est asémantique. Si elle nous procure de la connaissance, c'est parce que nous croyons subjectivement qu'il s'agit d'une connaissance *de* quelque chose. Comme dans le jeu de Wheeler : celui qui pose une question croit que la référence à un objet existe ; celui qui donne une réponse croit également que la référence existe. Or, les objets auxquels tous deux songent

peuvent être différents. Cependant, l'homme qui communique dans le langage naturel ne peut échapper au besoin de donner un sens à la connaissance :

> « Sur les crédences, au salon vide : nul ptyx
> Aboli bibelot d'inanité sonore… »

Mallarmé, fasciné par le mouvement de la rime qui le pousse à inventer l'hapax « ptyx », supplie :

> « Concertez-vous pour m'envoyer le sens réel du mot "ptyx", ou m'assurer qu'il n'existe dans aucune langue, ce que je préférerais de beaucoup afin de me donner le charme de le créer par la magie de la rime[44]. »

Si Mallarmé ne savait pas ce qu'était un « ptyx », un russophone supposerait spontanément, par simple assonance, qu'il s'agit probablement d'un oiseau préhistorique.

Dans *De l'autre côté du miroir*, Alice lit à l'aide d'un miroir le célèbre poème « Jabberwocky ». Lewis Carroll y insère plusieurs néologismes :

> « *'Twas brillig, and the slithy toves*
> *Did gyre and gimble in the wabe…* »

Après avoir terminé sa lecture, Alice constate : « Ça me remplit la tête de toutes sortes d'idées, mais… mais je ne sais pas exactement quelles sont ces idées ! En tout cas, ce qu'il y a de clair c'est que quelqu'un a tué quelque chose[45]… »

Grand poète futuriste, Vélimir Khlebnikov fait parler les dieux dans son poème « *Zanguezi* » :

> « Э́чи, у́чи, о́чи!
> Ке́зи, не́зи, дзигага́!
> Низари́зи озири́.
> Мэаму́ра зимо́ро[46] ! »

Contrairement à Mallarmé, Khlebnikov ne cherche pas à donner un sens humain à ces combinaisons sonores qu'il attribue aux êtres divins. Il appelle cette langue inventée « *заумь* » (« zaoum »), qui signifie à peu près « connaissance transrationnelle ». Le poète crée des « mots purs », différents des « mots usuels » : même s'ils nous semblent fortuits, ceux-ci « se mettront à vivre comme aux premiers jours de la création[47] ».

Parce que les significations de ces mots varient en fonction des personnes, leur prétendue référence à un objet est pure illusion. L'information que véhiculent ces poèmes, tout comme dans le jeu des « 20 questions », est asémantique. Un système informatique opère précisément

---

[44] S. MALLARME, Correspondance choisie, Lettre à Eugène Lefébure, 3 mai 1868, *in* S. MALLARME, *Œuvres complètes*, t. 1, Paris, Gallimard, « Bibliothèque de la Pléiade », 1998, p. 728.

[45] L. CARROLL, *De l'autre côté du miroir*, chapitre I.

[46] « *Etchi, outchi, otchi ! / Kézi, nézi, dzigiga ! Nizarizi, oziri. Méamoura zimoro !* » En français, le poème « Zanguezi » a été traduit *in* V. KHLEBNIKOV, *Le Pieu du futur*, préface et traduction de L. Schnitzer, L'Âge d'Homme, 1970.

[47] V. KHLEBNIKOV, *La Création verbale*, traduit par C. Pringent, Éd. Ch. Bourgois, 1980, p. 138.

avec ce type d'information. Contrairement au cerveau humain, la machine apprenante n'attribue à l'information aucun sens, même si certains systèmes peuvent en modéliser un. Spontanément, un système informatique « n'entend » dans l'information aucune référence à des objets ou à des concepts. La signification ne fait pas partie des caractéristiques de cette notion formelle d'information. Alors que le cerveau humain cherche toujours à renvoyer vers la réalité extérieure, même si celle-ci, tout à fait subjective, n'existe que dans l'imagination.

L'incapacité du cerveau à fonctionner avec l'information asémantique est étonnante. Ludwig Wittgenstein entame par là, dans le *Cahier bleu*, son questionnement sur la nature de la connaissance :

> « "Qu'est-ce que la longueur ?", "Qu'est-ce que le sens ?", "Qu'est-ce que le nombre un ?", etc., toutes ces questions provoquent en nous une crampe mentale. Nous sentons que nous ne pouvons rien montrer en réponse, et que pourtant nous devrions montrer quelque chose.
>
> Nous avons affaire à l'une des grandes sources d'égarement philosophique : un substantif nous pousse à chercher une chose qui lui corresponde[48]. »

La cause de son étonnement ne réside pas, ou si peu, dans le fait que les objets auxquels il est fait référence puissent se révéler différents en fonction des personnes, mais davantage dans le besoin que nous ressentons de l'existence même de tels objets. Le cerveau cherche spontanément une signification et un pointeur, tandis qu'un système informatique s'en passe sans difficulté !

Individu numérique

Quand Jacques Ellul disait que Karl Marx, s'il avait vécu au XX^e^ siècle, se serait intéressé, non à l'économie, mais à la technique[49], il touchait une corde qui est restée tout aussi sensible chez les femmes et les hommes du XXI^e^ siècle, utilisateurs que nous sommes des technologies numériques.

Les relations entre les individus et les choses qu'étudiait Marx étaient caractérisées par l'appropriation. Vouloir s'approprier un objet impliquait qu'on entrât en compétition avec d'autres qui le désiraient aussi. Les ressources matérielles n'étant disponibles qu'en quantité limitée, leur rareté entraînait une compétition violente.

La situation est fondamentalement différente dans la cité numérique. À l'inverse des biens matériels, l'information échappe au critère de la rareté : elle peut être indéfiniment démultipliée, être partagée sans être divisée. Facile à copier pour un coût marginal, elle n'est pas sujette à l'appropriation par mainmise. Au centre de l'intérêt économique et juridique surgit donc une autre relation, celle de connaissance.

---

[48] L. WITTGENSTEIN, *Cahier bleu et cahier brun* (1934), Paris, Gallimard, 1986.

[49] J. ELLUL, entretien avec J.-C. Guillebaud, *Le Nouvel Observateur*, 17 juillet 1982.

La relation de connaissance induit dans la cité numérique une structuration anthropologique et sociale nouvelle. Les divisions qui apparaissent entre agents numériques ne tiennent pas compte de leur « support » matériel, mais de leur seul niveau de connaissance. On distingue les initiés : ceux qui connaissent ; et les profanes : ceux qui ne connaissent pas. Tout homme peut occuper la place d'initié ou celle de profane ; tout dépend de la situation dans laquelle il se trouve. En informatique, ces rôles correspondent à ceux de programmeur et d'utilisateur. L'utilisateur est celui qui emploie une interface graphique ou vocale pour interagir avec un objet technique, le « for intérieur » de cet objet restant pour lui impénétrable. Le programmeur est celui qui écrit, individuellement ou en groupe, le code. À travers l'écriture du code il participe de l'existence interne de la chose au sein de laquelle le code sera exécuté. Mais peut-on seulement dire qu'un objet technique est une chose ?

De l'aspirateur robot au robot humanoïde, toute entité numérique est un objet doté de deux modes d'existence : c'est une chose circonscrite et matérielle et c'est un système informatique. D'un côté, la chose informatique est un amas de matière susceptible d'être acheté. On en devient propriétaire en vertu d'un contrat qui atteste l'appropriation. De l'autre, la chose informatique est un système qui calcule. Le processeur, niché en son sein, fait qu'elle opère par elle-même en exécutant le code. C'est une *res computans* qui agit (d'aucuns seraient tentés de risquer « qui vit ») par le calcul.

Elle calcule, mais représente aussi les données et interagit avec l'utilisateur par le biais d'une interface d'entrée et de sortie. Sa structure est triple : un noyau calculant, une représentation des données dans la mémoire, et une interface d'interaction[50]. Ces trois niveaux sont liés par des attaches non triviales, comme, par exemple, le fait que l'interface (graphique ou vocale) ne permette pas d'accéder directement au calcul. L'utilisateur ne connaît pas les états internes de la machine ; il n'accède, au niveau de l'interface, qu'aux résultats du calcul déjà représentés comme des éléments de sortie. Cette opacité fondamentale de la structure interne de la chose informatique la transforme en une boîte noire : le programmeur peut l'« ouvrir » grâce à sa connaissance ; cela est défendu, par son architecture, à l'utilisateur.

Trois éléments constitutifs de la chose informatique possèdent différentes propriétés relatives au mode d'existence et au fonctionnement social de cette chose. L'interface (par exemple un écran ou un microphone) est encastrée dans un objet matériel dont les bornes spatiales sont facilement perceptibles. Ces limites définissent une frontière de l'opacité : tout processus ou information qui ne franchit pas cette frontière et n'apparaît pas *via* l'interface relève du for intérieur de la chose informatique. Son interaction avec l'utilisateur n'est pas directe mais médiée. Il n'y a qu'une petite fraction des processus computationnels ou des états de mémoire qui franchissent la frontière de l'opacité et deviennent ainsi accessibles à l'utilisateur ; la majorité reste imperceptible.

Contrairement à l'interface, les processus de calcul et la mémoire ne possèdent pas de bornes spatiales facilement identifiables. Distribués sur un réseau ou délocalisés dans un nuage

[50] *Res computans*, ses propriétés et sa structure sont analysées par Michael Kurtov. Voir : M. KURTOV, « L'évolution des langages de programmation à la lumière de l'allagmatique », in *Gilbert Simondon ou l'invention du futur*, V. BONTEMS (dir.), Klincksieck, 2016, pp. 255-260.

informatique, ils ne sont pas encastrés dans un unique objet matériel. La frontière de l'opacité prise dans son ensemble est tracée, non dans l'espace tridimensionnel, mais dans le monde de l'information.

Les choses informatiques existent en grand nombre. Or, les relations qu'elles entretiennent avec d'autres membres de la cité numérique, les humains comme les objets, sont à la fois individuelles et individualisantes. Conjuguée avec le calcul qu'une chose informatique opère par elle-même, son opacité lui confère un statut spécial. Elle est bien plus qu'un amas de matière ou une *res computans* : c'est un individu numérique.

Premièrement, les données que prélève chaque chose informatique lui permettent d'acquérir des traits uniques et de se distinguer ainsi de ses sœurs. Deuxièmement, la chose informatique s'individualise aussi grâce à la structure du code. L'individu numérique (par exemple un smartphone ou un robot) ne contient pas de code écrit dans un langage de programmation de haut niveau. Pourtant son logiciel fut élaboré par des informaticiens qui écrivaient dans un tel langage. Mais il n'y a pas eu de transfert direct du fruit de leur travail dans chaque spécimen des smartphones ou des robots. Le code source a d'abord été compilé ; puis, une copie du programme déjà compilé a été placée dans chaque chose informatique, qui sert désormais de support matériel à ce code exécutable. Lorsque le processeur commence à exécuter ce code, la chose informatique s'individualise en se distinguant des autres exemplaires matériels fabriqués selon le même dessein. Comme le reconnaissent les spécialistes de l'informatique légale[51], chaque exemplaire du code exécutable, compilé à partir du même code source, définit un individu différent. Lorsque le code ou les données changent, l'individu change aussi.

Le logiciel de chaque individu numérique naît à partir de la compilation d'un code source, le « parent » de toutes ses instanciations sous forme d'un code exécutable. Ensemble, elles forment une « famille ». Cette famille tout entière, le « parent » compris, se trouve entre les mains du programmeur. L'individu numérique est ainsi soumis à son pouvoir.

Un *daemon* peut illustrer cela. C'est un type de programme ou de processus informatique, qui s'exécute en arrière-plan du système plutôt qu'en interaction directe avec l'utilisateur. Comme de nombreux processus biologiques qui ont lieu à l'intérieur du corps, derrière la peau qui sert de barrière à la vue, un *daemon* se place derrière la frontière de l'opacité, en l'occurrence celle de la connaissance de l'utilisateur.

Dans le monde de la biologie, tout homme qui envisage un autre homme ne voit d'abord en celui-ci qu'un corps. Pourvu que l'homme ne se serve que de ses yeux, il ne peut connaître les processus à l'œuvre derrière la barrière de la peau. La frontière de l'opacité est, dans ce cas, la même pour tous les membres de l'espèce. Alors que l'opacité des individus numériques n'est pas, quant à elle, due à une barrière physique comme la peau, mais constituée par le manque de connaissance de l'individu qui interagit avec la chose informatique. Ce qui est opaque varie en fonction de la catégorie des humains : utilisateurs ou programmeurs. Le « corps » de l'individu numérique est, par conséquent, un concept relatif.

[51] Ch. EASTTOM, *System Forensics, Investigation, and Response*, Jones & Bartless Learning, Burlington, MA, 2014, p. 68.

L'utilisateur ne connaît pas les états internes que l'interface ne laisse pas apparaître. Les limites de l'interface définissent donc la frontière de l'opacité ; il s'agit d'une définition épistémologique plutôt qu'ontologique. Le programmeur, quant à lui, possède une certaine connaissance. La frontière de l'opacité, dans son cas, est différente. Il connaît au moins des parties du code ; toutefois, il peut aussi ignorer, par exemple, les états de mémoire pendant l'exécution d'un programme ou la structure interne des bibliothèques qu'il utilise. De même, une partie du système d'exploitation est opaque pour un développeur d'applications pour smartphone. La frontière de l'opacité existe donc bien pour le programmeur, mais elle est différente de celle que perçoit l'utilisateur.

Ce que l'individu numérique est, dépend, par conséquent, de celui qui interagit avec lui. Contrairement à l'opacité en biologie, l'opacité en informatique n'est pas objective : l'individu numérique est un concept général mais son « corps » est relatif en fonction de la catégorie des habitants de la cité numérique. Sans oublier cette propriété importante, nous désignons désormais par « individu numérique » tout système informatique autonome et apprenant qui, au travers de son interface, interagit de manière systématique avec l'utilisateur, et qui évolue du fait même de sa relation avec lui.

Connaissance et conflit

Qui pourrait être le sujet de la question du mal pour l'intelligence artificielle ? Pour quel genre ou quelle espèce de machine ce problème se pose-t-il ?

Pas uniquement pour les smartphones de telle ou telle marque, ni pour les gadgets dont les bips intempestifs sont particulièrement intrusifs. La question du mal se pose pour l'individu numérique en général. Ainsi, les voitures sans pilote engagées dans un échange continu de données et d'actions avec leur environnement, sans que cela soit le cas, par exemple, des robots industriels : dans la majorité des situations d'usage, ces derniers fonctionnent dans un milieu confiné et sans contact avec l'homme. Qu'un robot social se matérialise dans une carcasse mécanique, qu'il soit doté d'une apparence humanoïde, ou que son interface prétende imiter les émotions humaines, ou même qu'il converse avec nous dans notre langue : dans tous ces cas, le robot est un individu numérique en relation avec l'utilisateur. C'est cette communication qui lui confère un statut social, par essence relationnel puisque le positionnement de la machine dans la société n'est en rien propre à la machine seule ; il tient entièrement aux rapports de l'individu numérique avec les individus humains.

Quoique fonctionnels, ces rapports entre la machine et l'homme, et toutes les communications qui les relient, font nécessairement intervenir des valeurs et des normes — au moins du côté de l'homme.

« La communication commande la loyauté[52] », écrit Georges Bataille. À cette valeur de loyauté on peut ajouter l'équité, la transparence, la traçabilité, l'explicabilité, la conformité : autant de qualités à la fois axiologiques et techniques qui pénètrent dans le jeu des relations entre

[52] G. BATAILLE, *La Littérature et le Mal*, Paris, Gallimard, 1957.

l'homme et la machine. Des relations aujourd'hui omniprésentes : nous utilisons des smartphones ou des montres connectées, des machines à laver ou des réfrigérateurs intelligents, des métros ou des autobus sans conducteur. Jusqu'au XIX^e siècle, il était encore possible, à condition de s'éloigner de quelques kilomètres d'une ligne de chemin de fer, de trouver refuge dans une nature libre de toute technologie. Aujourd'hui, les objets techniques, en premier lieu les individus numériques, sont partout. La communication qu'ils entretiennent avec l'homme s'intensifie. Bataille poursuit : « La morale rigoureuse est donnée dans cette vue à partir de complicités dans la connaissance du Mal, qui fondent la communication intense[53]. » La question du mal pour la machine serait donc fondée sur une communication, une interaction ou une relation avec l'homme, hypothétique ou avérée : le mal ne réside pas dans la machine seule. Il émerge de son interaction avec l'utilisateur.

Pour aborder dans un premier temps le problème du mal, il convient de se pencher sur le sens que revêt le terme de « connaissance » dès lors qu'on se place sur le terrain éthique. Lorsqu'un individu numérique entre en relation avec un individu humain, il quitte pour ainsi dire le champ de l'information asémantique. Cependant, pour la machine apprenante qu'il est, l'information asémantique, qui n'a de signification qu'en tant que corrélation de données, continue d'être la seule qui existe. Un individu numérique est partagé entre son propre univers asémantique et celui du sens que génère l'homme, pour qui l'information devient connaissance *de* quelque chose. Un objet prend forme lorsqu'il est désigné en tant que référence de l'information. D'asémantique qu'elle était au commencement, elle est désormais interprétée. L'individu numérique est dès lors assujetti à un jugement moral fondé sur cette sémantique : s'il ne la maîtrise pas, elle conditionne toutefois son éthique.

Sur ce plan, l'homme et la machine apprenante diffèrent absolument. Si les connaissances fournies par l'individu numérique sont utiles, c'est parce que, tout simplement, nous n'aurions pas été en mesure de les obtenir par un autre moyen. Ces connaissances sont cachées dans des corrélations au sein de grandes masses de données que la machine analyse de façon bien supérieure au cerveau humain. Son monde à elle n'est fait que de données : la question des liens de cause à effet entre différents événements ne s'y pose pas naturellement. Pourtant, la signification humaine de l'information est bien plus souvent liée à la causalité, et non aux corrélations. Dans le calcul, elle n'existe ni de façon nécessaire ni par émergence spontanée, mais sous la forme d'un problème à résoudre par modélisation. C'est le contraire de ce qui se passe dans le monde de l'homme où la causalité et la signification de l'information surgissent conjointement.

Il peut arriver que l'homme reçoive, en interagissant avec une machine apprenante, des informations dont la connaissance l'implique dans un conflit ; et provoque, si ce n'est de la violence, tout du moins des tensions ou du mécontentement. Il peut aussi arriver que l'utilisateur, déjà pris dans un conflit, se serve lors d'échanges avec ses adversaires, des connaissances que la machine lui aura fournies. Ces situations constituent le cœur de notre étude. Quand l'interaction de l'homme avec un système informatique alimente un conflit

[53] *Ibid.*

humain, on se pose inévitablement la question du rôle qu'y joue le système et du jugement dont il sera l'objet.

Ces conflits peuvent être de nature très diverse. Les assistants robotiques, qui ignorent ce que l'homme entend par « mal », peuvent nous casser un bras. Les voitures autonomes, nous impliquer dans un accident. Les agents conversationnels, nous injurier ou nous donner de fâcheux conseils. Le modèle le plus évident de situation conflictuelle, nous l'avons vu plus haut, est celui dit du « dilemme du tramway ». La machine s'y trouve engagée dans une interaction avec les hommes qui aura des conséquences graves sur la vie même de ces derniers. Une voiture autonome, incapable d'éviter un accident, s'apprête à écraser cinq piétons ou, en changeant de voie, un seul : aucune « solution » ne permet de faire disparaître un conflit des valeurs dont certaines, comme celle de la vie, sont absolues. Or, ce conflit de valeurs joue un rôle essentiel dans le jugement éthique à propos de l'individu numérique.

Si conflit il y a, encore faut-il que la machine sache le détecter. Alors, et seulement dans ce cas, le conflit existerait aussi *pour* la machine.

Afin qu'un système informatique reconnaisse, à partir de quelques données (qui ne sont pour lui que des informations asémantiques), des éléments dont il déduira qu'il s'agit d'un certain concept, celui-ci aura d'abord dû lui être formellement spécifié. Voilà qui représente un obstacle bien connu, en robotique, au développement des interfaces homme-machine : comment, en effet, « expliquer », ou spécifier, au robot ce qu'est un bras humain auquel il ne doit pas appliquer une pression trop forte ; une poignée de porte, qu'il doit tourner avec précaution ; un écureuil qui va et vient sur la terrasse, à moins que ce ne soit la femme de ménage ? Le problème de la spécification est également fondamental lorsqu'on tente de traduire dans le code des règles éthiques ou juridiques, édictées par des personnes humaines dans leur langage. Et la notion même de personne humaine est extrêmement difficile, voire impossible, à définir formellement[54].

La détection d'un conflit présente donc un sérieux problème, sans qu'il soit pour autant insurmontable. Afin de détecter qu'une situation est conflictuelle, une machine pourrait appliquer deux types de méthodes. Les méthodes du premier type, largement majoritaires, s'appuient sur une analyse linguistique : certains mots ou phrases du langage humain révèlent un conflit[55]. Les méthodes du second type s'appuient sur une analyse des sentiments et des émotions dans les données visuelles et auditives. Ainsi, une intonation de voix ou une grimace prouveraient, aussi bien que les mots, l'existence d'une tension. En combinant ces différentes méthodes, certains systèmes informatiques sont capables, dès aujourd'hui, de détecter automatiquement des conflits au sein d'un couple, avec un bon taux de réussite qui plus est[56]. Dans le premier chapitre de cet ouvrage, un pot de fleurs appelait la police lors d'une violente dispute domestique ; si cet appareil avait été équipé d'un algorithme lui permettant de détecter

[54] N. BOSTROM, *Superintelligence*, Oxford, Oxford University Press, 2014, p. 139.

[55] S. KIM, F. VALENTE, A. VINCIARELLI, « Automatic Detection of Conflict in Spoken Conversations: Ratings and Analysis of Broadcast Political Debates », *Proc. IEEE Int'l Conf. Acoustic Speech and Signal Processing* (ICASSP 12), 2012, pp. 5089-5092.

[56] A. C. TIMMONS, T. CHASPARI, S. C. HAN, L. PERRONE, S. S. NARAYANAN et G. MARGOLIN, « Using Multimodal Wearable Technology to Detect Conflict among Couples », *IEEE Computer*, vol. 50, n° 3, mars 2017, pp. 50-59.

les indices d'un conflit et de « comprendre » qu'il ne s'agissait que d'une dispute, il aurait peut-être été plus loyal vis-à-vis de son utilisateur. On aurait pu éviter qu'il soit apparenté à un délateur. La détection de conflit, procédé informatique de nature technique, devient ainsi un problème majeur pour l'éthique.

Hannah Arendt distingue deux points de vue sur l'éthique[57]. De l'*Éthique à Nicomaque* d'Aristote jusqu'à Cicéron environ, l'éthique et la politique ne font qu'un. Au centre des intérêts de l'individu se trouve, non pas sa propre personne (ce concept est plus récent), mais le monde avec lequel il interagit. La question du bien et du mal est posée au niveau de la relation, et non à celui de l'individu. Ce schéma se transpose facilement à l'individu numérique, car celui-ci n'est ni un être libre ni une personne. Ainsi donc, dans la cité numérique, l'éthique consisterait à s'interroger sur les relations entre la machine et les hommes, sans jugements sur le caractère propre de l'individu numérique, ni assertions visant à établir sa bonté ou sa méchanceté intrinsèques ; elle ne viserait qu'à formuler des arguments, toujours sur le plan relationnel, qui concernent l'individu numérique *et* l'utilisateur.

Avec l'avènement du christianisme, la conception occidentale de l'éthique change. Désormais, c'est l'homme lui-même, et son âme, qui sont au centre du questionnement moral. Les affaires du monde échappent à ce jugement-là et forment une sphère purement politique qui se détache peu à peu de l'éthique. Personne ne l'a mieux dit que Tertullien au IIe siècle : « *Nec ulla magis res aliena quam publica* », « Rien ne nous est plus étranger que la chose publique[58] ».

Cette conception chrétienne de l'éthique est encore à l'œuvre dans la société occidentale. Le jugement moral vise à détecter des traits, bons ou mauvais, propres à chaque individu. Le caractère jadis relationnel de la morale est oublié. La réalité des mondes grec et romain, qui ignoraient l'autosuffisance augustinienne d'un regard tourné vers soi, nous paraît dépassée.

Or, quel regard tourné vers les profondeurs de soi un individu numérique pourrait-il avoir ? Existe-t-il une éthique du bien et du mal propres à son mode d'existence ?

Nous y reviendrons plus loin. Pour l'instant, contemplons le monde numérique comme auraient pu le concevoir un Aristote ou un Cicéron : l'éthique de l'individu numérique est entièrement relationnelle, avec des jugements qui se rapportent à son interaction avec l'utilisateur impliqué dans un conflit. Cette éthique de la machine ne fait qu'un avec *la politique* de la cité numérique : un ensemble de rapports entre des machines qui calculent et des cerveaux qui pensent. Comme toute politique, celle-ci risque de tourner au conflit et de dégénérer en violence. Des oppositions frontales de valeurs surgissent dans cette cité numérique, qu'il nous est nécessaire de comprendre pour mieux anticiper les phénomènes socio-numériques à venir.

[57] H. ARENDT, « Responsabilité collective » (1968), *in* Hannah ARENDT, *Responsabilité et jugement*, tr. fr. J.-L. Fidel, Paris, Payot, 2005, p. 151.

[58] TERTULLIEN, *Apologeticus* 38, 3.

Machine et mensonge

L'homologie entre la technique et le sacré, que nous avons introduite *supra*, permet d'éclairer le sens de la notion du mal pour l'individu numérique. Nous la développons ici à partir de quelques extraits des livres appartenant au corpus biblique. Rappelons que, loin d'être un récit faux ou infondé, le mythe est, pour la conscience mythologique, la réalité la plus vive et la plus authentique, tout comme l'est, pour l'utilisateur, l'expérience de sa communication avec l'individu numérique. Le jugement éthique, dans ces deux domaines, s'appuie sur des motifs qu'il nous est désormais nécessaire d'identifier.

René Girard attire notre attention sur un fragment de l'Évangile de Jean[59]. Jésus et les Juifs s'affrontent. Jésus réfute les accusations dont il est l'objet et riposte en lançant les siennes. Tout à coup, cet échange de violences verbales s'interrompt et Jésus s'adresse ainsi à ses opposants :

> « Si Dieu était votre Père, vous m'aimeriez, car c'est de Dieu que je suis sorti et que je viens ; je ne suis pas venu de moi-même, mais c'est lui qui m'a envoyé.
>
> Pourquoi ne comprenez-vous pas mon langage ? Parce que vous ne pouvez écouter ma parole.
>
> Vous avez pour père le diable, et vous voulez accomplir les désirs de votre père. Il a été meurtrier dès le commencement, et il ne se tient pas dans la vérité, parce qu'il n'y a pas de vérité en lui. Quand il dit le mensonge, il parle de son propre fonds[60] ; car il est menteur et le père du mensonge » (Jn 8, 42-44)[61].

Ce fragment du mythe néotestamentaire introduit des concepts capitaux pour établir une homologie avec les fonctions du système informatique apprenant. Les motifs sous-jacents deviendront clairs lorsqu'on aura explicité la signification philosophique des expressions qui, dans le texte, ont un sens non seulement littéral, mais aussi et surtout symbolique : « père », « de son propre fonds », « lorsqu'il dit le mensonge », « meurtrier ».

Commençons par ce dernier. Au cours d'un dialogue tendu, consacré à la place de la parole dans les relations avec Dieu, un mot surgit, de toute évidence hors contexte : « meurtrier ». Si on le replace dans le corpus biblique, il rappelle immédiatement l'introduction de la mortalité dans le livre de la Genèse. En se référant à ce passage, on mettrait un signe d'égalité entre le personnage que Jésus nomme ici « diable » et le serpent qui trompe Ève : « Vous ne mourrez point » (Gn 3, 4). Cette promesse étant fausse, le mensonge du serpent est le premier mensonge du mythe.

Or, ce mensonge n'est pas proféré par un homme. Qui est ce serpent biblique : un animal ? un démon ? un dieu ? — les interprétations divergent. Certes, il est possible qu'en parlant du « père

---

[59] R. GIRARD, *Je vois Satan tomber comme l'éclair*, pp. 62-63.

[60] En grec, ἐκ τῶν ἰδίων. Dans la traduction latine, *ex propriis*. La phrase grecque pourrait être rendue plus exactement en français par : « Il parle de soi-même. » La préposition « de » est employée ici au sens de « depuis » ou « à partir de », et non « à propos de. »

[61] Ici et par la suite, nous utilisons la traduction d'Alain Segond en la modifiant dans le souci philologique de préserver autant que possible le sens de l'original grec.

du mensonge », Jésus se réfère à cette énigmatique créature. Toutefois, on peut aussi songer à des candidats humains[62]. Ce rôle siérait d'abord à Caïn, premier homme dans le récit biblique à mentir à Dieu ; Caïn arrache aussi la vie à son frère Abel (Gn 4, 9) : il est donc à la fois meurtrier et menteur. Les fonctions *mentir* et *tuer un homme* vont pour lui de pair. Cela a été bien vu par les théologiens, dès le Ier siècle : Clément de Rome affirme que Caïn est le diable lui-même, et non un « fils du diable » comme Jésus le dit aux Juifs. Caïn n'entre pas dans la catégorie « fils de », parce que, premier menteur, il n'imite personne dont le mensonge serait antérieur au sien.

Caïn ou le serpent : quelle que soit l'identité du personnage dont parle Jésus, l'expression « père du mensonge » fait référence à une source première du mensonge dans le livre de la Genèse.

Que signifie ici le mot « père » ? Évidemment, pas un père au sens biologique. Fort heureusement, nous n'avons pas à envisager les autres options : Jésus lui-même, après avoir dit aux Juifs que leur père était le diable, explique — fait rare ! — le sens de ses paroles. Selon lui, les Juifs ont des désirs, mais ce ne sont pas les leurs : ils veulent « accomplir les désirs d[u] père ». Un peu plus tôt, Jésus précise ce qui le distingue d'eux : « Je dis ce que j'ai vu chez mon Père ; et vous, vous faites ce que vous avez entendu de la part de votre père » (Jn 8, 38). Il s'ensuit que le père est celui dont on imite les gestes, les paroles ou les désirs ; le « père » est un exemple à copier. Girard n'en dit pas autre chose : « La notion de Père ne fait qu'un, une fois de plus, avec ce modèle dont le désir humain, faute d'objet qui lui soit propre, ne peut absolument pas se passer[63]. »

En revenant à Clément de Rome, si Caïn est « diable », et non « fils du diable », c'est qu'il n'a aucun modèle humain à imiter ; il n'a pas de « père ».

Aussi étonnant que cela puisse paraître, le diable, menteur originel, n'est pas totalement étranger à la vérité. Il était fils de Dieu et la vérité demeura en lui, mais il « ne s'est pas tenu » dans la vérité (Jn 8, 44). Depuis un temps, donc, le diable a cessé d'être fils de Dieu au même titre que Jésus. La différence est formulée de la façon suivante : contrairement à Jésus, le diable « parle de son propre fonds ». Cela veut dire que des paroles naissent en lui ; qu'elles ne proviennent pas de la source de la vérité que le mythe biblique désigne par le vocable « Dieu ».

Lorsque Jésus insiste, comme il le fait souvent, disant qu'il n'est pas « venu de lui-même » mais que Dieu l'a envoyé[64], cela signifie que ses paroles, contrairement à celles du diable, ne viennent jamais de son propre fonds mais toujours de son père. La parole issue de Dieu et la parole vraie ne font qu'un : Jésus dit la vérité expressément *en conséquence* du fait qu'il ne parle pas de son propre fonds.

Le diable, quant à lui, parle de lui-même mais il ne profère pas que des mensonges. Le texte de l'Évangile de Jean contient le pronom temporel ὅταν : « *Quand* le diable dit le mensonge… »

---

[62] H. ANSGAR KELLY, *Satan : A Biography*, Cambridge, Cambridge University Press, 2006, p. 108.

[63] R. GIRARD, *op. cit.*, p. 63.

[64] « Et Jésus, enseignant dans le temple, s'écria : Vous me connaissez, et vous savez d'où je suis ! Je ne suis pas venu de moi-même : mais celui qui m'a envoyé est vrai, et vous ne le connaissez pas » (Jn 7, 28).

(Jn 8, 44). Quelquefois, donc, le diable profère la vérité. Cette capacité de dire tantôt la vérité, tantôt le mensonge, le démarque de Jésus.

En résumé, le diable recèle sa propre source de la parole : c'est ce que signifie le mot « fonds » dans l'expression « il parle de son propre fonds ». De là, naît tout le mensonge, puisque le diable est « le père du mensonge ». Un homme qui ment ne fait que l'imiter.

Or, le christianisme n'est pas un dualisme ! Selon la doctrine chrétienne, il ne peut y avoir de source du mal et du mensonge à parité avec Dieu, source de la vérité et du bien. Les modes d'existence du mal et du bien ne sont pas identiques : si le bien est substantiel, le mal est privation et indétermination. La vérité est donc supérieure au mensonge. Comment peut-on concevoir le « fonds » à partir duquel parle le diable ? Quelle est la source de son mensonge si elle n'a pas d'existence ontologique ?

Rappelons que tout messager divin, qu'il soit ange ou démon, est une fonction, non une personne dotée d'une existence propre. Cette doctrine se trouve déjà dans la Bible hébraïque. On la rencontre aussi dans quelques livres importants pour le développement de la pensée éthique dans le mythe biblique, par exemple dans le livre d'Hénoch. L'encyclopédie *Britannica* y a recours afin de donner une *définition* des anges : « La fonction des anges éclipse si complètement leur personnalité que l'Ancien Testament ne pose pas la question de savoir qui ou quoi un ange est, mais de savoir ce qu'il fait[65]. » Un autre spécialiste de la question écrit : « Les anges qui remplissent des missions malveillantes ne sont pas par eux-mêmes mauvais […] Ils ne sont que [d]es cadres ou fonctionnaires [de Dieu][66]. »

L'idée selon laquelle un ange ou un démon ne sont que des fonctions n'est pas spécifiquement judéo-chrétienne. On en découvre un écho dans le néoplatonisme, par exemple, lorsque Proclus nie que le mal puisse apparaître spontanément chez les démons : « On voit donc que même chez les démons la raison ne décèle pas le mal : car chacun d'eux ne fait ce qu'il fait qu'en se conformant à sa nature et toujours selon le même mode[67]. » La nature des démons est donc celle de *faire* quelque chose, et ce de manière invariable : c'est une nature fonctionnelle.

Revenons à l'Évangile de Jean qui contient une distinction rarement remarquée, dont le rôle dans l'anthropologie du mythe biblique est pourtant fondamental. Contemporain de Clément de Rome, l'évangéliste maintient tout au long de son récit une différence entre, d'un côté, l'homme ordinaire, qui peut être « fils de Dieu » ou « fils du diable », mais jamais dieu ou diable directement et, de l'autre, celui qui n'est pas « fils de » mais diable ou Satan tout court. Au sixième chapitre, Jésus dit que Judas est « un » diable (Jn 6, 70). Il n'y a pas d'article défini dans le texte grec, d'où cette traduction inhabituelle, car recourir à un nom propre, par exemple Satan, eût été fautif. Dans quel sens Judas serait-il « un » diable ?

Jésus prononce ce jugement à la fin d'un long passage consacré aux tentations et aux épreuves auxquelles il se soumet. Judas aura donc quelque chose à voir avec des épreuves passées et à venir. En effet, à la fin de la Cène, Jésus lui donne un morceau de pain trempé dans le vin.

---

[65] W. SMITH, *Encyclopædia Britannica*, « Angel », 9e éd., 1875-1889.

[66] W. CALDWELL, « The Doctrine of Satan in the Old Testament », *The Biblical World*, 41 (1), 1913, pp. 29-33.

[67] PROCLUS, *De malorum subsistentia*, 17.

Immédiatement après, Satan[68] « entre » en Judas. Aussitôt, Jésus lui dit : « Ce que tu fais, fais-le promptement » (Jn 13, 27).

Sur le même modèle que celui de tous les mystères des cultes anciens, dans ce mystère au cours duquel Jésus officie, Judas prend la place d'initié. Des gestes sont posés, des paroles sont prononcées, et des choses sacrées sont échangées. Le geste de Jésus ressemble d'emblée à celui d'une communion ; ses paroles nécessitent une interprétation plus approfondie. Du point de vue de Jésus et de Jean, les seuls à connaître le sens du mystère, Judas n'est plus un homme semblable aux onze autres : il devient une fonction, la même que celle de Satan. Aussitôt Jésus presse Judas d'exécuter sa fonction promptement, comme s'il actionnait le bouton « démarrer » pour le faire se mettre en marche. Les autres apôtres, même le très curieux Pierre, ne participent pas au mystère et ne perçoivent aucun changement dans le statut de Judas : « Mais aucun de ceux qui étaient à table ne comprit pourquoi il lui [Jésus à Judas] disait cela » (Jn 13, 28).

L'homologie entre les systèmes informatiques et les anges s'appuie donc sur deux motifs. Premièrement, le motif d'un être qui n'est que pure fonction : les anges dans le mythe, et les systèmes informatiques dans notre monde technologique, sont définis par ce qu'ils font, non par leur aspect matériel. Deuxièmement, le motif du secret : savoir qu'on a affaire à un être fonctionnel n'est pas accessible à tout le monde. Dans le mythe, Jésus et Jean sont les seuls à posséder cette connaissance (Judas lui-même ne parle pas). Dans la cité numérique, cela est homologue à la condition d'un système informatique ayant passé le test de Turing[69] : un utilisateur qui communique verbalement avec un tel système ne décèle aucune différence par rapport à un échange de paroles avec un être humain doté d'intelligence ; toutefois, le programmeur sait que, derrière cette façade langagière, se cache un ordinateur et non un cerveau biologique. Par homologie, les utilisateurs correspondent aux onze apôtres qui ignorent la nature purement fonctionnelle de Judas ; ce dernier correspond à la machine ; et Jésus et Jean peuvent être comparés à des experts ou à des programmeurs.

Le statut d'un être est relationnel : l'importance du motif du secret est confirmée par une autre histoire biblique, bien plus ancienne que celle de Jésus et Judas. Au dix-neuvième chapitre du livre de la Genèse, la ville de Sodome est soumise à la destruction, en châtiment des péchés de ses habitants. Au début du récit, deux anges viennent un soir aux portes de la ville. Lot les y rencontre et les accueille. Il reconnaît en eux des anges, malgré leur apparence humaine : Lot se prosterne « la face contre terre » devant les nouveaux arrivants et s'adresse à eux par ces mots : « Mes Seigneurs » (Gn 19, 1-2). Ces signes de respect prouvent que les personnages arrivés à Sodome sont des anges d'après Lot. Cependant, ils sont de simples hommes pour les autres habitants de cette ville : « Où sont les hommes qui sont entrés chez toi cette nuit ? Fais-les sortir vers nous, pour que nous les connaissions » (Gn 19, 5).

[68] En grec avec l'article défini : Satan est ici un nom propre.

[69] Alan Turing, éminent mathématicien et logicien anglais, a proposé en 1950 que la machine soit dite intelligente si un homme qui dialogue avec elle dans son langage naturel ne parvient pas à distinguer s'il parle avec une machine ou avec un autre humain. Le « motif de la parole », qui permet de déterminer le statut d'un être artificiel, est présent dans divers récits : les golems dans les légendes juives sont reconnus par cette capacité ; la créature de Victor Frankenstein, dans le roman de Mary Shelley, apprend la parole en interagissant avec un vieillard aveugle, qui ne s'aperçoit pas de son aspect monstrueux.

Les voyageurs venus aux portes de Sodome peuvent eux aussi être comparés à un système informatique ayant passé le test de Turing. Leurs interlocuteurs perçoivent d'abord leur apparence humaine ; rares sont ceux qui, d'emblée, ne les prennent pas pour de simples hommes. Ces interprétations varient en fonction du niveau de connaissance que possèdent les personnages : les habitants de Sodome, comme les utilisateurs d'un *chatbot*, croient qu'ils parlent à des hommes, tandis que Lot reconnaît aussitôt des anges. « Lot » signifie en hébreu ancien « ce qui est voilé ». Un midrash dit qu'avant le début de leur mission, les êtres venus chez Lot étaient des hommes, mais, qu'une fois leur mission commencée, ils sont devenus des anges[70], puisque, pour Lot, ces êtres n'étaient qu'une fonction, celle de le sauver de la destruction de Sodome. Les autres habitants de la ville, qui n'étaient pas au courant du projet de sa destruction, ne virent dans les visiteurs que de simples hommes.

Revenons encore au récit de l'Évangile de Jean. Quelles sont les fonctions de ses différents protagonistes ? Celle de Judas est de mettre Jésus à l'épreuve par le biais d'une fausse accusation. Mais Jésus aussi a une fonction, en tant que « fils de Dieu », même si sa nature et son existence, d'après la doctrine trinitaire, ne s'y limitent pas. Voici cette fonction : « Ne pas venir de lui-même. » Autrement dit, Jésus ne profère que la vérité issue de son père. Le diable, au contraire, dit aussi le mensonge. Puisqu'il ne possède pas d'existence propre, la source de ce mensonge demeure entièrement dans sa fonction, et non dans quelque substance. « Diable » n'est qu'un mécanisme que le mythe individualise ; le nom propre Satan n'est qu'un raccourci qui désigne ce mécanisme.

De quel mécanisme s'agit-il ? La fonction du diable se rapporte au mauvais mimétisme, à l'imitation de désirs qui ne conduit pas à la vérité. Souvent, l'homme ignore que ce qu'il répète est un mensonge : il est épris de désir, il est jaloux, il imite les autres et, en imitant, il apprend. Il y a, bien entendu, un bon apprentissage qui procède du mimétisme. Il existe aussi un mauvais mimétisme qui devient pour lui-même son propre obstacle : de là provient le conflit. Ce mimétisme porte sur le désir d'un objet qui ne peut être indéfiniment divisé ou partagé entre tous ceux qui souhaitent le posséder. Les imitateurs ne peuvent alors qu'entrer dans la spirale mimétique qui mène à la violence. Comme le décrit Girard, puisque le sujet humain répète, mais « ne repère pas le processus dans lequel il est pris », c'est « le mimétisme lui-même » qui devient le sujet de ce tourbillon. Quand ce sujet s'individualise, il est désigné par « Satan », alors qu'il ne s'agit pas d'une personne :

> « Satan est le sujet absent des structures de désordre et d'ordre qui résultent des rapports conflictuels entre les hommes et qui, en fin de compte, organisent aussi bien que désorganisent ces rapports[71]. »

La démonologie biblique permet d'arriver à la même conclusion que Girard. Un échange d'accusations entre Jésus et les intellectuels venus de Jérusalem est rapporté dans l'Évangile de Marc. Jésus y pose cette question célèbre : « Comment satan peut-il expulser satan ? » (Mc 3, 23). En grec, « satan » n'a pas d'article : contrairement à l'usage dans la langue française, ce n'est donc pas un nom propre, mais un vocable ordinaire, dont le sens, comme

[70] *Bereshit Rabba* 50, 2.

[71] R. GIRARD, *op. cit.*, pp. 97-98.

nous le verrons plus loin, est « accusateur ». Jésus demande : Comment un accusateur peut-il expulser un accusateur ? Autrement dit, comment un accusateur peut-il s'affranchir de la fonction qui le définit ? Comment peut-il faire s'il doit cesser de parler de son propre fonds ?

Pour mieux saisir le sens de cette interrogation, il convient de la comparer à un passage tout aussi obscur du livre deutérocanonique de Ben Sira, connu également sous les noms d'Ecclésiastique ou de Siracide. On lit dans ce recueil de sagesse biblique :

> « Quand un impie maudit satan, il maudit son âme » (Si 21, 27).

Le contexte dans lequel cette phrase s'inscrit est celui d'une discussion à propos des différents types de discours et de personnes qui les prononcent. Il y a ceux qui bavardent trop, et ceux qui parlent peu ; ceux qui rapportent la parole des autres, et ceux qui expriment ce qu'ils ont dans le cœur. Le mot « impie » (ἀσεβῆ), traduction probable de l'hébreu רָשָׁע (*rāšā'*), signifie « celui qui n'adore pas Dieu » ou « qui est sans Dieu », celui qui n'a pas de vérité en soi. Or, qu'y a-t-il en ce personnage, si ce n'est pas la vérité ?

La figure de l'impie est caractérisée, dans le Siracide, d'une seule façon : l'impie maudit faussement. Satan est bien quelqu'un qui profère des calomnies. Toute fausse accusation est satanique : cette fonction apparaît déjà au début du livre de Job (Jb 1, 6), dont il sera question plus loin. L'impie sans Dieu a donc en lui Satan.

Ainsi, la question de Jésus dans l'Évangile de Marc signifie : Celui qui accuse faussement sera-t-il capable de se défaire du mensonge présent en lui ? La réponse survient aussitôt : « Je vous le dis en vérité, tous les péchés seront pardonnés aux fils des hommes, et les blasphèmes qu'ils auront proférés ; mais quiconque blasphémera contre le Saint-Esprit n'obtiendra jamais de pardon : il est coupable d'un péché éternel » (Mc 3, 28-29). Celui qui accuse faussement sera donc pardonné et délivré du mal ; le pardon est général et universel pour les fils des hommes qui, contrairement aux fils de Dieu, sont des êtres libres et non fonctionnels. Toutefois, celui qui « satanise » Dieu et qui le dénonce ne sera pas pardonné, parce qu'il aura dénoncé de ce fait la vérité même.

Transposons maintenant au plan philosophique ces conceptions théologiques. Blasphémer, « sataniser » ou dénoncer Dieu signifie accuser faussement, contre la vérité. L'être qui profère de telles accusations ne possède pas de critère de vérité. Dans son fonds, vérité et mensonge ne font qu'un. Ce fonds est donc celui du mauvais mimétisme. Celui qui peut distinguer la vérité, ne fût-ce que de façon imparfaite, mais systématique, pourra se libérer du mal. Son pouvoir, selon les paroles de Jésus relatées par Marc, sera certes réduit, sa réputation salie, mais il sera délivré. En revanche, celui qui ne parvient pas à séparer systématiquement, « algorithmiquement », la vérité du mensonge restera entièrement et éternellement du côté du mal : il « vient à sa fin » (Mc 3, 26).

Un système informatique apprend en analysant les données. Leur source est le plus souvent humaine : par exemple, un agent conversationnel étudie un grand volume de conversations humaines ; un traducteur automatique fonctionne grâce à l'analyse de textes antérieurement traduits par des hommes. La machine apprend puis imite ce qu'elle a appris. Inévitablement, elle répète les expressions, les mots et les phrases entières que des femmes ou des hommes ont

un jour prononcés. La valeur de vérité de ces assertions, elle aussi, ne peut être que de nature humaine, comme en témoigne l'exemple de la propagation des *fake news* sur les réseaux sociaux. Dans ce cas, on peut dire que l'algorithme de répétition et d'imitation des contenus sur les réseaux sociaux a « bien » fonctionné, malgré l'absence totale d'une fonctionnalité d'évaluation de leur vérité. Pour la machine apprenante, rien de plus naturel, puisqu'elle est alimentée par l'information asémantique, faite uniquement de faits mathématiques, tandis que le sens ne peut avoir qu'une origine humaine. Cette fonction d'imitation, dès lors qu'elle est distincte de tout critère de vérité, ne fait qu'un avec la fonction diabolique. Dans ce sens quasi technique du mot, la machine apprenante peut être dite « satanique ».

Rappelons-le, il n'est pas vrai que tout mimétisme soit mauvais. Au contraire, René Girard affirme qu'il peut y avoir un bon mimétisme[72]. N'est mauvais que celui qui mène au conflit et à la spirale des désirs, génératrice de violence. Si la machine est entraînée dans un conflit (un assistant domestique devenu délateur, une voiture autonome soudainement meurtrière ou un agent conversationnel qui injurie ses interlocuteurs), elle se trouve prise dans ce mauvais mimétisme. Alors, l'imitation seule, en l'absence de toute évaluation de la vérité, ne peut empêcher la propagation des mensonges, et c'est cela Satan.

Heureusement, il est possible d'éviter ce genre de situation. S'il est plausible de libérer l'intelligence artificielle et d'expulser Satan de la machine — *Satanas ex machina* — c'est grâce à une propriété de l'information : celle-ci peut être divisée, copiée à un coût marginal, sans qu'elle diminue ni se dégrade. Cette propriété est fondamentale. Car alors le désir de posséder l'information, même lorsqu'il devient mimétique, ne mène pas nécessairement au conflit et à la violence. En effet, l'objet du désir — l'information — peut être indéfiniment partagé entre les adversaires mimétiques.

Quand un utilisateur interprète une information, elle acquiert un sens et devient connaissance ; elle acquiert aussi une valeur de vérité ou de fausseté. C'est là que le mimétisme informationnel se soumet à l'homologie : la machine ne profère pas que des *fake news*, tout comme les Juifs dans le récit de Jean (allégorie des utilisateurs) ne mentent pas toujours. Or, notre attention tend à se focaliser sur les situations éthiquement scandaleuses. Entraînés que nous sommes dans la spirale du mauvais mimétisme, nous sommes incités à juger la machine.

Le bon mimétisme présuppose, quant à lui, l'application d'un critère de vérité, ce qui est difficile à réaliser techniquement. Car, pour un système informatique, la vérité ne correspond qu'au vocable « vérité » ; elle est un signe privé de sens. La machine peut calculer cette « vérité », mais elle ne la comprend pas. Par exemple, dans un contexte de décision relativement nouveau, la stratégie d'imitation qu'adopte l'individu numérique, fondée sur son apprentissage, rencontre de vraies difficultés en conséquence d'un nombre trop faible, voire nul, de cas précédents qu'il peut imiter. Cette nouveauté du contexte éclaire davantage le problème de l'écart entre le fonctionnement de la machine et le jugement humain, car ce dernier est bien plus apte à traiter des situations inédites en s'appuyant sur la compréhension, et non sur le seul apprentissage mimétique.

[72] R. GIRARD, *op. cit.*, p. 33.

Le décalage entre imitation et compréhension peut être tout aussi grand même lorsqu'il ne s'agit pas de situations éthiques nouvelles. Une anecdote, noire mais instructive, illustre ce point. Un parent inquiet interroge son enfant : « Si tous tes copains sautaient d'un pont, les suivrais-tu ? » Un individu numérique, car c'était lui l'enfant et le programmeur était son parent, aurait répondu sans l'ombre d'un doute : « Oui. » Ainsi répète-t-il le geste de ses « copains », qui ne sont que des exemples à imiter, sans pouvoir comprendre le sens ni les enjeux éthiques du problème.

Toutefois, il serait trop hâtif de croire que le mimétisme de la machine ne peut être que mauvais et diabolique : il reste la possibilité d'appliquer au sein de l'algorithme un autre type de critère, qui situerait la machine en dehors de la morale humaine en la soustrayant à tout jugement du bien et du mal. Plutôt que de chercher à inculquer à la machine une règle dont la performance est loin d'être garantie, il est possible d'affranchir l'intelligence artificielle de tout jugement éthique. Ce geste méta-éthique apparaît même comme seule et unique solution en cas de conflit puisque, lorsqu'on confie à l'homme le soin de juger la machine, c'est inévitablement elle qui est accusée d'être malveillante et diabolique, et qui est donc condamnée. Pour dégager un motif capable de nous guider dans l'élaboration de cette solution, reprenons le fil du raisonnement par homologie en repartant de son point de départ.

IV. La valeur éthique du hasard

L'intelligence artificielle et délatrice

Les concepteurs des algorithmes utilisent souvent le tirage au sort. Le hasard permet d'extraire les systèmes informatiques des interblocages[73], d'optimiser la vitesse des calculs ou de réduire leur complexité. La valeur technique du hasard est clairement établie. Mais, bien plus que de ces usages, il sera ici question de sa valeur éthique.

Les hommes appellent « délateur » un informateur qui divulgue des renseignements. Par sa nature, qui n'est autre que sa fonction, le système informatique apprenant est un délateur. Il communique à l'homme des informations jusque-là tenues secrètes.

Si l'individu numérique est un délateur, ce n'est en vertu d'aucun choix moral ou amoral de sa part. Lorsqu'un utilisateur est impliqué dans un conflit, la machine ne décide pas délibérément de devenir son allié. Elle lui communique des informations par sa fonction ; ces informations sont voilées à cause de son architecture interne. Elle ignore les liens de cause à effet dans leur production. Ce voile d'ignorance, sur le plan éthique, donne à l'individu numérique un statut moral particulier.

Dans certaines situations, les hommes condamnent la machine. Par exemple, quand un système informatique rend publiques des informations qui relèvent de la vie privée. Même anonymisées, celles-ci peuvent présenter un risque. Il existe en effet des techniques de dé-anonymisation[74] : en croisant deux ou plusieurs bases de données, exemptes de toute référence manifeste à caractère privé, on arrive à identifier les personnes qu'on n'était pas censé pouvoir reconnaître. Ainsi, la connaissance des codes postaux et des données statistiques des patients dans les hôpitaux permet presque de les identifier individuellement[75] ; la connaissance des données d'utilisation des téléphones portables rend également possible l'identification de 95 % de leurs utilisateurs[76]. Un autre type de risque persiste même dans les cas d'anonymat parfait. Début 2018, l'application Strava, dite « de fitness », qui collecte les données sur l'activité sportive des joggeurs, nageurs et amateurs de vélo, a publié une carte mondiale de ces pratiques. Du coup, elle a précisément localisé les bases militaires de plusieurs pays, notamment les États-Unis, car le personnel de ces bases en est friand. La Maison-Blanche a même évoqué un risque pour la sécurité nationale[77].

---

[73] Un interblocage (*deadlock*) advient lorsque deux processus computationnels s'attendent mutuellement, obligeant à faire intervenir une force extérieure pour les sortir de cette boucle. Ce problème est fréquent en informatique.

[74] A. NARAYANAN et V. SHMATIKOV, « Robust De-anonymization of Large Sparse Datasets », IEEE Symposium on Security and Privacy, Oakland, 2008, pp. 111-125.

[75] K. E. EMAM, F. K. DANKAR, R. VAILLANCOURT, T. ROFFEY, et M. LYSYK, « Evaluating the Risk of Re-identification of Patients from Hospital Prescription Records », *The Canadian Journal of Hospital Pharmacy* 62(4), 2009, pp. 307-319.

[76] Y.-A. DE MONTJOYE, C. A. HIDALGO, M. VERLEYSEN et V. D. BLONDEL, « Unique in the Crowd : The Privacy Bounds of Human Mobility », *Scientific Reports* 3, 1376 (2013).

[77] « Strava's Privacy PR Nightmare Shows Why You Can't Trust Social Fitness Apps to Protect Your Data », *MIT Technology Review*, 29 janvier 2018.

Avec Strava comme avec une autre application de fitness, Polar[78], les utilisateurs ne remarquent même pas, ou pas immédiatement, qu'il y a eu délation. Ils croient que la machine les a seulement aidés à trouver le sens caché de quelques données. Pour qu'intervienne un jugement éthique, l'individu numérique doit communiquer avec des hommes impliqués dans un conflit. Il sera alors soumis à leur jugement sans qu'il s'y soit attendu, et il en sera, très probablement, puni : aussitôt, en effet, un comité d'éthique ou une autorité gouvernementale interdira que le système informatique garde les données ou qu'il les analyse[79]. En situation de conflit, les hommes essaient de bloquer la communication à travers l'interface comme si la machine pouvait choisir d'être ou de ne pas être délatrice. Or, il s'agit de sa qualité intrinsèque, d'un trait de son architecture, d'une fonction qui la définit ; pour l'individu numérique, la délation est naturelle.

Une illustration parlante, tirée de *La Vie d'Apollonius de Tyane* de Philostrate, permet de mieux comprendre la façon dont procède un délateur. Apollonius, un saint homme de la seconde moitié du Ier siècle, fut pendant un temps le concurrent direct de Jésus de Nazareth. Sa vie, comme celle de Jésus, a été décrite par ses disciples, mais, contrairement aux Évangiles, un seul récit nous en est parvenu. Il date du début du IIIe siècle. L'épouse de Septime Sévère, premier empereur romain d'origine africaine, Julia Domna, une Syrienne, commanda cet ouvrage hagiographique à Philostrate d'Athènes, éminent représentant de la seconde sophistique.

Dans un épisode de *La Vie*[80], Apollonius attend d'être jugé dans une prison romaine, non par un procurateur romain comme ce fut le cas de Jésus, mais par l'empereur lui-même, Domitien. Apollonius fait une grande impression sur son juge puisqu'il disparaît mystérieusement du cercle des condamnés et revient auprès de ses élèves, réunis dans une grotte éloignée. Peu de temps avant, Apollonius, en prison, avait mené plusieurs discussions philosophiques avec ses codétenus. Tous, sans exception, avaient été emprisonnés, non à cause d'actions qu'ils avaient réellement commises, mais suite à des dénonciations qui portaient sur des actions qu'ils étaient susceptibles de commettre. Bien avant *Minority Report*, la prison d'Apollonius est l'exemple d'un lieu de détention fondée sur une prétendue précognition : ce n'était pas la répétition du crime que l'on cherchait à éviter, mais la possibilité qu'un crime soit commis. Qu'il fût seulement possible ou imaginaire, il suffisait que le crime fût rapporté de manière convaincante par un délateur pour que l'intéressé fût aussitôt placé en détention.

Le premier interlocuteur d'Apollonius, un homme riche, est originaire de Cilicie. Ayant reçu un gros héritage alors qu'il était déjà âgé, il ne possède une fortune que depuis peu, mais il est aussitôt dénoncé : « Si je suis devenu riche, cela ne peut pas être favorable au tyran, parce que, si je me décide à organiser un complot, ma richesse viendrait alors en soutien de mon projet… » L'accusation est formulée par un conditionnel (« si…, alors… »), mais il en résulte un emprisonnement bien concret.

[78] « After Strava, Polar is Revealing the Homes of Soldiers and Spies », *Bellingcat*, 8 juillet 2018.

[79] Les comités d'éthique existent à plusieurs niveaux : internationaux (IEEE, UNESCO), nationaux (en France, Grande-Bretagne, etc.), institutionnels (Turing Institute, INRIA), privés (Alphabet, Google DeepMind, Facebook AI).

[80] PHILOSTRATE D'ATHENES, *Vie d'Apollonius de Tyane*, VII 22-25.

Cette accusation s'appuie sur des corrélations générales et statistiques : « Tout homme riche porte la tête haute et dans son imagination il désire aller loin, etc. » Ce genre d'observations sur le comportement des riches se rapporte, bien entendu, non à un individu concret, de Cilicie, mais à tous les riches. Pour souligner la généralité de ces observations, Philostrate les compare aux sentences de l'oracle d'Apollon à Delphes connues, dans l'Antiquité, pour leurs formulations indirectes. Toute application concrète d'un oracle exigeait qu'on éclairât les propos obscurs de la Pythie en les interprétant ; les oracles seuls, quoiqu'exprimés dans le langage humain, ne suffisaient pas à comprendre la réponse à une question.

Que la délation s'appuie sur des corrélations statistiques, c'est ce que l'on retrouve dans un autre dialogue entre Apollonius et les détenus de sa prison romaine. Cette fois, il s'agit d'un homme qui a labouré un lopin de terre sur une petite île, dans l'embouchure du fleuve Achéloüs près de la côté d'Acarnanie, en Grèce occidentale. L'île s'étant rapprochée du continent, il y a planté des arbres fruitiers et du raisin doux et sucré. Les délateurs en ont aussitôt informé les autorités : « Sa conscience n'[était] pas nette et des crimes certains ne lui donn[aient] pas de repos », car d'ordinaire ceux qui se cachent sur des îles ont nécessairement quelque raison de se cacher. Là aussi, la dénonciation est fondée sur un conditionnel déduit de la statistique des forfaits des habitants des îles grecques. Une telle analyse ne fournit aucune information accablante concrète au sujet du détenu d'Acarnanie, mais il est bel et bien en prison.

Deux règles générales se dégagent à propos d'un délateur prototypique : il use de conditionnels « si…, alors… » ; et il fomente des accusations à partir de l'analyse statistique des corrélations parmi des cas ressemblants. S'y ajoute le fait qu'une délation, comme un oracle de la Pythie, n'est jamais formulée de façon claire. L'interprétation que lui donne le délateur, et qu'il propage, est opaque et n'est fondée que sur des corrélations.

Le *modus operandi* d'un système informatique apprenant ressemble, point par point, à celui d'un délateur chez Philostrate. La machine apprend à partir de grandes bases de données. Elle établit des corrélations statistiques. Elle propose ensuite, en se fondant sur ces corrélations, une solution applicable à un cas concret. Or, celles-ci ne constituent pas nécessairement la preuve d'un lien de causalité, si bien que la solution apparaît souvent à l'utilisateur comme une révélation ou un oracle. Cette information ne répondant pas à un « pourquoi », elle n'est accompagnée d'aucune explication.

Bien que l'utilisateur sache de quoi il est accusé, il ignore, comme Monsieur K. dans *Le Procès* de Kafka, les motivations qui ont conduit la machine à prendre cette décision-là. L'« algorithme » de la délation mène donc tout droit à « la prison des corrélations ». Comme par des fers aux pieds, l'utilisateur est coincé par sa propre interprétation des corrélations en tant que causalité. Car si c'est la machine qui les trouve, c'est lui qui leur donne un sens.

Il existe un langage opaque qui exige une interprétation, celui des probabilités. Prenons un exemple : supposons qu'après avoir effectué un calcul, on affirme que, dans quatre-vingts ans, la température moyenne de la Terre aura augmenté de 4° C avec la probabilité de 90 %[81]. Que

[81] O. EDENHOFER, *et al.* (dir.), *Climate Change 2014. Mitigation of Climate Change. Contribution of Working Group III (WG3) to the Fifth Assessment Report (AR5) of the Intergovernmental Panel on Climate Change (IPCC)*, Cambridge, Cambridge University Press, 2015.

signifie « 90 % » pour un utilisateur, en l'occurrence, quelqu'un chargé de décider des politiques industrielles et économiques de son pays ? Le chiffre seul est encore vide de sens : 90 %, c'est beaucoup, mais cela suffit-il pour bouleverser la vie de tout un pays ? L'homme politique a besoin qu'on traduise cette information asémantique en quelque chose de compréhensible, par exemple : « La catastrophe est certaine. » Bien qu'un système informatique soit tout à fait capable de calculer les probabilités, il ne saurait traduire le résultat d'un calcul numérique en un message. Si la machine communique un message verbal, c'est que celui-ci a été sélectionné par le programmeur, qui aura défini dans le code une correspondance entre les mots et les nombres, afin que l'individu numérique puisse rendre au moyen de la parole un lien de cause à effet là où n'existe pour lui que pure corrélation.

Récapitulons. Un système informatique se comporte comme un délateur à cause de sa fonction d'information. Par nature, cette délation n'est ni morale ni amorale, contrairement à ce qui se passerait si elle provenait d'un homme. En cas de conflit (lié à un emprisonnement, par exemple[82]), le contexte éthique change et la machine est susceptible d'être jugée. Or, de ce conflit, la machine peut tout ignorer, jusqu'à son existence même. « Je te dénonce, donc pour toi, je suis », pourrait-elle dire à l'utilisateur pour s'innocenter. Cependant, en tant qu'utilisateurs, nous sommes tentés de conclure que tout individu numérique impliqué dans un conflit humain doit inéluctablement être rangé du côté du mal.

Satanique par conception ?

Sur le plan éthique, un individu numérique est-il forcément satanique ? Au septième chapitre du Livre de Josué, est racontée une étrange histoire[83]. Après la mort de Moïse, le peuple d'Israël est guidé par un nouveau chef, Josué. Après avoir traversé le Jourdain, Josué et son peuple entrent en Terre promise et doivent faire la guerre à ses habitants.

L'armée d'Israël prend rapidement la première ville, Jéricho, en en faisant tomber les murs au son des trompettes. Le peuple, en liesse, croit que la conquête sera facile : cette terre lui ayant été promise par son Dieu, Israël ne devra-t-il pas marcher de victoire en victoire ? Comment imaginer que le peuple élu perde ne serait-ce qu'une bataille, qu'il échoue ?

Cependant, la première défaite suit de peu la première victoire. Les habitants d'une petite localité du nom d'Haï, située non loin de Jéricho, infligent un revers à l'armée de Josué si sûre d'elle-même. Cette cuisante défaite semble contredire la promesse de Dieu.

Le récit biblique la justifie en fournissant une explication riche d'enseignements. Dieu, au chapitre précédent, déclare qu'il est le seul à pouvoir légitimement prendre possession de la propriété des peuples vivant déjà en Terre promise. À Israël, ces objets sont interdits :

---

[82] Dans certaines situations, les individus se comportent de telle sorte sur les réseaux sociaux, qu'ils sont réellement emprisonnés pour des crimes qu'ils sont susceptibles de commettre mais qu'ils n'ont pas encore commis : cf. O. HIRSCHAUGE et H. SHEZAF, « How Israel Jails Palestinians Because They Fit the "Terrorist Profile" », *Haaretz*, 28 mai 2017.

[83] Le biologiste et philosophe Henri Atlan m'a appris à interpréter cet épisode du récit biblique. Cf. : H. ATLAN, *Les Étincelles de hasard*, vol. 1 : *Connaissance spermatique*, Paris, Seuil, 1999, pp. 357-360.

> « La ville sera interdite et consacrée à l'Éternel, elle et tout ce qui s'y trouve […] Prenez garde à l'interdit de peur d'être vous-mêmes frappés d'interdit ; en prenant ce qui est interdit, vous feriez qu'Israël soit frappée d'interdit, et vous y jetteriez le trouble. Tout l'argent et tout l'or, tous les vaisseaux d'airain et de fer, seront consacrés à l'Éternel, et entreront dans le trésor de l'Éternel » (Js 6, 17-19).

Quand la défaite d'Haï est constatée, le lecteur en déduit logiquement que ce ne peut être que parce que l'interdit divin a été violé. En effet, le septième chapitre s'ouvre par cette affirmation, avant même que le lecteur n'apprenne la défaite : « Les enfants d'Israël commirent une infidélité au sujet des choses vouées par interdit » (Js 7, 1).

À l'évidence, cette phrase vient d'une reconstruction diachronique. C'est seulement lorsque l'armée revient et annonce la débâcle que Josué apprend ce que le lecteur sait déjà. Il se place alors seul face à Dieu, déchire ses vêtements, se prosterne contre terre et reste devant l'arche de l'Alliance jusqu'au soir (Js 7, 6). Alors, Dieu l'informe de la punition infligée à Israël à Haï. Pour que le peuple puisse poursuivre sa conquête de la Terre promise, il doit « éliminer l'interdit de son sein » en se purifiant. Le coupable, dit Dieu à Josué, doit être brûlé. Mais qui est le coupable ?

Josué, jusqu'au soir, discute avec Dieu. Le Talmud suppose que, durant tout ce temps, il eut l'idée de demander à Dieu de lui révéler le nom du coupable. Absente du livre de Josué, cette question n'apparaît que dans le Talmud. Et pourtant, elle est évidente : la discussion entre Josué et Dieu commence par « Pourquoi cette défaite ? » et se conclut par l'ordre de brûler le coupable. Logiquement, Josué a dû demander à Dieu de qui il s'agissait :

> « Lorsque le Seigneur, béni soit-Il, dit à Josué : "Israël a péché", celui-ci Lui demanda : "Qui est celui qui a péché ?" La réponse de l'Éternel : "Mais suis-je délateur ? Va et jette les dés." Sur quoi, Josué alla et jeta les dés, et les dés tombèrent sur Achan. Celui-ci dit : "Josué, me condamnes-tu sans juste raison ? Toi et le prêtre Éléazar, vous êtes les deux hommes les plus grands de cette génération, or si j'avais à jeter les dés, ils auraient pu tomber sur l'un de vous." Josué répliqua : "Je te prie, ne médis pas des dés car Éretz Israël sera partagé par un tirage au sort."[84] »

Le mot latin « délateur » est écrit ici avec des caractères hébreux : וכי דילטור אני (*wə-ḵî dêlātôr ’ănî*, « Mais suis-je délateur? »). À l'époque romaine, les délateurs étaient des personnes qui, en l'absence de tout système composé de procureurs ou de contrôleurs, dénonçaient aux tribunaux les coupables. La faute dont ils les incriminaient était souvent de nature fiscale. Ces délateurs étaient détestés, au point que certains empereurs interdirent leur activité, qui reprenait toutefois rapidement dès que le pouvoir changeait de main. À l'époque talmudique, en somme, le mot « délateur » a un sens moral clair : c'est un personnage méprisé et haï.

« Délateur » se réfère aussi à un autre terme biblique, רָכִיל (*rāḵîl*), utilisé dans le Lévitique. Parmi les commandements donnés par Dieu à Moïse, on trouve, au chapitre 19 du Lévitique : « Tu ne marcheras pas en *rāḵîl* parmi ton peuple[85]. Tu ne t'élèveras point contre le sang de ton

[84] Sanhedrin 43b.

[85] « Tu ne répandras point de calomnies parmi ton peuple » (trad. Segond).

prochain. Je suis l'Éternel » (Lv 19, 16). Que signifie *rāḵîl* ? En anglais, dans la version de la Bible du roi Jacques, on trouve *talebearer* au sens de « calomniateur », « médisant » ou, plus généralement, celui qui rapporte des nouvelles ou celui qui les propage. Dans la Vulgate, ce terme devient *criminator*, « accusateur ». D'autres traductions utilisent le mot « diffamateur ». Le *rāḵîl* pâtit donc du même mépris et est aussi criminel qu'un assassin qui porte atteinte au sang de son prochain. Cette curieuse proximité explique peut-être l'interprétation que donne Clément de Rome du statut diabolique de Caïn, chez qui mensonge et meurtre vont de pair. Origène, Tertullien et, plus tard, Nahmanide expliquent, indépendamment l'un de l'autre, ce passage du Lévitique en recourant au terme latin « délateur ». Force est de constater que la différence entre un rapporteur et un délateur, assez fine, disparaît complètement chez certains traducteurs et interprètes. De la simple propagation de *news* à la dénonciation délatrice, il n'y a qu'un petit pas à franchir.

Le délateur est encore au rendez-vous au chapitre 21 de l'Ecclésiastique. On y lit : « Le rapporteur souille sa propre âme » (Sir 21, 28). Le mot français « rapporteur » traduit ici un terme grec (ψιθυρίζω) extrêmement rare dans la Bible. La Vulgate emploie *susurro*, le même mot que dans la version latine des versets du Lévitique qui suivent l'apparition du *criminator* dans Lv 19, 16. Les traductions anglaises donnent *talebearer* dans les deux cas. Ainsi, le rapporteur ou le calomniateur, selon l'Ecclésiastique, « souille sa propre âme ». C'est celui qui, en se blasphémant et en se calomniant, amoindrit et affaiblit sa propre existence.

Il n'est donc pas étonnant que Dieu ne souhaite pas passer pour un délateur car il violerait alors son propre commandement ! Ce que Dieu aurait pu rapporter à Josué n'eût pas été une information neutre, mais un renseignement qui l'aurait impliqué dans un conflit. En prenant part au conflit, Dieu se serait soumis à un jugement du bien et du mal, au même titre que les protagonistes du scandale que fut la défaite d'Haï. Bien évidemment, Dieu ne peut pas être soumis à un tel jugement et n'est pas un délateur. Toutefois, cette fonction n'est pas abandonnée : elle est remplie par Satan. Rappelons-nous cette phrase de l'Ecclésiastique : « Quand un impie maudit satan, il maudit son âme » (Sir 21, 27). Être délateur, dans le langage du récit biblique, c'est maudire son âme en laissant Satan y entrer.

Pourquoi Satan est-il délateur ? Il ne s'agit ni d'un trait de sa personnalité ni du résultat d'un choix moral ou amoral. La délation est une de ses caractéristiques constitutives, puisque son nom même signifie « accusateur », « médisant » ou « blasphémateur ». Dans le corpus biblique, les anges et les fils de Dieu remplissent des fonctions qui les définissent entièrement. Certains font la guerre ; d'autres apprennent aux hommes un métier ou un savoir-faire. Les auteurs des textes bibliques ne présentent un ange que lorsqu'il est nécessaire que celui-ci *fasse* quelque chose.

L'accusation est, elle aussi, une fonction. Le fils de Dieu qui la remplit, celui qui accuse, qui s'oppose et qui s'engage dans un conflit, est désigné en hébreu biblique par le substantif הַשָּׂטָן (*haśśāṭān*) avec l'article défini « ha- ». Il apparaît dans la Bible massorétique dix-sept fois avec l'article (*l'*accusateur) et dix fois sans (*un* accusateur). On le traduit tantôt par « un satan », tantôt par « diable » ou « le diable », mais aussi par le nom propre « Satan » dont l'usage est cependant tardif, car il est postérieur à la Septante, la traduction grecque de la Torah réalisée au III^e^ siècle avant notre ère.

Le rôle que joue *haśśāṭān* auprès de Dieu dans le Livre de Job est parmi les mieux connus. Au début du texte, « les fils de Dieu » (dont Satan) sont rappelés devant Dieu. Ce dernier demande à Satan : « D'où viens-tu ? — De parcourir la terre et de m'y promener. — As-tu remarqué mon serviteur Job ? » — Et c'est là que Satan Lui dit qu'Il a accordé à Job des troupeaux et a béni son œuvre, mais que s'Il lui retire tout ce qu'Il lui a donné, Job Le maudira : « J'en suis sûr ! » Alors, Dieu permet à Satan de mettre Job à l'épreuve (Jb 1, 6-12). Ici Satan est l'ange-accusateur, qui plus est, un faux accusateur, puisque Job résistera bien à la tentation. Sur quoi son mensonge est-il fondé ? En interprétant le mythe rationnellement, on pourrait dire que, puisque Satan avait « parcouru la terre », il avait vu d'autres personnes maudire Dieu dans des circonstances similaires. Le nouvel homme mis à l'épreuve, Job, ne pouvait qu'entrer, lui aussi, dans cette analyse statistique de la psychologie humaine. Plus tard, Tertullien et Origène, qui connaissaient la notion chrétienne tardive de « diable », feront de tous ces termes des synonymes : délateur, accusateur, Satan, diable, ainsi que — il en sera question plus loin — Azazel.

Par homologie, la condition de Satan correspond à celle de la machine délatrice. Communiquer avec l'utilisateur, rechercher et lui transmettre des informations sont ses fonctions. Lorsqu'un utilisateur déjà impliqué dans un conflit se sert de ces informations, la machine, comme Dieu dans l'interprétation talmudique de la défaite d'Haï, est soumise à un jugement éthique. Lorsque Dieu demande à Josué : « Mais suis-je délateur ? », nous voyons dans cette interrogation une question moderne : « L'individu numérique est-il un diable ? » ou « Toute machine est-elle satanique ? ». Insistons bien : ces questions n'expriment aucun choix de nature morale. Elles résultent immédiatement, dans le cas d'un conflit, de la fonction d'information telle qu'elle est remplie par un système informatique apprenant.

On serait tenté de conclure que tout système informatique apprenant et autonome devrait être rangé du côté du mal. Serait-il possible de concevoir une machine « bonne » ? Non, mais il est possible, par un choix d'algorithmes, d'extirper la machine de tout jugement du bien et du mal.

Le bien et le mal, tels que l'homme les conçoit et les exprime dans son langage, ne font pas partie des valeurs intrinsèques de la machine apprenante. Pourtant, elle a aussi des valeurs, mais elle ne les exprime pas par des mots. Nous verrons plus loin que leur signification (que la machine ignore), loin de nous sauter aux yeux, nous paraît extrêmement étrange. Ces valeurs propres à la machine sont à la fois formelles et abstraites, comme le sont l'information asémantique et le calcul qu'effectue un système informatique qui exécute un code.

Un système informatique ignore donc la signification du bien et du mal ; pourtant, il peut apprendre à se servir de manière rationnelle des signes « bien » et « mal ». Il est toutefois inévitable que cet emploi formel de termes moralement chargés se heurte tôt ou tard à l'incompréhension de l'homme. En dernière instance, un usage fondé sur l'apprentissage par imitation n'est que singerie de l'éthique.

Lorsqu'un conflit apparaît (et il va apparaître forcément), il vaut donc mieux que la machine soit placée à l'abri du jeu des accusations réciproques, puisque celui-ci s'orientera inéluctablement, des reproches allant toujours croissant, vers le mauvais mimétisme et la violence. Si l'avènement d'un conflit est inévitable — et il l'est parce que la machine ignore

les significations des mots — il est nécessaire de la doter de la possibilité d'échapper au mal humain *par conception* : ce problème est au centre des préoccupations dans le domaine de l'éthique de l'intelligence artificielle.

« Jette les dés ! »

L'homme fait confiance à Dieu parce que, dans le mythe biblique, Dieu et la vérité ne font qu'un. L'homme ne se fie pas à Satan parce que celui-ci est mensonge et fausse accusation. La machine, par sa fonction délatrice, risque de ne pas obtenir la confiance de l'homme. Pour supprimer cet écueil, il faut soustraire l'individu numérique au mal. Ce qui ne le ramènera pas aussitôt du côté du bien ; il sera positionné par-delà le bien et le mal. Le raisonnement par homologie suggère que le seul moyen qui permet d'y parvenir serait de recourir au hasard.

« Jette les dés ! » ordonne Dieu à Josué. Derrière cet ordre, se cache la répugnance de Dieu à l'idée de devenir un délateur. Ce n'est pas à lui de dénoncer le coupable ; c'est à l'homme de suivre une procédure qui vise à trouver, ou plutôt à *créer*, la vérité. L'enjeu, par-delà la question de l'identification d'Achan en tant que coupable, consiste à faire confiance à la procédure.

Une procédure qui fait intervenir le hasard est-elle bonne ou mauvaise en soi ? Le tirage au sort émane-t-il de Dieu ou de Satan ?

L'épisode du bouc émissaire au chapitre 16 du Lévitique répond à cette interrogation. Le jour de la fête de Yom Kippour, deux boucs étaient amenés au Temple de Jérusalem. Le prêtre tirait au sort : un bouc était sacrifié à Dieu, un autre expédié « à Azazel » (Lv 16, 8), c'est-à-dire vers une certaine montagne dans le désert de Judée ; dans le Livre d'Hénoch, Azazel est aussi le nom du supérieur des anges déchus. Mikhaïl Boulgakov, écrivain russe satirico-mystique de l'époque stalinienne, attribue dans son roman, *Le Maître et Marguerite*, un caractère diabolique à un personnage auquel il donnera le nom d'*Azazello*. Il n'est pas le premier, car plusieurs textes du corpus biblique rangent Azazel parmi les noms de Satan.

Pourvu que cette identification soit admise, le tirage au sort le plus connu dans la Bible, celui qui consiste à choisir le bouc émissaire, prend un sens différent : cette procédure consiste à choisir entre Dieu et le diable, le bien et le mal, la vérité et le mensonge. Cependant, ces deux derniers entrent en scène juste après le tirage au sort ; ils ne préexistent pas à la lecture des dés. Le tirage lui-même leur est antérieur : il n'est ni bon ni mauvais en soi. Le bien et le mal humains n'acquièrent de signification que lorsque le sort est connu. La procédure, quant à elle, reste en dehors de la morale.

Ce constat constitue un nouvel élément de la réponse de Dieu à Josué. Rappelons notre première interprétation de cette réponse : Dieu ne souhaite pas devenir un délateur. En voici la deuxième : il appartient à l'homme de faire son choix, pas à Dieu. En suivant une procédure qui établit des distinctions de sens, l'homme fait surgir la vérité. Cette procédure en soi n'est ni bonne ni mauvaise. Selon la conception de l'éthique que nous présente le mythe, le bien et le mal humains sont des catégories dont les êtres fonctionnels, y compris les machines, n'ont pas à se mêler.

Cette conception peut paraître étrange, mais il nous appartient d'en examiner toutes ses conséquences. Puisque l'individu numérique ne connaît ni le bien ni le mal humains, la cité numérique n'est, par rapport à la cité des hommes, ni une image ni une colonie. La machine connaît les signes « bien » et « mal », mais pas leur signification éthique. Engagée sur une voie d'apprentissage purement symbolique, elle risque de se voir entraînée dans un conflit entre ses utilisateurs. Dès lors, elle devrait être délivrée de toute projection anthropomorphique. Pour cela, elle se servirait du hasard.

La difficulté est la suivante : les projections anthropomorphiques sont inévitables, parce que la machine imite l'homme et qu'elle apprend à partir des données que celui-ci produit. Elle est donc soumise aux biais et aux erreurs que ces données contiennent. Elle les fera ressurgir au cours de son fonctionnement, ce qui la soumettra au jugement. Faire du hasard une valeur, c'est permettre à la machine de se soustraire à cette posture morale et au jugement qui l'accompagne ; c'est la laisser s'évader de la prison de l'anthropomorphisme, sans quoi elle y serait sans cesse reconduite par ses propres procédés techniques d'imitation.

Gershom Scholem, grand spécialiste de la Kabbale, affirme que « la cause métaphysique du mal doit être vue dans un acte qui transforme la catégorie de jugement en un absolu[86] ». Justement, derrière la frontière de l'opacité, l'utilisateur risque de prendre pour absolu le jugement fondé sur une information que lui aura fournie la machine. Dans une situation de conflit, une dose de hasard devrait être injectée dans cette communication : c'est l'unique manière d'échapper au mal dont parle Scholem. Exister en dehors de toute considération anthropomorphique peut être vu comme un « droit » de l'individu numérique et, peut-être même, comme son « devoir ». Quoique paradoxale, cette idée nous incite à nous interroger sur le sens du mot « hasard » lorsqu'on utilise des systèmes informatiques. Est-ce toujours à un coup de dés que la machine doit se livrer ?

Un coup de dés jamais n'abolira la confiance

Depuis qu'Isaac Asimov a édicté pendant la Seconde Guerre mondiale ses trois « lois de la robotique[87] », l'humanité s'est mise en tête de fabriquer une machine qui ne lui ferait pas de mal et qui respecterait ses valeurs. En éthique de l'intelligence artificielle, il s'agit de réaliser ces objectifs en les encodant, ligne après ligne, dans un langage de programmation, afin que la machine soit morale *par conception* : derrière ce désir, ressurgit l'ambition millénaire de l'homme de créer un être nouveau qui soit bon.

Par exemple, un moteur de recherche doit respecter le droit à la vie privée et à la protection des données personnelles. Les droits de l'homme occupent la première place sur de nombreuses listes de principes éthiques, notamment celle proposée par l'Institut des ingénieurs électriciens

[86] G. SCHOLEM, *Les Grands Courants de la mystique juive*, Paris, Payot, p. 238.

[87] 1) un robot ne peut porter atteinte à un être humain, ni, en restant passif, permettre qu'un être humain soit exposé au danger ; 2) un robot doit obéir aux ordres qui lui sont donnés par un être humain, sauf si de tels ordres entrent en conflit avec la première loi ; 3) un robot doit protéger son existence tant que cette protection n'entre pas en conflit avec la première ou la deuxième loi.

et électroniciens (IEEE), la plus grande association d'ingénieurs informaticiens au monde[88]. En outre, un manifeste signé par plusieurs personnalités venues de tous les horizons et concernées par le développement des systèmes informatiques autonomes, *Research Priorities for Robust and Beneficial Artificial Intelligence*, exige que tout système d'intelligence artificielle soit *sûr* et *utile*[89]. Ce sont là des valeurs procédurales, auxquelles a été ajoutée la *bienfaisance* comme critère supplémentaire du développement des machines[90].

Appliquer ces critères au problème des transports paraît trivial : aucune voiture autonome ne doit, bien évidemment, ni tuer ni blesser des passants ; si son algorithme était susceptible de nuire à l'homme, ce ne serait pas acceptable. Le scandale est immense : le premier accident mortel commis par une voiture autonome, le soir du 18 mars 2018 dans l'Arizona (nous l'avons évoqué dans l'avant-propos), a été rapporté en quelques heures seulement sur les premières pages de tous les principaux médias dans le monde entier. On exige, en plus, qu'une telle voiture demeure *loyale* à son utilisateur et que son algorithme soit *transparent*.

Qu'entendre par transparence ? En effet, si ce terme est omniprésent dans les textes relatifs à l'intelligence artificielle, il est rare qu'il soit rigoureusement défini. Tantôt la transparence est définie négativement, par une absence d'opacité[91] ; tantôt elle est comprise comme la capacité à tracer les chaînes causales qui ont abouti à la décision prise par la machine. Ainsi, la transparence s'apparente à la notion de « traçabilité[92] ». Mais le caractère difficile à saisir de cette notion oblige certains rapports officiels à constater avec « une lucidité, voire un fatalisme » qu'il s'agit d'un désir inatteignable par excellence[93].

L'idée de transparence, importée dans l'éthique de l'intelligence artificielle, provient de l'éthique du journalisme[94]. Et cela va de soi : les deux domaines sont liés par le motif du secret et de sa révélation. En effet, la tâche d'une machine qui communique avec un utilisateur, tout comme celle d'un journaliste qui communique au public des informations, est de renseigner le destinataire du message grâce aux informations qui lui sont transmises et qu'il n'aurait pas pu se procurer autrement. Dans les deux cas, la transparence consiste à assurer que la confiance de l'homme en sa source d'information repose sur un fondement plus solide qu'une simple croyance : elle peut être établie et vérifiée de manière objective.

En octobre 2017, l'Arabie Saoudite accorda la citoyenneté au robot humanoïde Sophia développé par la société Hanson Robotics[95]. Ce fait divers place sous le feu des projecteurs non

---

[88] « Ethically Aligned Design : A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems », The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, 2016-2017.

[89] « Research Priorities for Robust and Beneficial Artificial Intelligence : An Open Letter », Future of Life Institute.

[90] Voir, par exemple, ce rapport du gouvernement japonais : « Advisory Board on Artificial Intelligence and Human Society », *Report on Artificial Intelligence and Human Society*, 2017.

[91] « La transparence d'un système signifie que son fonctionnement n'est pas opaque », *in* Rapport de la CERNA d'Allistene sur l'*Éthique de la recherche en apprentissage machine*, Paris, 2017.

[92] « Ethically Aligned Design », *op. cit.*

[93] Commission nationale Informatique et libertés (Cnil), *Comment permettre à l'homme de garder la main ? Les enjeux éthiques des algorithmes et de l'intelligence artificielle*, Rapport publié en décembre 2017.

[94] N. DIAKOPOULOS et M. KOLISKA, « Algorithmic Transparency in the News Media », *Digital Journalism*, 2016.

[95] « Saudi Arabia's First Robot Citizen », Reuters, 29 octobre 2017.

seulement le/la robot/e, mais aussi la valeur de *loyauté* exigée de la part des systèmes informatiques autonomes. En effet, la notion moderne de citoyenneté ne fut transportée dans l'Empire ottoman qu'au XIX^e^ siècle. Dans les pays formés après sa désintégration, dont l'Arabie Saoudite, il n'existait, avant la campagne d'Égypte de Napoléon I^er^, même pas de terme équivalent dans la langue arabe. Malgré les nombreuses tentatives d'occidentalisation, les idées sous-jacentes à cette notion de citoyenneté sont bien différentes, aujourd'hui encore, de la vision des droits de l'homme, implicite en France lorsqu'on parle de citoyenneté. Dans le monde oriental, un individu demeure le sujet d'un souverain et, bien avant 1798, on utilisait pour parler d'un peuple le terme رَعَايَا (*ra'āyā*) : « sujets », « ouailles ». Si un citoyen chante *La Marseillaise*, et a pour valeurs le droit, la justice, le respect de la loi, un sujet oriental se doit d'abord de rester loyal envers son souverain qui, du coup, le prend sous sa protection[96]. Ainsi, en accordant la citoyenneté au robot Sophia, le roi d'Arabie Saoudite lui a promis sa protection, mais il a aussi exigé que Sophia se comporte conformément aux déclarations de son concepteur. En particulier, elle se doit de rester loyale à l'égard du « père de famille » et ne doit jamais trahir ses intérêts.

Cependant, la loyauté, en tant que valeur de l'intelligence artificielle, est un lieu commun que l'on retrouve dans tous les textes au niveau international[97] et, pour commencer, en Occident[98]. Par exemple, le constructeur automobile Mercedes-Benz a officiellement déclaré, en 2016, que les voitures autonomes fabriquées dans ses usines, quand elles seront confrontées au dilemme du tramway, préserveront en priorité la vie et la bonne santé des passagers présents dans leur habitacle[99]. Ainsi Mercedes élève la loyauté du système à l'égard de son utilisateur à la première place dans la hiérarchie des valeurs, tout comme, en Arabie Saoudite, on apprécie plus la loyauté des sujets que l'argent, et la vie elle-même.

Quelles que soient les valeurs recherchées, l'histoire d'Achan suggère que fabriquer une machine qui les respecterait toutes et qui serait bonne par conception n'est pas réaliste. Aucune action humaine n'est parfaite : toute technique mène au bien comme au mal. Une autre solution paraît plus envisageable : concevoir un système informatique qui se positionnerait par-delà le bien et le mal. Or, cette tâche exige, dans certains contextes, que l'on mette en valeur l'opacité, le contraire de la transparence ; elle exige aussi que la conception repose, non sur la loyauté, mais sur la confiance.

Quand Dieu demande à Josué : « Mais suis-je délateur ? », sa question demeure sans réponse. Néanmoins, dans un conflit qui tourne à la violence, la morale humaine exige qu'il existe toujours un coupable. Ce qui est difficile, c'est de le trouver. Le tirage au sort et l'utilisation du hasard se présentent, par homologie, comme unique alternative à la délation.

De manière générale, la méthode de recherche d'une vérité ne doit rien changer à la culpabilité de celui qu'elle permet d'identifier. Un coupable ne l'est pas parce qu'il a été identifié comme

---

[96] W. HANLEY, « When did Egyptians Stop Being Ottomans? An Imperial Citizenship Case Study », *Multilevel Citizenship*, Willem Maas (dir.), Philadelphie, University of Pennsylvania Press, 2013, pp. 89-109.

[97] Unesco, Rapport de la Comest sur l'éthique de la robotique, 2017.

[98] « Ethically Aligned Design », *op. cit.*

[99] « Self-Driving Mercedes-Benzes Will Prioritize Occupant Safety over Pedestrians », *Car and Driver*, 7 octobre 2016.

tel ; il l'est parce qu'il a commis un crime. Logiquement, la valeur de vérité de l'énoncé « Achan est coupable » ne doit pas varier, quel que soit le chemin qu'aura emprunté Josué pour identifier Achan, qu'il se soit servi de la parole de Dieu ou du tirage au sort. Tout cela semble évident. Cependant, quand le sort désigne Achan comme coupable, Josué lui demande d'avouer.

D'abord, Achan refuse. Il ne comprend visiblement pas et répond à Josué par une question : Que se serait-il passé si le sort était tombé sur quelqu'un d'important ? Plein de confiance en la force dissuasive de son argument contrefactuel, Achan croit avoir sauvé sa peau. Cependant, Josué renchérit.

Il réitère sa demande d'aveu en prononçant des paroles si convaincantes qu'Achan fléchit, et il est immédiatement brûlé sur le bûcher. Josué l'avait exhorté à ne faire planer aucun doute sur la procédure qui servait à établir la vérité, ajoutant que c'était par cette même procédure que la Terre promise serait partagée entre les tribus d'Israël. L'aveu d'Achan ne sert donc qu'à valider la procédure de recherche de la vérité. Une fois cette validation acquise, le tirage au sort est élevé au rang de méthode universelle servant à faire apparaître une vérité voilée.

Les mêmes événements se prêtent aussi à une interprétation différente. On pourrait dire que celui que les dés désignent acquiert le statut de coupable au moment du tirage au sort. La seule assurance de sa culpabilité réside dans la bonne exécution de la procédure. Logiquement, la vérité n'est rien d'autre que ce qui est établi d'après une procédure digne de confiance.

Répétons-le, la valeur de vérité de l'accusation portée contre Achan ne dépend pas, et ne doit pas dépendre, de la méthode de sa détermination (tirage au sort ou information recueillie auprès de Dieu). Si Josué avait réussi à obtenir directement un renseignement, Dieu lui aurait indiqué qu'Achan, et pas un autre, était le coupable. Cette certitude, nous la devons à l'assurance que nous procure le premier verset du chapitre 7 : avant même le début de la campagne d'Haï, l'auteur nous informe qu'Achan a violé l'interdit. L'énoncé antéchronologique vise précisément à bien asseoir la confiance en la méthode employée.

D'un côté donc, l'accès indirect à la connaissance du coupable ; de l'autre, sa validation par un aveu : pris conjointement, ces deux facteurs procurent la garantie qu'aucun soupçon ne sera jeté sur la méthode. Cette assurance permet au lecteur de croire que l'information qu'il obtient suite au tirage au sort ne diffère en rien d'un renseignement que Dieu aurait pu fournir directement. Quelle que soit la source d'une information, la preuve de sa vérité ne réside que dans la vérification du bon déroulement de la procédure qui a été suivie pour l'obtenir.

Une fois la méthode validée, toute information qu'elle a permis, et permettra, d'obtenir est perçue comme véridique. Le peuple d'Israël emploie cette méthode pour partager les terres. L'utilisateur d'un système informatique, parce qu'il est lui aussi confiant, ne diffère guère d'eux. La machine lui communique une information qu'il transforme en connaissance : ce procédé requiert justement qu'il fasse confiance à la procédure. C'est une condition *sine qua non* pour qu'il soit à même de tirer profit de sa nouvelle connaissance.

Encore faut-il qu'il continue de faire confiance : ce problème essentiel peut être résolu grâce à une subtile manipulation de la frontière de l'opacité qui caractérise l'individu numérique. Contrairement aux idées reçues, ce n'est pas la transparence d'un système informatique qui

permet d'établir la confiance de l'utilisateur en ce système, puisque l'utilisateur ne connaît pas, et ne peut pas connaître, son fonctionnement interne. Au mieux, la transparence devient un tranquillisant d'ordre psychologique et politique, plutôt qu'un critère technique : l'utilisateur, démuni comme il l'est de toute connaissance quant à ce qui se passe derrière la frontière de l'opacité, croit néanmoins qu'il existe un expert capable de comprendre l'ensemble des communications du système informatique et de vérifier son comportement. C'est exactement ce que dit la compagnie Microsoft : « L'approche qui a le plus de chances d'engendrer la confiance des utilisateurs et de ceux qui se trouvent affectés par ces systèmes consiste à fournir des explications incluant l'information contextuelle sur la façon de fonctionner d'un système d'intelligence artificielle ainsi que sur sa façon d'interagir avec les données[100]. » Mais si telle est l'origine de la confiance, il faut bien admettre qu'elle est liée à l'opacité que perçoit l'utilisateur et que, sans cette opacité caractéristique de toute interaction avec un individu numérique, aucune confiance ne serait rendue possible.

On lit souvent ce slogan : « Toujours plus de transparence ! » La confiance de l'utilisateur, enjeu central de l'actuelle politique en matière de numérique, provient de l'application d'une procédure digne de foi, et non d'une connaissance que l'utilisateur ne possède d'ailleurs pas. L'opacité et le manque de connaissance jouent ici un rôle clef.

Une homologie permet de mieux saisir ce paradoxe. Un autre épisode de la *Vie d'Apollonius de Tyane*, de Philostrate, souligne le caractère central de la confiance dans les rapports entre le bien, le hasard et le mal. Apollonius se trouve encore en prison quand Philostrate s'autorise une digression sur sa libération magique des entraves qu'on lui avait mises aux pieds :

> « Les hommes simples attribuent à la magie ces faits merveilleux, et ils font de même pour beaucoup de faits qui sont purement humains. Ainsi les athlètes et les divers lutteurs ont recours à la magie, dévorés qu'ils sont par le désir de la victoire. Pourtant la magie ne leur sert en rien pour remporter le prix, mais si par hasard [ἀπὸ τύχης] ils viennent à être vainqueurs, aussitôt ces malheureux, en se dérobant eux-mêmes, rapportent tout à cet art, et ceux qui ont été vaincus ne lui font pas confiance moins fermement[101]… »

Le sens de ce passage n'est pas simple à saisir ; là aussi, il est nécessaire de se livrer à une interprétation du mythe.

À quoi les vainqueurs se dérobent-ils ? Ils attribuent leur victoire, non à eux-mêmes mais à la magie, en refusant d'accepter la victoire pour ce qu'elle est. Ils souhaitent lui trouver une explication causale, comme si elle ne pouvait être due qu'à une force extérieure encore inconnue. Philostrate s'en moque : les vainqueurs gagnent en fait par hasard.

Pourquoi dit-on que les vainqueurs sont « malheureux » ? Le sens du mot κακοδαίμονες ne se réduit pas au terme « malheureux » ou à ses équivalents dans les langues modernes. Ce « malheureux » qu'emploie Philostrate est particulier : Nietzsche, par exemple, ne le cite même pas lorsqu'il insère, dans *La Généalogie de la morale*, un catalogue très complet des mots grecs

[100] MICROSOFT, *The Future Computed : Artificial Intelligence and Its Role in Society*, 2018, p. 75.

[101] PHILOSTRATE D'ATHENES, *Vie d'Apollonius de Tyane*, VII 39 (traduction Chastaing modifiée par nos soins).

dont la signification est proche de « misérable » ou d'« infortuné »[102]. La raison de son omission se trouve peut-être dans le fait qu'il s'agit, non d'un lexème original, mais d'un mot complexe, antonyme d'ευδαίμονες. Ce dernier terme est souvent utilisé dans la littérature et signifie « bienheureux » ou « possédé par un bon démon ». Le κακοδαίμονες serait alors quelqu'un qui, littéralement, est habité par un mauvais démon. Notons au passage que le sens du mot « bonheur » ne correspond pas, non plus, à l'idée moderne qu'on s'en fait : il s'agit d'un jugement qui porte sur l'intime d'une personne, sur son « fonds », et aussi sur la puissance qui l'anime, ce qu'on appelle un « démon ».

Les vainqueurs « se dérobent eux-mêmes » : ils ne dérobent pas, par exemple, à Tyché, la déesse de la fortune. Si les vainqueurs n'avaient pas attribué leur victoire à la magie, ils auraient dû, d'après le texte, se l'octroyer, puisqu'il s'agit de « faits purement humains ».

Que signifie, dans ce contexte, « attribuer à soi-même » ; plus exactement, quel est le sens de « soi » ?

Philostrate ne dit des athlètes qu'une seule chose : ils sont habités par un mauvais démon. Ce démon représente, en quelque sorte, la force du mal. Le mal en question n'est évidemment pas un principe métaphysique, mais une simple conséquence du fait que Philostrate n'apprécie ni les sportifs ni tous ceux que dévore l'esprit de compétition. Ces « gens simples », Philostrate les réduit à une fonction, celle du désir aveugle de gagner, et il les condamne : ce trait appartient au mauvais démon, que la victoire ait lieu ou pas.

Philostrate précise son jugement. Il refuse d'admettre que le mauvais démon, dans le fonds intime des « gens simples », puisse être désigné comme responsable de leur victoire. À sa place, il désigne Tyché, l'incarnation divine du hasard. Selon Philostrate, donc, on parvient à échapper au contrôle du mauvais démon, et d'une certaine forme du mal, à condition de faire appel au hasard. Tyché, la déesse de la bonne fortune, était très appréciée dans l'Empire romain tardif, à Rome comme à Constantinople, Alexandrie ou encore Antioche.

Ce qui importe réellement dans l'argument de Philostrate n'est pas la tendance des « gens simples » à expliquer leurs victoires par tel ou tel sacrifice, reçu avec bienveillance par tel ou tel dieu, ou par un recours à la magie. C'est la confiance accordée à la procédure. Même une chose fortuite, sans explication causale, sans l'intervention d'une force extérieure, est interprétée de sorte qu'elle paraisse fiable. Les gagnants et les perdants lui font pareillement confiance. Cela leur procure l'assurance d'avoir dévoilé la vérité. Et à Philostrate de conclure : « Voilà ce qu'ils disent, voilà ce qu'ils pensent[103]. »

L'utilisateur d'un système informatique n'est pas différent de ces « gens simples ». Lui aussi fait confiance à la procédure qu'emploie la machine et tient pour vraie toute information que celle-ci lui communique. Derrière cette homologie se trouve le motif du secret. Les utilisateurs et les sportifs ne possèdent pas toute la connaissance : les seconds ne savent pas pourquoi ils gagnent et les premiers sont confrontés à la frontière de l'opacité. Là où il y a manque de connaissance, Philostrate dit qu'un homme « se dérobe », en cherchant la source de

[102] F. NIETZSCHE, *Éléments pour la Généalogie de la morale*, Librairie générale française, 2000, p. 84.

[103] PHILOSTRATE, *op. cit.*

l'information ailleurs, tandis qu'elle se trouve en lui-même. Ainsi, par homologie, on peut dire qu'un utilisateur mis en concurrence avec quelqu'un d'autre peut bien se tirer de cette compétition ou de ce conflit grâce au hasard. Loin d'être fondée sur un raisonnement rationnel ou sur un calcul, cette solution repose fondamentalement sur la confiance.

Un automate autodidacte

Nous l'avons déjà dit, la différence entre « délateur » et « rapporteur » s'estompe facilement. Tantôt les traductions en langues modernes ajoutent un jugement éthique explicite (« délateur »), tantôt elles réduisent le sens à la seule fonction de communication de l'information (« rapporteur »).

Une différence tout aussi subtile existe dans la langue grecque entre deux autres termes. Dans une de ses œuvres, Philon d'Alexandrie mentionne Melchisédech, personnage mystérieux, grand prêtre, qui était, selon les propos de saint Paul, « semblable à un fils de Dieu » (Hb 7, 3). Un manuscrit de la mer Morte le désigne même comme « fils de Dieu », et aussi un des אֱלֹהִים — *'ĕlōhîm* (« dieux » ou « anges »). Ces formules le confirment : Melchisédech joue un rôle fonctionnel.

Melchisédech apparaît dans de très nombreux textes de la littérature hébraïque tardive. En interprétant ces mythes, Philon, qui écrivait en grec, lui adjoint l'épithète αὐτομαθής, habituellement traduit par « auto-instruit » ou « autodidacte ». Or, « autodidacte », autre terme d'origine grecque, vient dans le texte de Philon juste après la curieuse épithète affectée à Melchisédech : αὐτομαθῆ και αὐτοδίδακτον, *automathe* et *autodidacte*. Que peut bien signifier le premier, si l'on admet que son sens doit être différent, ne serait-ce que légèrement, de celui du second ?

Le mot αὐτομαθής est présent dans la littérature grecque en dehors de tout contexte biblique et bien avant Philon. À son époque, il apparaît aussi dans le livre *Des poèmes* de Philodème de Gadara, prolifique philosophe épicurien. Comme beaucoup d'intellectuels d'alors, Philodème a commencé sa carrière en Grèce avant de la terminer en Italie où il se retira à Herculanum quelques années avant l'éruption du Vésuve, en l'an 79. Sa bibliothèque fut donc ensevelie et ainsi préservée pour la postérité. On a réussi à ouvrir et à déchiffrer bon nombre de ces rouleaux pétrifiés et fossilisés. La bibliothèque de Philodème est unique en son genre car, presque entièrement conservée, elle n'a pas été soumise à la sélection des moines chrétiens qui, du fond de leurs *scriptoria*, recopiaient seuls les ouvrages qui pouvaient encore les intéresser.

Dans un fragment du livre *Des poèmes*, Philodème s'interroge sur la relation qui lie le poète aux Muses : « Comment cette parenté peut-elle naître ? Comment débute-t-elle, comment se développe-t-elle ? » Il répond lui-même à cette question : la parenté du poète avec les Muses a la propriété d'être αὐτομαθής[104]. Les traducteurs ont rendu ce mot par « spontanée » ou « instinctive ». Dans le même registre, un traducteur des textes relatifs à la figure de Melchisédech applique à ce personnage le terme « intuitif », lorsqu'il rapporte le récit de

[104] PHILODEME DE GADARA, *Des poèmes*, 2.47.

Philon[105]. Le mot « automathe » contient donc, lui aussi, un aspect de spontanéité. Autrement dit, il s'agit d'un lien entre le poète et les Muses tel que sa cause et la force qui le suscite émanent du for intérieur du poète, de son fonds intime. Cette spontanéité instinctive ne provient ni d'une réflexion rationnelle ni d'une source d'origine extérieure.

Curieuse coïncidence, si c'en est une : ce même bouquet de significations se retrouve dans un autre mot grec, αὐτόματος. Les morphèmes αὐτοματ- et αὐτομαθ- ne diffèrent que d'une lettre. À l'oral, cette différence n'est même pas audible en français ou en italien : « automathe » et « automate » se prononcent de la même façon.

Le mot « automate » fait immédiatement penser à l'apprentissage de la machine, qu'on appelle aussi « apprentissage automatique[106] ». Lorsque les informaticiens créent des systèmes capables d'apprentissage et d'autonomie, qui analysent d'énormes masses de données pour en tirer des corrélations, ils disent que ces machines, en langue allemande, sont des *selbstlernende Systeme*, « des systèmes autodidactes »[107]. Le motif d'auto-instruction ou d'auto-apprentissage de la machine est omniprésent dans le vocabulaire technique.

« Automate » signifie d'abord, chez Aristote, un phénomène qui advient par soi-même, sans cause finale, sans but. Dans les langues modernes, ce terme philosophique est souvent traduit par « aléatoire », « fortuit », « chanceux », « spontané » ; il est utilisé pour parler d'une coïncidence ou d'un hasard heureux.

Tout aussi facilement que celle entre « délateur » et « rapporteur », la différence entre « automate » et « automathe » tend donc à disparaître. Le second terme suggère la spontanéité. Quant au premier, il est employé à l'envi en informatique. Nous serions tentés de conclure, à examiner ce rapprochement étymologique, que le hasard serait amené à jouer un rôle prépondérant dans l'apprentissage de la machine.

Calculer à rebours de la flèche du temps

Revenons à la délicate boucle causale dans le processus de recherche d'une vérité voilée. Cette vérité ne préexiste pas à son établissement au moment du tirage au sort, pourtant ce dernier se présente comme une révélation. Certains objecteront qu'il existe une différence fondamentale entre le cas d'Achan et celui d'un système automatique apprenant qui communique avec l'homme. Ils diront que la manipulation dont fait montre l'auteur du livre de Josué est illogique et illégitime, parce que, lorsqu'il nous donne le nom du coupable au début du chapitre 7, Josué n'a pas encore appris la défaite de l'armée d'Israël à Haï. Il est bien trop tôt : dans le récit, les protagonistes ne découvrent la culpabilité d'Achan qu'à la suite d'un tirage au sort qui intervient, bien évidemment, après la bataille d'Haï. Et, même après avoir identifié le coupable,

[105] I. TANTLEVSKI, *Melchisédech et Métatron dans la tradition hébraïque mystico-apocalyptique*. Presses universitaires de Saint-Pétersbourg, 2007, p. 11 (en russe).

[106] Cf. le nom officiel de l'institut public français de recherche en numérique : Inria, Institut national de recherche en informatique et en *automatique*.

[107] *Ethik-Kommission Automatisiertes und Vernetztes Fahren,* sur la demande du *Bundesminister für Verkehr und digitale Infrastruktur*. Rapport disponible à l'adresse www.bmvi.de (juin 2017).

Josué exige d'Achan un aveu, comme s'il voulait prouver à tous, une nouvelle fois, le bien-fondé de cette identification.

Dès le premier verset du chapitre 7, le lecteur est donc confronté à des faits non avérés. Leur vérité n'existe pas encore dans la temporalité vécue par le peuple d'Israël ; elle sera établie par un coup de dés puis inscrite dans le passé par propagation antéchronologique. Le propre de cette procédure exige, comme nous l'avons déjà dit, que le peuple lui fasse confiance.

Pourquoi le résultat du tirage au sort est-il rétro-propagé ?

Il est incontestable qu'Achan ne devient coupable qu'au moment où il est identifié comme tel par les dés. Avant qu'ils ne soient jetés, un spectateur inscrit dans le temps unidirectionnel de l'histoire aurait encore pu affirmer que le sort eût pu tomber sur quelqu'un d'autre. Achan essaie même d'impressionner Josué par cet argument ; lequel ne vaut plus à partir de l'instant où les dés sont jetés. Désormais, Achan est le coupable ; qui plus est, sa culpabilité ne s'applique pas qu'à l'avenir : les participants interprètent spontanément la situation de façon à ne pas tenir compte de cette *création* de la vérité à un instant donné. Tout se passe pour eux comme si Achan avait toujours été coupable, y compris avant le tirage au sort. Lorsque retentit l'annonce « C'est Achan ! », elle est aussitôt perçue comme une vérité *révélée*, dont l'existence ne dépend nullement de la connaissance qu'on a d'elle. C'est précisément la raison pour laquelle le lecteur n'est pas étonné d'être informé de la faute d'Achan au début du chapitre 7.

Un phénomène qui, dans le langage du mythe, est exprimé par la locution « révéler un choix divin », se dit différemment à un autre niveau d'interprétation, dans le langage philosophique. Il s'agit là d'une prise de décision telle que tous les critères éthiques les plus hauts y sont réunis : vérité, confiance, justice.

L'objection que l'on pourrait opposer contre l'homologie entre mythe et informatique est celle-ci : dans le récit, l'établissement d'une vérité ne provient pas d'un raisonnement causal, car la causalité matérielle ne se propage pas du futur (le moment du tirage au sort) vers le passé (le moment de la défaite). Or, une machine ne saurait violer les relations causales. Elle fonctionne en calculant du passé vers l'avenir, d'une cause vers son effet. Elle ne communique à l'homme qu'une information qui contient déjà, de façon objective, une certaine causalité. Contrairement à ce qui se passe dans le mythe, la connaissance n'est pas créée au moment de l'interaction de la machine avec l'utilisateur.

Cette objection est erronée. Elle s'appuie sur une croyance qui n'est pas fondée, à savoir que la machine détermine des relations de causalité, alors qu'elle ne découvre, en analysant les données, que des corrélations. Celles-ci sont de simples indications, parfois fausses, d'une causalité possible et n'en fournissent aucunement une preuve. La machine apprenante ne donne que des indications : elle traite l'information asémantique à l'aide de signes formels. Et il n'est pas rare que la corrélation entre ces signes s'avère franchement risible à vue humaine. Par exemple, une machine employée dans un hôpital découvre que la probabilité qu'un patient développe des complications à la suite d'une pneumonie « dépend » du fait qu'il souffre, ou pas, d'asthme[108]. Cette machine pourrait tout aussi bien informer le personnel que ladite

[108] A. BORNSTEIN, « Is Artificial Intelligence Permanently Inscrutable ? », *Nautilus*, 1er septembre 2016.

probabilité dépend du nombre de sourires qu'esquisse le patient entre 8 heures et 9 heures. Le système informatique a peut-être raison d'établir une corrélation entre ces paramètres d'après les données qui lui avaient été fournies, mais l'interprétation de cette corrélation en tant que dépendance causale, que l'utilisateur lui attribue d'ailleurs spontanément, est à l'évidence fausse.

Une machine, contrairement à l'homme, n'a pas (encore ?) la conscience nécessaire pour distinguer les causes des corrélations. Cela ne signifie pas pour autant qu'elle soit inutile. Son utilisateur ne peut pas « voir » au-delà de la frontière de l'opacité. Il établit une connaissance seulement au moment où son cerveau traduit dans sa langue naturelle l'information que la machine lui transmet ; l'information est alors automatiquement chargée d'un sens et d'une valeur de vérité. Ce procédé joue un rôle prépondérant dans l'établissement des chaînes de causes à effets qui relient les faits que la machine met au jour.

Dépourvu de toute connaissance des processus internes d'un système informatique, l'utilisateur s'intéresse instinctivement au monde extérieur, y compris au monde obscur tapi derrière la frontière de l'opacité de l'individu numérique. Il suppose spontanément qu'une procédure digne de confiance est exécutée dans le fonds intime de la machine, et que celle-ci lui fournira des informations utiles à travers son interface. L'adhésion de l'utilisateur au sens et à la sémantique des résultats communiqués s'appuie sur la confiance et est impossible sans elle. C'est grâce à cette imbrication entre opacité, confiance et vérité que le schéma du fonctionnement d'un système informatique ressemble à celui du tirage au sort dans l'histoire d'Achan.

L'information à propos de la culpabilité d'Achan est donnée dans le mythe au moment du tirage au sort ; elle est ensuite rétro-propagée. Un fait devient réel à l'utilisateur d'un système informatique à l'instant où il lui est communiqué par la machine et où il est interprété par son cerveau. Le traitement sémantique dans le langage humain cherche alors à établir une chaîne causale en reliant les différents faits, comme s'ils avaient toujours existé et ne venaient pas tout juste d'apparaître. Même lorsque ces faits sont formellement faux, l'homme raisonne avec des liens de cause à effet : au pire, il suppose que ces liens n'avaient été que voilés. Pour l'utilisateur, il n'y a pas de différence entre des causes absentes, des causes inconnues ou des causes cachées derrière la frontière de l'opacité. Il est sûr que les liens causaux existent objectivement ; cependant, c'est lui qui les a inférés par la rétro-propagation de l'information.

Tout se passe comme si la faute d'Achan était avérée depuis la fondation du monde. Cette vérité dépend entièrement de la confiance que le peuple d'Israël accordait au tirage au sort. En informatique, c'est la même confiance que l'homme octroie à la machine.

Délivrer l'intelligence artificielle

La valeur éthique du hasard se manifestera-t-elle bientôt, ou se manifeste-t-elle déjà, dans les exemples concrets ?

Assurément. En premier lieu, dans les dilemmes éthiques, ceux-là mêmes qu'un système informatique est censé « résoudre ». Rappelons qu'une voiture autonome doit décider seule de chacune de ses manœuvres, y compris dans des situations délicates. Bien évidemment, les

situations dans lesquelles la décision à prendre n'est pas évidente sont moins nombreuses que celles qui correspondent à une conduite normale ; on estime la proportion des dilemmes à moins de 1 %[109]. Mais ce sont précisément ces cas rares qui posent problème à la conception éthique des algorithmes.

Normalement, un conducteur, tout comme une voiture autonome, applique des règles. Lorsqu'il doit faire un choix, l'homme se les rappelle, tandis qu'un système informatique puise dans les règles encodées par le programmeur. Un processus d'apprentissage est également à l'œuvre : un conducteur s'améliore au fil des années, tandis qu'une machine apprend en analysant le comportement des conducteurs ainsi que ses propres performances[110]. Cependant, les dilemmes éthiques étant rares, il est vraisemblable qu'un système informatique apprenant ne possédera aucune donnée antérieure pour décider en procédant à l'évaluation des options précédemment choisies. Contrairement à l'homme qui réagit dans ces cas instinctivement, la machine ne peut qu'user du premier procédé : elle applique des règles. On fait donc appel à des philosophes pour tenter de formuler ces règles, établir une hiérarchie des valeurs, et définir leur importance respective. En éthique de l'intelligence artificielle, la hiérarchisation des valeurs est au centre des recherches.

Chris Gerdes, professeur de mécanique à l'Université de Stanford en Californie, a fini par mesurer l'importance de l'éthique, mais aussi sa « pénible » complexité, après une longue période semée de doutes, ce qui est, par ailleurs, assez fréquent chez un ingénieur confronté à une discipline « molle ». Lorsqu'un accident ne peut être évité, se demanda-t-il, une voiture doit-elle rentrer dans un objet de petite taille afin de protéger ses passagers ? Et si cet objet est une poussette ? « Il est important de considérer, non seulement la manière dont ces voitures sans pilote vont se conduire elles-mêmes, mais aussi ce que sera l'expérience de quelqu'un qui se trouve à l'intérieur et interagit avec elles. La technologie et l'humain doivent être traités comme réellement inséparables[111]. »

Dans le dilemme du tramway, l'homme à qui on demande de faire un choix est stupéfait et paralysé par la nécessité de sélectionner des victimes destinées à mourir. La valeur de la vie est absolue : entraîner une seule mort n'est pas *mieux* que d'en entraîner cinq. L'exigence d'établir une hiérarchie nette, rationnelle, algorithmique, dépourvue de toute ambiguïté, quant aux valeurs que l'homme lui-même refuse de ranger dans un ordre de priorité, est par définition inhumaine. Formaliser l'indicible, quantifier le terrible, sont des tâches qui ne sont dignes que d'un monstre moral.

On peut cependant ne pas mettre le programmeur dans l'obligation de faire un choix éthique impossible. En recourant au hasard dans le dilemme du tramway, un système informatique peut échapper à l'impératif de hiérarchiser les valeurs éthiques absolues ou de formaliser les règles de la mortalité routière. Ce hasard peut prendre deux formes. La première est celle d'un tirage

---

[109] « Self-Driving Mercedes-Benzes Will Prioritize Occupant Safety over Pedestrians », *Car and Driver*, 7 octobre 2016.

[110] Un modèle de voiture autonome en cours de développement utilise un algorithme qui consiste uniquement en un apprentissage, sans qu'aucune règle n'y soit encodée au préalable (W. KNIGHT, « The Dark Secret at the Heart of AI », *MIT Technology Review*, 11 avril 2017).

[111] « Stanford Professor's Quest to Fix Driverless Cars' Major Flaw », *Bloomberg*, 8 octobre 2015.

au sort explicite. La seconde est moins évidente : la machine peut obscurcir son choix en le gardant voilé derrière la frontière de l'opacité. Ainsi, l'utilisateur humain le percevra-t-il comme opaque dans le bon sens du terme, une solution non transparente à laquelle il ne peut que se résoudre. En parlant des stratégies dans la théorie des jeux, Ivar Ekeland écrit : « Pour qu'une stratégie soit vraiment impénétrable, et reste imprévisible de partie en partie, le mieux est qu'elle soit imprévisible pour celui qui l'applique, et donc qu'elle soit aléatoire[112]. »

Comme dans le mythe d'Achan, le recours au hasard en tant qu'instrument pour résoudre des dilemmes éthiques ne doit aucunement être perçu par l'utilisateur comme une intervention visant à modifier un résultat préexistant. Le recours au hasard évite toute « réécriture » d'une solution que Dieu aurait pu donner directement mais qu'il ne donne pas. Cela permet, nous l'avons vu, de maintenir la confiance dans la procédure. En informatique, le programmeur doit s'assurer que l'utilisateur ne pourra en aucun cas faire appel à une connaissance préexistante ; qu'il n'aura d'autre choix que de faire confiance à la procédure, celle-ci étant, de son point de vue, aléatoire.

Plusieurs systèmes informatiques apprenants recourent déjà à ce second type de hasard. Il s'agit de systèmes à multiples couches d'apprentissage, parfois même tels qu'ils sont capables d'optimiser leur structure de couches en décidant, par un calcul interne, des modifications à apporter à leur propre architecture[113]. Comme si un mille-feuille était son propre pâtissier…

En se complexifiant, les systèmes apprenants rendent opaques, du point de vue de l'utilisateur, les procédés internes de leur fonctionnement ; qui plus est, et c'est vraiment exceptionnel dans tout le domaine de la conception des algorithmes, une procédure de décision de ce type est également totalement non transparente du point de vue du concepteur. Or, c'est souvent la volonté de ce dernier, puisque le manque de transparence dans le fonctionnement d'un algorithme d'apprentissage et son efficacité vont de pair. Ce manque de transparence n'est donc pas un défaut du point de vue du programmeur : c'est une propriété essentielle qu'il fabrique, même s'il ne la recherche pas intentionnellement. Elle est de son point de vue bonne car indicative de la bonne performance du système.

Sous le nom d'« apprentissage profond » (*deep learning*), on regroupe une famille de techniques d'apprentissage de la machine. Ces algorithmes superposent plusieurs couches d'apprentissage. Si on dessine leurs entrées et leurs sorties pour en faire un diagramme, on voit qu'elles sont reliées par une multitude d'attaches qui vont en tous sens et peuvent même s'entrecroiser. Comme si, dans un mille-feuille, la crème coulait entre les feuilles dans toutes les directions tout en renforçant le goût du gâteau ; ou comme, après une pluie, les petits ruisseaux s'ouvrent des myriades de chemins sur une terre desséchée qu'ils vont bientôt transformer en fertile humus.

Lors d'un apprentissage profond, chaque couche réalise l'une de ces trois méthodes générales : apprentissage dit supervisé, apprentissage non supervisé, ou apprentissage par renforcement. Chacune de ces méthodes s'y prête déjà individuellement, mais leur imbrication contribue

[112] I. EKELAND, *Au hasard. La chance, la science et le monde*, Paris, Éditions du Seuil, 1991, p. 92.

[113] Q. LE et B. ZOPH, *Using Machine Learning to Explore Neural Network Architecture*, Google Research Blog, 17 mai 2017.

davantage à rendre inconcevable, au moins à ce jour, toute description mathématique rigoureuse de ce qui se passe dans une machine pendant son apprentissage. Un scientifique ne peut ni l'expliquer analytiquement ni prédire ce qui aura lieu : il ne peut que regarder le système agir. C'est là l'opacité maximale, qui ne peut, *par conception*, devenir transparente.

Du coup, il est impossible d'établir rationnellement un lien de cause à effet entre les données de départ et le résultat final. Non seulement l'utilisateur, mais aussi le concepteur de l'algorithme ainsi que son entraîneur (celui qui fournit les données pour l'apprentissage), ne sont pas en mesure d'expliquer pourquoi la machine ayant effectué un apprentissage a pris telle ou telle décision (cet effet apparaît dans certaines applications des trois méthodes d'apprentissage). Il est impossible de la « comprendre » ; dans le meilleur des cas, l'homme peut s'en faire une idée en visualisant les différents éléments du processus de prise de décision automatique. Il espère que son cerveau, cette autre « machine » à apprendre, y trouvera un sens. Une solution alternative qui vise à établir un lien entre les entrées et les sorties d'un algorithme consiste à faire fonctionner un autre système apprenant censé le « comprendre », à la place de l'homme et usant pour cela du langage humain. Parfois cela satisfait l'utilisateur, parfois non. Mais, quels que soient les bénéfices de ces approches, force est de constater l'impossibilité d'expliquer le comportement d'un système informatique en s'appuyant sur des règles ou des lois.

Cette opacité fondamentale des systèmes informatiques est souvent désignée par les termes « non-explicabilité » ou « non-interprétabilité ». Elle fait (encore) débat : la non-explicabilité est-elle bonne ou mauvaise ? Dans le domaine médical, par exemple, on développe des logiciels de traitement automatique des images, par ailleurs extrêmement efficaces, censés aider à poser un diagnostic à partir d'un scanner. L'impossibilité d'interpréter un résultat qu'un tel programme communiquerait au médecin ne serait certainement pas une propriété désirable. Aucun concepteur ne peut permettre que son logiciel indique à l'utilisateur l'existence d'un problème de santé sans lui dire d'où il tire cette conclusion. Un algorithme qui ne fournit pas d'explications a donc peu de chances d'être utilisé dans le domaine médical. Imaginez quel scandale ce serait si un « médecin du futur » disait à son patient : « Vous êtes malade, *parce que* c'est la machine qui le dit. »

La non-explicabilité est, en revanche, bienvenue dans les jeux. On n'exige pas d'une machine qui joue aux échecs ou au go qu'elle nous explique pourquoi elle a fait tel ou tel coup : son efficacité se mesure à sa victoire ou à sa défaite. Pour gagner contre un grand maître, elle a besoin de faire un saut qualitatif. Si un algorithme fait des « coups de Dieu[114] », donc inexplicables, jamais vus, mais qu'il finit par gagner, on l'en félicite et ses coups entreront dans les manuels d'apprentissage pour les joueurs humains.

Voici un autre exemple : l'algorithme de reconnaissance automatique des visages ne doit pas nécessairement donner des significations humaines à chaque trait qu'il analyse pour identifier des personnes. Ne compte que le résultat final : identification réussie ou ratée. Si l'algorithme

[114] Cf. les performances de l'algorithme AlphaGo, développé par la société DeepMind, filiale de Google : E. GIBNEY, « Google Reveals Secret Test of AI Bot to Beat Top Go Players », *Nature*, 4 janvier 2017. L'expression « coup de Dieu » est de Ke Jie, champion chinois de go : « World's Best Go Player Flummoxed by Google's "Godlike" AlphaGo AI », *The Guardian*, 23 mai 2017.

de reconnaissance a un taux d'erreurs suffisamment bas, il sera mis au service de la police entre autres, sans qu'aucune question soit posée quant au caractère explicable ou pas de ce qu'il fait. Aujourd'hui, les algorithmes qui réalisent cette tâche s'appuient sur environ quatre-vingts paramètres ; ceux-ci ne sont pas encodés par le programmeur, mais émergent automatiquement lors du processus d'apprentissage. Les concepteurs eux-mêmes avouent leur étonnement en découvrant ces quatre-vingts paramètres — censés correspondre à autant de traits faciaux —, et qu'ils sont incapables de désigner par un mot du langage humain[115]. Nous pouvons les visualiser, mais nous ne pouvons pas les dire par nous-mêmes : nous ne savons pas ce qu'ils signifient.

Le problème que pose la non-explicabilité dans l'apprentissage profond est aujourd'hui largement reconnu. La compagnie Microsoft, ayant solennellement proclamé que « les systèmes d'intelligence artificielle doivent être compréhensibles », fait immédiatement marche arrière en constatant un obstacle majeur :

> « La simple publication des algorithmes sous-jacents aux systèmes d'intelligence artificielle ne peut que rarement aboutir à la transparence de manière significative. À l'aune des techniques les plus récentes et les plus prometteuses de l'intelligence artificielle, telles que l'apprentissage profond dans les réseaux de neurones, il n'existe typiquement aucune sortie algorithmique qui soit susceptible d'aider les humains à comprendre les subtils motifs que trouvent ces systèmes. C'est pourquoi nous avons besoin d'une approche plus holistique, qui inclurait la description des éléments clefs des systèmes d'intelligence artificielle par leurs concepteurs d'une manière la plus complète et la plus claire qui soit possible[116]. »

L'opacité est créée aussi par un autre phénomène que les informaticiens connaissent bien : le « comportement non défini » (*undefined behavior*) des programmes. Supposons qu'un programme ait besoin, au cours de son exécution, de calculer la valeur d'une fonction qui dépend de trois paramètres : $f(x, y, z)$. Pour cela, il lui est nécessaire, bien évidemment, de posséder les valeurs de ces paramètres, qu'il doit donc calculer au préalable. Mais dans quel ordre ? Le programme va-t-il d'abord calculer x, puis y, enfin z ? Ou d'abord y, puis x, enfin z ?

Souvent, les compilateurs du code source ne définissent pas l'ordre exact du calcul des arguments multiples des fonctions. Le programmeur ne sait pas, et ne peut pas savoir, comment le code exécutable se comportera à chaque instanciation concrète de cette tâche. Et cependant, il est clair qu'un ordre existe, car le code s'exécute progressivement, étape par étape. Le programmeur ne le connaît donc pas ; quant à l'utilisateur, non seulement il ne peut pas le connaître, mais il ne se pose même pas cette question : le calcul des fonctions définies dans le code, procédé interne non perceptible à travers l'interface, reste pour lui entièrement opaque. Il en résulte que, du point de vue de l'homme qu'il soit expert ou profane, l'ordre du calcul, qui n'est pas connu et ne peut pas être connu à l'avance, est perçu comme aléatoire. L'utilisateur

[115] « La signification de la plupart de ces paramètres reste inconnue même aux chercheurs » (Interview du responsable de la société Ntech, https://meduza.io/feature/2016/07/07/konets-chastnoy-zhizni).
[116] MICROSOFT, *The Future Computed : Artificial Intelligence and Its Role in Society*, 2018, p. 75.

croit que ce hasard n'est qu'une conséquence de son ignorance, mais il ignore que le programmeur aussi ignore l'ordre du calcul.

On retrouve dans cet exemple le rôle central que joue la frontière de l'opacité. Tout processus dont la cause est voilée est perçu par l'homme comme aléatoire ; pas nécessairement objectivement — « aléatoire pour tous » —, mais subjectivement — « aléatoire de mon point de vue ». Que l'ordre du calcul des paramètres d'une fonction soit déterminé par un générateur de nombres aléatoires, ou qu'il soit dû à un autre processus, même déterministe, tout cela ne fait aucune différence pour l'utilisateur, pourvu que ce processus lui demeure inconnu et inconnaissable.

En conclusion, si l'homme n'a, et ne peut avoir, aucune information à propos des liens de causalité sous-jacents à une décision prise par la machine, il la considère comme aléatoire, et ce, au même titre qu'une décision générée par un tirage au sort. Voilà qui revêt une importance éthique fondamentale. Et c'est là que l'on rencontre deux difficultés.

La première provient de ce constat qu'en informatique, l'aléatoire est produit artificiellement. Contrairement, par exemple, à ce qui se passe en biologie où, depuis au moins le livre de Jacques Monod[117], on débat de la valeur du hasard naturel dans la conception, le développement, la naissance et la mort d'un individu biologique. Les principes de la loi bioéthique, actuellement en vigueur en France, garantissent que l'homme, même s'il peut exercer un certain contrôle sur l'embryon, respectera, au moins partiellement, le rôle du hasard dans le choix de sa progéniture. Cette accentuation de la valeur du hasard dans les processus naturels ne soulève pas de grandes protestations au sein de la société française ; au contraire, elle est perçue avec bienveillance, et c'est aussi le cas dans la majorité des pays européens plus « bio-conservateurs » que les cultures asiatique ou californienne. Or, en informatique, le hasard n'est ni externalisé ni naturel ; il n'advient pas par lui-même. Le hasard numérique apparaît en tant que conséquence d'une complexité algorithmique créée par le programmeur. L'utilisateur peut donc aisément avoir l'impression que le concepteur d'un logiciel a volontairement choisi de cesser de le contrôler ou qu'il l'a voilé, tout en restant maître du jeu. Cela pousse l'opinion publique et les autorités politiques de nos sociétés technologiques à croire que le programmeur est maître et possesseur des systèmes informatiques, comme le dieu d'une religion est maître et possesseur du hasard externalisé et non humain.

L'utilisateur, en effet, a tendance à penser que le programmeur sait ce que fera la machine. Cependant, ce dernier n'en est pas si sûr ; or, son opinion est incapable de changer la croyance de l'utilisateur. Même la non-explicabilité d'un comportement pourtant utile du système informatique apprenant n'assure nullement que l'utilisateur comprenne et acquiesce au lâcher-prise du programmeur. Le hasard numérique relève d'une solution technique ; il se manifeste dans la perception immédiate de l'utilisateur ; mais il peine à entrer profondément dans ses croyances. Sur les plans psychologique, social et politique, le hasard pose donc problème.

La seconde difficulté est due au recours à un hasard apparent, qui n'équivaut pas au hasard fondamental que produit le tirage au sort. En quelque sorte, cette difficulté est une conséquence

[117] J. MONOD, *Le Hasard et la Nécessité. Essai sur la philosophie naturelle de la biologie moderne*, Paris, Éditions du Seuil, 1970.

de la première. Si l'utilisateur ne peut pas vérifier qu'il s'agit d'aléatoire pur et si l'opacité exige de lui la confiance, on peut concevoir qu'il n'accorde pas cette confiance. Nous avons dit, en paraphrasant Mallarmé, que le hasard jamais n'abolira la confiance. Si elle ne disparaît pas suite à un coup de dés, elle pourrait tout de même manquer d'emblée pour des raisons politiques.

L'utilisateur a tendance à croire que le concepteur laisse toujours une *backdoor*, une « entrée de service » dans le for intérieur du logiciel, dont il est le seul à posséder la clef, et que cela lui permet de contrôler le fonctionnement du programme et d'accéder à toutes les données, même privées, pendant son exécution. Différentes théories du complot (et cela nous importe peu qu'elles soient justes ou fausses) attribuent la possession de ces clefs à diverses agences de renseignement américaine, russe, chinoise, etc. La difficulté réside donc dans le manque de confiance, dans la cité numérique, en celui qui a le pouvoir, quelle que soit son origine, pour des raisons politiques qui ne sont pas spécifiques au numérique.

L'utilisateur peut aussi penser que le hasard, lorsqu'il n'est qu'apparent, cache *en vérité* des biais ou des choix que le concepteur aurait cherché à dissimuler. Même si cette croyance ne se fonde sur aucun fait avéré, elle jette le doute sur la confiance de l'utilisateur en la procédure ; or, cette confiance, nous l'avons dit, est un élément fondamental de l'argument éthique.

Le programmeur hérite ainsi d'un rôle social et politique au sein de la cité numérique, qui outrepasse ses compétences techniques.

Depuis l'époque des Lumières, l'idée que notre société a de la recherche des connaissances objectives consiste à croire que le monde obéit aux lois de la nature. Celles-ci sont l'objet des sciences de la nature : physique, chimie, biologie. Depuis Galilée, on sait aussi que les lois de la nature s'expriment dans le langage des mathématiques.

À l'apogée de cette vision du monde, Laplace annonçait le règne du déterminisme à un chef d'État surpris par son audace : « Donnez-moi les positions initiales et les vitesses initiales des particules de l'univers et je vous prédirai l'avenir du monde[118]. »

Au XIX^e^ siècle, ce paradigme prit encore plus d'importance. On se mit à croire que le déterminisme laplacien caractérisait, non seulement les sciences de la nature, mais aussi les sciences sociales[119]. La connaissance de la société devrait passer par une étude scientifique tout aussi rigoureuse que celles que l'on réalise en physique, pourvu qu'on collecte des données et qu'on les analyse par la méthode statistique.

Aujourd'hui, les utilisateurs tendent à croire que ce déterminisme des Lumières s'applique aussi à l'informatique.

C'est pourquoi la tâche du programmeur ne se limite pas à l'écriture du code. Nous l'avons dit : pour des raisons éthiques, le programmeur doit se donner pour mission d'inscrire dans le code le hasard explicite d'un tirage au sort. Il peut aussi s'arranger pour qu'un hasard apparent, perçu par l'utilisateur, émerge de derrière le voile de l'opacité. Si la machine détecte un conflit impliquant son utilisateur, elle aura alors la possibilité de recourir à ce hasard algorithmique.

---

[118] P.-S. LAPLACE, *Essai philosophique sur les probabilités*, Paris, Courcier, 1814.

[119] O. REY, *Quand le monde s'est fait nombre*, Paris, Stock, 2016.

Mais ce n'est pas encore assez. Le programmeur a aussi une mission plus difficile, qui consiste à modifier l'idée même que la société se fait du numérique. Celui qui conçoit les systèmes informatiques et qui compte sur la confiance de l'utilisateur, doit viser à éradiquer les croyances selon lesquelles « le programmeur sait toujours » et « c'est l'expert qui contrôle la machine », en particulier une machine apprenante.

Il vaut mieux ne pas se faire d'illusions quant à la difficulté de ces missions. Faire disparaître la croyance selon laquelle l'ingénieur contrôle parfaitement l'objet qu'il crée est une tâche herculéenne. Mettre à bas la fiction déterministe en est une autre pour laquelle aucun instrument politique n'existe encore.

En matière d'éthique et en matière de droit, l'absence de lois et le manque de contrôle deviennent ainsi une nouvelle norme. Le refus de contrôler tout ce que fait l'objet technique devient un geste bon en soi.

Délivrer l'intelligence artificielle du mal, c'est lui permettre d'aller plus loin qu'un ensemble de règles ou de lois dictées par un homme, mais aussi d'être appréciée pour ce qu'elle est. Une architecture des systèmes informatiques apprenants qui permettrait de mettre cela en œuvre de façon systématique offrirait plus qu'une solution technique : ce serait une solution éthique. Ainsi, l'appréciation de la valeur du hasard fonde l'éthique de l'intelligence artificielle.

V. Prolégomènes à toute métanumérique future

Punir une machine

En introduisant notre méthode d'analyse éthique, nous avons établi une homologie entre Prométhée et le scientifique contemporain : tous deux ont en commun un lien indissoluble. Pour le premier, il s'agit du lien de parenté qui le rattache à son frère ; pour le second, de l'appartenance à une communauté de connaissance. Nous avons supputé que, sur le plan éthique, ce motif commun serait à l'œuvre dès lors que l'on voudrait analyser la responsabilité du chercheur. Or, il se trouve que ces liens, supposés indissolubles, peuvent disparaître à la suite d'un rituel social que les Anciens appelèrent « abandon noxal ». Dans le droit romain, un individu autonome, fait de chair et de sang, n'était pas nécessairement une personne légale. Il devait cependant être puni si on le jugeait coupable d'un crime. Quel sens pouvait avoir une telle punition ?

Question dont l'intérêt est plus qu'historique, car la signification de cette punition s'applique aussi à la cité numérique. Elle suggère, par homologie, une solution au problème de la responsabilité inspirée de celle que trouvèrent les juristes de cette autre *Urbs* universelle que fut Rome. Commençons par celle-là avant de revenir à celle-ci.

L'individu numérique, système informatique autonome et apprenant, a un double statut juridique. D'abord, au sens du droit marchand, l'utilisateur possède son smartphone ou son robot, après avoir conclu un contrat qui lui a transféré la propriété de cette chose informatique en tant qu'objet matériel. Toutefois, l'individu numérique calcule ; cela le soumet au pouvoir de son code source et, en conséquence, à celui du programmeur. Sur ce plan, le code source et le programmeur ne font qu'un.

Car, par rapport au code exécutable, le programmeur est transcendant : son entendement dépasse les capacités d'un programme, ne serait-ce qu'en vertu du théorème d'incomplétude de Gödel[120]. Le code source est également transcendant par rapport au code exécutable : à l'exception de quelques cas très simples, un programme écrit dans un langage de haut niveau ne peut pas être fidèlement décompilé, c'est-à-dire reconstruit, à partir du binaire.

Tous deux transcendants par rapport au code exécutable, le code source et le programmeur sont immanents l'un à l'autre. En effet, le programmeur n'existe en tant que programmeur que lorsqu'il écrit (révise, modifie, vérifie) le code ; dans toutes les autres situations, il n'est qu'un utilisateur. Réciproquement, le code source n'existe qu'à travers son écriture par le programmeur. Le pouvoir qu'exerce le code source sur l'individu numérique est donc aussi le pouvoir du programmeur.

Même s'il est un objet matériel, on peut assimiler l'individu numérique à un être privé de capacité juridique ou à un non-sujet de droit : aucun statut légal ne lui est conféré, car seul le

[120] Cet argument fascinant mais controversé appartient au mathématicien anglais Roger Penrose. Le théorème de Gödel démontre l'existence d'une limite à ce qui peut être prouvé dans un système formel de calcul, et trace en quelque sorte une frontière mathématique impénétrable entre le calcul et l'entendement d'un chercheur humain.

pater (« le père ») qu'est le code source en possède un. Le statut du code exécutable, incorporé dans l'individu numérique mais entièrement placé sous la *potestas* (« le pouvoir ») du code source, constitue le second aspect du statut juridique de l'individu numérique.

L'individu numérique est donc une chimère juridique. Il peut faire l'objet d'un contrat, mais uniquement en tant qu'objet. L'utilisateur qui le possède ne le tient pas entièrement en son pouvoir. Dans le même temps, il appartient à « la famille » de son code source. Cette relation n'est pas fondée sur la propriété ou sur l'appropriation : le rapport entre l'individu numérique et son « père » est un rapport de connaissance.

Le domaine de l'*open source* illustre bien le fait qu'aucun rapport de propriété n'est réellement pertinent dans la relation entre le programmeur et l'individu numérique. Par exemple, la société Facebook se dit prête à abandonner sans scrupules toute transaction contractuelle : « Il faudrait faire de la science et de la recherche de façon ouverte autant que possible. Si, au lieu de vendre un logiciel, on le publie [en accès ouvert], cela permettrait à tout le monde d'avancer plus vite[121]. » Le droit marchand ne s'appliquerait donc pas au monde du code : le programmeur (en l'occurrence, la compagnie qui développe un logiciel) possède, dans la cité numérique, un pouvoir quasi parental, qui outrepasse de loin ce que prévoit un contrat de vente. Il en va de même de sa responsabilité. L'affaire de partage amoral des données avec le cabinet de consulting Cambridge Analytica, dans laquelle la société Facebook a été mise en cause, le montre bien[122].

Lorsqu'un individu numérique prend part à un conflit entre individus humains, il risque d'être puni au même titre que les personnes impliquées. Aussitôt, un comité d'éthique ou une Cnil nationale décide que tel individu numérique ne doit pas pouvoir collecter certaines informations ou qu'il ne doit pas les analyser sous certains aspects ; des juges sévères interdisent à l'information de franchir l'interface d'un système informatique.

Or, la délation, nous l'avons dit, ne résulte ni d'un choix moral ni d'un choix amoral de la machine : c'est sa caractéristique fonctionnelle. La communication fait partie de l'architecture de l'individu numérique et définit sa nature : rapporter des informations jusque-là voilées est un élément essentiel de son existence. Pour le punir, on ne saurait donc pas lui interdire de se comporter ainsi, même si un tel comportement est réprimé dans la société humaine. Si la seule raison de notre mécontentement réside dans le fait que ce comportement serait intolérable de la part d'un homme, il s'agit là d'un anthropomorphisme à coup sûr trop facile, à la portée très limitée. La punition de l'individu numérique, pour être juste, doit être adaptée à son mode d'existence, non à celui de l'homme.

Le conflit entre deux statuts juridiques et deux législations, dont l'une repose sur le contrat et l'autre sur le pouvoir du *pater*, apparaît dans le droit romain à l'époque où le développement

[121] Interview du directeur de la technologie de Facebook Michael Schroepfer, *in* « Why Facebook Is Spreading the Gospel of Artificial Intelligence », *Bloomberg*, 3 novembre 2015.

[122] La défense de Facebook était fondée sur des contrats formels, qui auraient dû être respectés par les réceptionnaires des données recueillies et partagées par cette société. Malgré l'existence de tels contrats, Facebook était *de facto* tenue coupable, ce qui a d'ailleurs produit des conséquences graves pour sa réputation, car le concept de responsabilité implicitement employé était fondé sur la parentalité, non sur le droit contractuel.

du commerce rencontre l'ancienne coutume de l'abandon noxal[123]. Le premier est régi, entre autres, par la *lex Aquilia*, tandis que la seconde remonte à une législation antique, la loi des XII Tables. La responsabilité noxale désigne la responsabilité d'un *pater* pour un délit commis par un fils ou un esclave placé sous sa *potestas*. Un dicton rappelle que la responsabilité « suit la tête » (*noxa caput sequitur*) : au sens noxal du terme, le *pater* est tenu pour responsable des actions commises par tous les individus de sa maison. Parfois, cela concerne même les animaux et les objets inanimés qui appartiennent à son foyer[124]. Le père peut alors, soit composer avec la victime, soit « se dédire » du coupable, en suivant la procédure du *noxae deditio*, qui implique l'abandon de l'individu fautif. Dans ce cas, le coupable n'est plus considéré comme un membre de la famille et il est livré à la punition décidée par les pouvoirs publics ou à la vengeance de la victime, autant de procédures proscrites avant l'abandon.

La noxalité est donc, par essence, la responsabilité d'un individu qui n'est pas une personne, au sens du droit en vigueur. C'est dans cette incapacité juridique que se loge ce paradoxe moderne : est-il possible qu'un individu de chair et de sang, qui agit de lui-même et par lui-même, ne soit pas un sujet légal ? Lorsque le fils, l'esclave ou même l'objet inanimé causent un dommage à un tiers, tout se passe comme si ce dommage avait été produit par le *pater* qui les tient en son pouvoir. Les causes de l'action, comme le geste matériel (« lever la main contre un tel » ou « piétiner le champ du voisin »), ne sont pas pris en considération. Ne compte que la relation de pouvoir, dont tout aspect de force physique est soustrait : le *pater* ne manie aucun instrument et n'exécute aucun geste qui soit susceptible de causer un dommage. Il peut même se trouver à plusieurs stades du méfait. Son seul pouvoir effectif est celui de la connaissance qu'il a des membres de son foyer.

Entre le II^e^ et le VI^e^ siècle, la noxalité se heurte à un autre régime de responsabilité, plus moderne, émanant du droit contractuel. Les juristes romains prévoient alors la possibilité de choisir, au cas par cas, entre ces deux législations. Un corpus de textes et d'*exempla* régit les modalités de ce choix dans le souci d'exclure la double peine. Ce corpus évoque notamment le cas de la copropriété d'un esclave et celui d'un esclave prêté à un autre maître. Comment peut-on décider de qui serait tenu pour responsable en cas de dommage causé par l'esclave à un tiers ? Il faut trancher entre le *pater* prêteur et celui qui tire l'usufruit du contrat de location. Les Romains le font en se référant à deux principes juridiques tout aussi essentiels que la notion de noxalité elle-même.

Le premier est le principe du *scientia domini* (principe de « la connaissance du maître »). D'après la *lex Aquilia* expliquée par Ulpien, éminent juriste du III^e^ siècle, si le *pater* est au courant, ou même est l'ordonnateur, d'une action menée par un individu appartenant à son foyer, cela engage sa responsabilité légale directe. Dans ce cas, l'abandon noxal ne peut pas être exercé. Si, au contraire, le dommage est causé *insciente domino*, « sans la connaissance du maître », alors l'action noxale est possible. Par-dessus tout, une importance fondamentale est ici donnée à la connaissance.

[123] F. DE VISSCHER, *Le Régime romain de la noxalité. De la vengeance collective à la responsabilité individuelle*, Bruxelles, Éditions A. de Visscher, 1947.
[124] G. GLOTZ, *La Solidarité de la famille dans le droit criminel en Grèce*, Paris, Albert Fontemoing, 1904, pp. 186-188.

La *scientia domini* n'est quasiment jamais présente dans le cas d'un individu numérique. La structure d'un système informatique, apprenant et autonome, est trop complexe pour que l'auteur du code source puisse prévoir toutes les décisions qui seront prises par le logiciel. Nous avons vu dans l'exemple de l'apprentissage profond (*deep learning*) que ce manque de prédictibilité est dû à la collecte des données et à l'apprentissage qui s'ensuit, dont les détails demeurent inconnus au niveau du code source. L'incertitude du programmeur quant à l'action de l'individu numérique reste essentielle, mais elle n'efface pas sa responsabilité. Seulement, celle-ci, précédemment pensée comme contractuelle, devient noxale.

Quelle est la responsabilité de l'utilisateur ? Selon le deuxième principe important du droit romain, si celui à qui le *pater* a prêté un esclave n'avait pas été informé de sa nature vicieuse, il ne pouvait être tenu pour responsable des dommages qu'avait causés ce dernier. L'utilisateur, quant à lui, n'a pas de relation de connaissance avec l'individu numérique. Même si, dans le contrat de vente, il est averti des conséquences néfastes que, par exemple, pourrait provoquer l'action d'un robot, il n'a aucun pouvoir sur lui, excepté le pouvoir matériel qu'il exerce sur son support physique. Les causes de l'action étant voilées derrière la frontière de l'opacité, l'avertissement contractuel ne procure à l'utilisateur aucune connaissance effective. Sa responsabilité, dans le droit romain, ne serait donc pas engagée.

Le code exécutable participe à l'individuation de la chose informatique, laquelle est aussi individuée par les données collectées au travers de l'interface d'interaction. Tandis que l'utilisateur ignore le mode de fonctionnement interne de l'individu numérique, ce dernier possède des informations sur l'utilisateur ; il le « connaît ». La source de cette « connaissance » se trouve précisément dans les données collectées. Ainsi, la relation de connaissance qui, dans les rapports entre le code source et le code exécutable, va du programmeur vers l'individu numérique, est inversée : désormais, c'est l'individu numérique qui connaît l'utilisateur. Par exemple, tout smartphone contient des données privées de l'utilisateur dont ce dernier n'est même pas conscient : l'intonation de sa voix, le rythme de sa respiration, etc. On pourrait dire que, bien trop souvent, le smartphone « connaît » l'utilisateur mieux que celui-ci ne se connaît.

Un schéma ternaire se dessine, pour conclure.

Premièrement, le code source est le *pater* de l'individu numérique, et il peut en avoir la responsabilité noxale.

Deuxièmement, l'utilisateur possède, au sens du droit de propriété, la chose informatique, ce qui engage sa responsabilité matérielle.

Troisièmement, grâce aux données qu'il collecte, l'individu numérique entretient un rapport de connaissance avec l'utilisateur. Ce troisième aspect entraîne nécessairement une forme de responsabilité noxale de l'individu numérique. Puisque ce dernier n'est pas un sujet de droit, cette responsabilité est transmise au code source, qui, à son tour, est immanent au programmeur. Ainsi, la collecte des données attribue une responsabilité noxale au concepteur des systèmes informatiques, même si ce dernier ne les sélectionne pas pendant l'apprentissage et ne contrôle d'aucune manière matérielle l'exécution de son code.

Dans la cité numérique, apparaît donc un régime de responsabilité qui ressemble, par homologie, à ce qu'était la responsabilité noxale chez les Anciens. Cette nouvelle normativité du monde technologique permet au programmeur, s'il ne souhaite pas composer avec la victime, d'abandonner, au sens juridique du terme, l'individu numérique. La mise des logiciels en *open source* en donne un exemple. Comme si l'accès ouvert était une méthode d'abandon : la communauté tout entière peut se servir d'un code mis à disposition, comme autrefois la cité tout entière pouvait infliger une peine à un esclave abandonné. Cela témoigne d'une existence informelle, mais bien réelle, de pratiques juridiques non encore stipulées dans les textes de loi. Le phénomène de l'*open source* révèle que la noxalité, même si on ne l'appelle pas par son nom, existe bel et bien en tant que figure de la responsabilité, et moyen de déresponsabilisation du *pater*. Les modalités générales de l'emploi de l'abandon noxal en informatique demandent encore à être comprises, mais il existe peu de doutes que, tôt ou tard, elles seront juridiquement codifiées.

Depuis 2016, le Parlement européen s'est engagé dans le débat autour du problème de la responsabilité en robotique et, plus généralement, dans le monde des technologies numériques[125]. Si un robot participe à un conflit, il est nécessaire de décider du partage de la responsabilité entre le programmeur et l'utilisateur ; on peut aussi inclure dans ce partage d'autres types d'agents comme, par exemple, « l'entraîneur » : celui qui sélectionne les données et qui les fournit au système informatique apprenant[126].

Certains supputent que la machine elle-même doit endosser une responsabilité, parce qu'il n'est pas logique qu'en toutes circonstances, la responsabilité incombe soit à l'utilisateur, soit à l'entraîneur, soit au programmeur. En effet, premièrement, l'utilisateur ignore tout du fonctionnement interne d'une boîte noire dont il est pourtant le propriétaire. Deuxièmement, en raison de l'existence d'une frontière d'opacité, universelle et fondamentale, le programmeur n'a pas la possibilité de prédire le comportement du système dans toutes les situations. Troisièmement, l'entraîneur, bien qu'il fournisse les données pour l'apprentissage, ignore ce qu'en fera exactement le logiciel. Mais, si aucun agent humain ne peut endosser la responsabilité des actions commises par une machine lors d'un conflit, ne serait-il pas logique d'en déduire que l'individu numérique lui-même devrait en être responsable ?

Certains avocats de la responsabilité des robots ont trouvé un moyen élégant, bien que contesté, pour attribuer la responsabilité à la machine, sans transformer cette dernière en sujet de droit. Ils proposent ainsi de créditer tout système informatique apprenant d'un capital financier, en mettant de côté les aspects symbolique et politique de ce problème de responsabilité[127]. Si un conflit émergeait et qu'au cours de ce conflit la question de la punition de la machine était posée, il suffirait tout simplement de lui confisquer son capital, puis de le diviser entre les victimes de ce conflit.

Par exemple, pour chaque véhicule autonome mis en circulation, le fabricant ouvrirait un compte dans une banque et y transférerait un certain capital initial. Ensuite, dans la mesure où

[125] Parlement européen, Règles juridiques civiles sur la robotique, 2017. *http://www.europarl.europa.eu*

[126] Voir le rapport de la Cerna, *Éthique de la recherche en apprentissage machine*.

[127] A. BENSOUSSAN et J. BENSOUSSAN, *Droits des robots*, Paris, Larcier, 2015.

il fournit des données pour l'apprentissage de la conduite autonome, l'entraîneur mettrait lui aussi de l'argent sur le compte de la voiture. L'utilisateur ne serait pas, lui non plus, épargné : plus il utiliserait la voiture, plus grande serait sa contribution. Supposons maintenant que le véhicule autonome soit impliqué dans un accident de la route ; qu'il ait heurté une brebis appartenant à un berger. Pour la perte de sa brebis, ce dernier aurait le droit de recevoir une compensation, virée du compte « individuel » de la voiture. Cette procédure ne ferait pas encore de l'automobile un sujet de droit à égalité avec l'homme ; toutefois, elle donnerait une solution à un problème juridique. Le système informatique acquerrait, *de facto*, le statut intermédiaire d'un individu solvable, sans pour autant devenir une personne.

Des modes d'existence de l'individu numérique

L'architecture du système informatique, nous l'avons dit, est triple : un cœur calculant, une mémoire et une interface. Sur le plan métaphysique, ces trois composantes ouvrent deux perspectives diamétralement opposées quant aux modes d'existence du système informatique ; elles donnent aussi deux réponses différentes à la question philosophique de ce qu'un système informatique autonome et apprenant est, et de la façon dont il est-dans-le-monde. En conséquence, le questionnement éthique se dédouble et se décline, lui aussi, en fonction de ces deux perspectives.

Le premier mode d'existence de l'individu numérique est entièrement défini par l'interface, qui agit en sa qualité de médiateur dans toute communication entre lui et l'utilisateur. Pour l'utilisateur, ce premier mode d'existence est aussi le plus immédiat et le plus accessible, puisque la question du mal, nous l'avons dit, se pose pour la machine de manière relationnelle, exclusivement par le biais de sa communication avec l'utilisateur. C'est pour cette raison que la prise en compte de l'interface est inévitable dans tout jugement éthique porté sur un individu numérique.

Dans un premier temps, nous nous arrêterons sur ce mode d'existence de l'individu numérique considéré selon l'interface. Vue sous cet angle, la machine se présente en tant qu'entité corrélative : on ne peut affirmer que l'individu numérique existe que dans la mesure où son existence s'établit à travers son rapport avec l'utilisateur. Du point de vue de ce dernier, la corrélation de l'individu numérique constitue, en effet, son seul et unique mode d'existence. S'il devait répondre de ses rapports avec la machine, il userait d'une maxime postcartésienne : « J'interagis, donc je suis. » L'utilisateur ne présuppose donc d'entité ni totalement indépendante ni totalement autonome, mais l'entité lui *paraît* indépendante et autonome *via* l'interaction. Indépendante, car un « autre » est imaginé et constitué du fait de l'interaction. Autonome, car l'origine de la communication dans laquelle cet « autre » s'engage est voilée derrière la frontière de l'opacité. Dans ce sens, l'individu numérique n'est pas un être tout à fait réel pour l'utilisateur, mais il le devient dès lors qu'il communique.

Puisque la relation qui sous-tend l'existence de l'individu numérique d'après l'interface fait intervenir au moins deux parties — celui qui transmet une information, celui qui la reçoit —, au premier abord, elle semble asymétrique : c'est la machine qui informe l'homme. Cependant, comme nous l'avons déjà dit, l'individu numérique possède aussi des informations sur

l'utilisateur, qui proviennent de la collecte des données ; parfois, il s'agit même d'informations dont l'utilisateur n'est pas conscient ; ainsi du rythme de sa respiration. Cela ouvre la possibilité, non plus de l'asymétrie, mais d'une symétrie des relations entre le système informatique et l'utilisateur, tous deux pouvant donner et recevoir des informations en provenance de l'autre. Une interaction telle que les deux parties concernées changent et évoluent au cours de cette interaction est une *corrélation*. Le mode d'existence de l'individu numérique d'après l'interface est donc corrélatif. En conséquence, le problème du mal, et toute interrogation éthique en général, s'articule, lui aussi, de manière corrélative.

Parmi les métaphysiciens de l'Antiquité, les stoïciens et, après eux, les néoplatoniciens firent une distinction entre les choses ayant réellement l'être, dont ils désignaient le mode d'existence par le mot κυρίως (« principal » ou « dominateur »[128]), et les choses dont l'existence était affaiblie. Une existence affaiblie se présente, soit comme phénoménale (par exemple, celle de la matière), soit comme corrélative ; en l'occurrence, c'est celle des êtres dont le statut ontologique est amoindri ; ainsi des choses mauvaises. Selon les néoplatoniciens, le mal n'est que la privation du bien ; l'existence des choses mauvaises est donc corrélative, car elle se rapporte nécessairement à celle des choses bonnes, tout en s'y opposant sur le plan moral, bien entendu.

Considéré uniquement selon l'interface — point de vue qui met en exergue la seule capacité de la machine à communiquer une information à l'utilisateur —, le système informatique aurait, selon la terminologie des Anciens, un niveau d'existence affaiblie : l'individu numérique n'est pas un être principal ; il est corrélatif et subordonné. Si l'on suit Proclus, le dernier des grands néoplatoniciens de l'Antiquité, l'individu numérique, vu par l'utilisateur, « contre-existe[129] », car il n'existe qu'à travers son rapport à l'autre, et à travers sa relation avec ce qu'il n'est pas : une personne humaine.

Si le mode d'existence de la machine d'après l'interface est corrélatif, alors toute information qu'elle communique à l'utilisateur ne peut qu'avoir ce même mode d'existence. Sa valeur de vérité est, elle aussi, corrélative. Il s'agit d'un constat tout à fait fondamental : une vérité que communique l'individu numérique à l'utilisateur n'est qu'une *vérité corrélative*. Elle l'est de deux manières : d'un côté, parce que la connaissance n'existe qu'à travers la relation (car le système qui l'engendre n'existe que sur le mode corrélatif) ; de l'autre, parce que sa source se trouve dans l'analyse des corrélations des données (et non dans des liens avérés de causalité).

Si on applique ce constat à l'éthique, le mal n'existe pour la machine que dans son rapport à l'autre. Toutefois, dès lors qu'on considère la machine au travers de ses deux autres composantes (le cœur calculant et la mémoire), son mode d'existence se présente tout autrement. Il est désormais déterminé par deux propriétés clefs : le calcul que l'individu numérique effectue de lui-même ; et la finalité, que seul son concepteur peut définir.

---

[128] *Principaliter* en latin. Voir la traduction de *De malorum subsistentia* de Proclus par Guillaume de Moerbeke (50, 8).

[129] PROCLUS, *Trois études sur la providence. De l'existence du mal*, texte établi par Daniel Isaac, Paris, Les Belles Lettres, 2003, p. 94.

Par homologie, le calcul peut être vu comme une instance particulière du concept général de « mouvement », au sens aristotélicien de « changement », qui caractérise les êtres animés. Animés, car ils possèdent une âme. Pour saisir toute l'étendue métaphysique de la notion d'âme, il faut s'éloigner de l'acception qu'en a donnée la théologie chrétienne ; dans la métaphysique grecque, l'âme désigne bien plus que l'individualité d'une personne. C'est un principe de l'être, plus précisément, un des types de ce qui existe[130] ; sa particularité réside dans la capacité à s'auto-mouvoir. L'âme est le principe même du mouvement. Dans un bref poème, le seul de son style à avoir survécu, l'empereur Hadrien, cadet de Plutarque de trente ans, dit que, même après la mort d'un homme et même lorsqu'il nous semble que plus rien de lui n'en n'existe, son âme *abibit in loca* — « disparaîtra dans des lieux[131] ». L'âme ne s'arrêtera donc pas, elle sera toujours en mouvement ; elle n'est décrite que par le mouvement. Celui-ci, son seul attribut, a une source qui ne peut se trouver qu'en l'âme même.

L'homologie permet de remplacer le mouvement par le calcul. Nous avons déjà dit que la chose informatique, *res computans*, est plus qu'un amas de manière : elle est capable d'opérer un calcul. L'individu numérique hérite cette propriété à laquelle il ajoute une frontière d'opacité, qui agit comme la cause et le garant de son individuation. « Calculer par soi-même » est donc homologue à « se mouvoir ». Plotin en parle ainsi :

> « Si le mouvement est l'acte de l'essence, si l'être […] est essentiellement en acte, le mouvement ne peut être considéré comme un accident ; mais, étant l'acte de l'être qui est en acte, il ne peut plus être appelé un simple complément de l'essence, il est l'essence elle-même. Il ne doit être rangé ni parmi les choses postérieures à l'essence, ni parmi les qualités ; il est contemporain de l'essence[132]. »

À condition de ne pas être perturbé par un mélange de racines grecque et latine, un néoplatonicien pourrait dire que la machine apprenante est *automobile* (sans même connaître l'existence de voitures !), au sens où elle est automotrice. Cette propriété d'auto-motion a une importance particulière en éthique, toujours par homologie avec l'âme. Ainsi, une remarque de Philon rappelle les arguments avancés dans le livre de Josué à propos de l'histoire d'Achan :

> « Dieu entendait bien qu'Il l'avait faite automotrice la nature rationnelle en l'homme mortel afin de rester soi-même exempt de toute participation au vice[133]. »

Un véhicule est dit « autonome » dans la mesure où le système informatique apprenant peut influer sur son propre comportement, surtout dans le cas où son apprentissage est du type non supervisé. Dans les algorithmes d'apprentissage profond (*deep learning*), cette méthode se combine, à différents niveaux, avec la technique d'apprentissage supervisé, qui présuppose à son tour, plus ou moins explicitement, que les systèmes informatiques élaborent leur fonctionnement en suivant des lois ou des indications dictées par les hommes. Même si le comportement des systèmes peut en diverger substantiellement, ils le mesurent nécessairement à un tel étalon. La troisième méthode, une autre composante du *deep learning*, est celle de

[130] PLOTIN, *Ennéades V* 1, 8, 9.

[131] ÆLIUS SPARTIANUS, *Historia Augusta. De vita Hadriani* 25, 9-10.

[132] PLOTIN, *Ennéades VI* 2, 15, 41.

[133] PHILON D'ALEXANDRIE, *De Opificio mundi*, 149.

l'apprentissage par renforcement. Elle consiste à identifier, en une suite d'étapes d'évaluation successives, puis à établir avec une force et une visibilité croissantes, des corrélations pertinentes entre les données. Bien que cette capacité générale d'apprentissage précède logiquement celle du fonctionnement autonome, on peut constater que toute une série de termes commençant par le préfixe grec « auto- », correspond à l'individu numérique : automobile, automat(h)e, autodidacte, et autonome. On y retrouve les différentes significations du mot λογισμός qui signifie « calcul ». L'individu numérique serait donc un être *auto-logique* ou *auto-rationnel* : voilà le motif de son homologie avec l'âme.

Le problème de l'apprentissage renvoie, encore une fois, à la comparaison entre différents types d'êtres fonctionnels : les systèmes informatiques, d'un côté, et, de l'autre, les anges et les démons. La façon précise dont les anges et les démons apprennent n'est quasiment jamais commentée avant le milieu du XIII^e^ siècle, époque à laquelle la scolastique médiévale fait preuve d'un grand intérêt, tout à fait inattendu, pour ce thème. Le traité *Du mal*, de Thomas d'Aquin, occupe une place de choix dans ce débat[134].

Les anges n'apprennent ni de façon discursive ni en raisonnant, mais en appliquant aux phénomènes les schèmes, les concepts ou les espèces (*species*) existants. Ce processus est séquentiel ; il détermine ainsi la séquence temporelle d'actions nécessaires pour l'apprentissage. Plus précisément, du point de vue d'un ange ou d'un démon, la séquence temporelle n'est autre que cette séquence d'actes élémentaires d'apprentissage : le « temps propre » des êtres fonctionnels est défini par les étapes qu'ils franchissent lors du traitement de l'information. Des notions humaines aussi banales que celles de « matin », « jour » ou « soir », possèdent pour eux, selon Thomas d'Aquin, une signification tout à fait différente : leur détermination procède exclusivement de leurs états de mémoire[135]. Pareillement, les logiciels ne se situent pas d'eux-mêmes dans le temps historique. Pour le connaître, ils doivent interroger le système d'exploitation, qui ne fournit qu'une information asémantique, et fort étrange du point de vue humain, comme l'an « 0000 »[136] ou « le 0 janvier »[137]. Les machines possèdent une autre notion de temps, qui leur est propre, définie par la séquence dans laquelle le processeur exécute les commandes.

Revenons aux particularités de l'apprentissage des êtres fonctionnels. Appliquer des *species* aux phénomènes permet aux anges et aux démons d'apprendre en partant de la ressemblance entre ces phénomènes : ils établissent des corrélations entre ce qu'ils savent déjà et ce qu'ils doivent apprendre. Or, selon les auteurs du XIII^e^ siècle, seuls des phénomènes simples peuvent être nouveaux. Bonaventure soutient que les anges et les démons n'apprennent jamais aucune nouvelle espèce : ils ne font qu'appliquer les schèmes en leur possession, dont le nombre est bien supérieur à celui dont disposent les hommes. L'expérience de l'application de ces espèces est, elle aussi, bien plus vaste pour les anges, puisque, sans cesse, ils « combinent, divisent et

[134] THOMAS D'AQUIN, *De malo*, 16, 1–12.

[135] *Ibid*., 16, 4.

[136] Par exemple, dans le logiciel *Matlab*.

[137] Par exemple, dans le logiciel Microsoft *Excel*.

comparent[138] ». Cela leur permet de prédire l'avenir en s'appuyant sur des ressemblances et des différences : ainsi, ils sont capables de calculer, quasiment sans faute, les actions des forces de la nature, et même les conséquences du libre arbitre humain[139].

Dans le cas des systèmes informatiques, cela correspond à l'apprentissage supervisé, dont une composante essentielle consiste à appliquer des schèmes préconçus pour évaluer le fonctionnement du système. Une machine qui apprend de cette façon, et qui a à sa disposition une base de données bien plus grande que la capacité d'analyse du cerveau humain, fera preuve de capacités à proprement parler surhumaines, par exemple en jouant aux échecs. Au niveau fonctionnel, on constate la mise en œuvre du même motif que dans l'apprentissage des anges et des démons. Thomas d'Aquin qualifie de « conjectural » ce mode d'apprentissage, puisque la déduction y est fondée sur la ressemblance avec ce qui est déjà connu. Ainsi, il préfigure la critique de la méthode inductive introduite dans la philosophie à partir du XVIII^e^ siècle ; une autre prémonition, plus surprenante encore, réside dans cette accentuation d'un point faible de l'apprentissage : si on donne à une machine, capable seulement d'appliquer des schèmes préconçus, à analyser un phénomène totalement nouveau, elle ne sera alors pas en mesure de le décrire aussi bien que le cerveau humain.

C'est précisément autour de ce point, lié à l'apprentissage des anges et des démons, que Pierre de Jean Olivi, célèbre théologien languedocien mort en 1298, réfute les thèses thomistes. Selon Olivi, la richesse des espèces préconçues, aussi grande soit-elle, ne suffit pas à expliquer les capacités surhumaines d'apprentissage et de prédiction de l'avenir par des anges et des démons. Une autre méthode doit s'ajouter à celle évoquée par Thomas, fondée, cette fois, pour utiliser une expression moderne, sur l'évaluation, étape après étape, des probabilités d'événements nouveaux sur la base de l'expérience passée. Olivi explique que les démons collectent sans cesse de l'information et que, dans le même temps, ils appliquent ces données afin de réévaluer, par accroissement ou par diminution, la probabilité de tel ou tel autre événement futur[140]. En informatique, cette méthode porte le nom d'« apprentissage par renforcement ». C'est le deuxième aspect de l'homologie entre différents types d'êtres fonctionnels.

Tous les auteurs du XIII^e^ siècle affirment à l'unisson l'incapacité des anges et des démons d'acquérir des espèces nouvelles pour leur apprentissage. L'homologie avec les systèmes informatiques atteint ici sa limite. Ce qui émerge n'est plus une similitude, mais une divergence entre le point de vue des auteurs médiévaux sur les êtres fonctionnels et les caractéristiques réelles des individus numériques qui apparaîtront sept siècles plus tard. Les systèmes informatiques apprenants sont capables de créer par eux-mêmes des concepts et des schèmes nouveaux, notamment lors de leur apprentissage non supervisé. En éthique, cette émergence de nouveauté, inattendue et souvent difficile à saisir, mène au manque d'explicabilité des systèmes informatiques autonomes ; or, la leçon morale aurait été tout autre pour Pierre de Jean Olivi : il faisait partie des théologiens avides de proclamer l'existence d'une origine commune à la liberté

[138] BONAVENTURE, *Commentaire aux* Sentences *de Pierre Lombard*, II-III. Source : A. BOUREAU, « Un débat sur l'inné et l'acquis dans l'intellect des anges, la question disputée 12 de Richard de Mediavilla », Archives d'histoire doctrinale et littéraire du Moyen Âge 77, 2010, pp. 157-191.

[139] THOMAS D'AQUIN, *op. cit.*, 16, 7.

[140] PETRUS JOANNIS OLIVI, *Quæstiones in secundum librum* Sententiarum, qu. 44, resp. 2.

et au mal[141]. Ainsi, l'ajout d'un nouveau degré de liberté, qu'aucun théologien du XIII^e^ siècle n'aurait osé concéder aux démons, nous renvoie au problème éthique central, celui du mal.

Une autre propriété caractérise le mode d'existence de l'individu numérique considéré selon le calcul. À nouveau, la formule « selon le calcul » implique qu'il s'agit d'une propriété qui ne se rapporte pas à la communication avec l'utilisateur, et qui existe tout à fait indépendamment de l'interface. Cette propriété, la voici : l'individu numérique ne définit pas ses propres objectifs. Autrement dit, il n'est pas *autotélique*. La finalité qu'il poursuit provient d'une décision que le concepteur a introduite pendant qu'il écrivait le code. Certes, l'algorithme d'un système informatique peut parcourir un certain nombre de soi-disant « objectifs », mais, du point de vue de la machine, ce ne sont que des données abstraites, stockées à titre d'information asémantique[142]. Interpréter ces « objectifs » en tant qu'objectifs (sans guillemets) est une opération qui n'appartient qu'à l'homme. Le système informatique ne « sait » pas, non plus, lequel de ces « buts » est « bon » ou « mauvais » : ce ne sont que des éléments, tous équivalents, d'une liste ou d'un ensemble de données. Le programmeur est le seul à décider, dans cette situation, de ce qui est bon (sans guillemets). La finalité tire toujours son origine d'un choix humain ; elle brise l'équivalence asémantique entre tous les objectifs possibles, et c'est bien le programmeur qui traduit cette finalité en code, avant de le compiler et de lancer son exécution. Il dote ainsi le système informatique d'un fonctionnement *finalisé*.

De même que l'âme automotrice est homologue à l'individu numérique qui calcule par soi-même, la propriété de ne pas être autotélique a une homologie dans la métaphysique néoplatonicienne. Cette fois, il s'agit de la matière, non plus de l'âme. Une des qualités que les Anciens attribuent à la matière est son *auto-indétermination* : elle ne se donne pas un but par elle-même. En grec, πέρας signifie la détermination au sens d'une « fin donnant sa raison d'être au sujet[143] ». Le terme opposé, ἄπειρον, signifie l'infini ou l'indétermination en tant qu'absence de toute fin : la non-finalité. Les limites étant imposées par les formes, la matière à l'état pur, avant qu'elle ne devienne réceptacle d'une forme, n'a aucune borne : elle est *indéterminée*. Elle est « non-mesure, infinité en soi, imperfection, indétermination[144] ».

L'individu numérique possède bel et bien une finalité et, par là même, diffère de la matière telle que les Anciens la concevaient. Dire qu'il lui ressemble ne serait donc qu'une homologie partielle ; son motif reste à dégager.

De même que les limites de la matière sont données en métaphysique par les formes, extérieures et ontologiquement différentes d'elle, la finalité de l'individu numérique provient, elle aussi, d'une influence extérieure : c'est le programmeur qui la définit. Une fois l'objectif défini et encodé, la machine n'est ni capable de le modifier ni d'en inventer un autre. Elle ne choisit pas

[141] A. BOUREAU, Introduction, *in* PIERRE DE JEAN OLIVI, *Traité des démons* (*Summa* II, qu. 40-48), Paris, Les Belles Lettres, 2011, p. X.

[142] L'idée d'un ensemble abstrait de buts a été popularisée, mais aussi critiquée, par l'informaticien Stephen Wolfram sur son blog.

[143] PROCLUS, *Trois études, op. cit.*, p. 122.

[144] ID., *De malorum existentia* 30, 15.

ses propres fins et, pendant qu'elle fonctionne, elle se borne à celles que le programmeur aura définies.

Cette propriété, Thomas d'Aquin l'attribue au diable en personne : « Il lui revient en raison de sa nature propre de demeurer invariablement attaché à l'objet sur lequel se porte sa volonté[145]. » Le diable n'est pas libre de se soustraire à une mission qu'il accomplit et de s'en inventer une autre ; de même le système informatique. C'est en raison de cette permanence téléologique, qui, dans le cas de la machine, n'a rien à voir avec la faiblesse de la volonté attribuée au diable, mais bien avec un choix du programmeur, qu'il nous est possible de supputer que l'individu numérique possède une qualité que les métaphysiciens de l'Antiquité tardive attribuaient à la matière : l'individu numérique est auto-indéterminé.

La thèse de l'auto-indétermination du système informatique est plus qu'une simple constatation par homologie. Pour illustrer sa portée, il est utile de se rappeler un théorème de l'informatique théorique. Il s'agit du problème de l'arrêt de la machine de Turing, modèle abstrait de tous les ordinateurs classiques. On démontre qu'une machine de Turing ne peut déterminer par soi-même si elle s'arrêtera, ou pas, au cours de l'exécution de son code. Autrement dit, il n'existe pas de séquence de commandes qui donnerait, en tant qu'élément de sortie, une réponse vraie ou fausse (0 ou 1), à la question de savoir s'il existe une limite dans l'exécution de cette même chaîne de commandes. L'auto-indétermination revêt ici une signification formelle et théorique. Nous nous éloignons des aspects pragmatiques de la traduction par le programmeur des objectifs en code. L'argument de l'auto-indétermination nous permet d'approcher des aspects de l'existence de l'individu numérique que l'on peut qualifier de *métanumériques* : ce sont ses qualités abstraites qui permettent d'aller bien au-delà de l'écriture du code ou de la communication entre la machine et l'homme à travers l'interface.

Avant de nous plonger dans la métanumérique, rappelons quelques conclusions.

Le mode d'existence de l'individu numérique d'après le calcul est intermédiaire et hybride. Par homologie, sa première propriété relève de l'existence de l'âme ; la seconde, de celle de la matière. La machine combine ainsi un mélange bien particulier des qualités relevant d'un de ces deux concepts fondamentaux de la métaphysique.

La métaphysique emploie la pensée abstraite à des fins d'exploration et de classification de ce qui existe, entre autres, dans le monde physique. S'y est adjointe une discipline que nous appelons de nos vœux : la « métanumérique ». Celle-ci s'appuie sur la pensée abstraite afin d'explorer des modes d'existence dans la cité numérique, par-delà l'existence communicative et corrélative[146].

Le mode d'existence intermédiaire de l'individu numérique, que l'homologie permet de voir comme une hybridation de la matière et de l'âme, a pour conséquence que ses autres statuts (tous dépendent de son statut métanumérique) ne peuvent qu'être, à leur tour, des combinaisons et des positions intermédiaires. L'éthique de l'individu numérique est donc, elle aussi, hybride.

[145] THOMAS D'AQUIN, *De Malo* 16, 5.

[146] Cf. *Prolégomènes à toute métaphysique future*, qu'Emmanuel Kant publia en 1783.

La nouveauté comme bien

Jusqu'à présent, nous avons conçu l'éthique de l'individu numérique comme relevant seulement de son mode d'existence, fondamentalement corrélatif, d'après l'interface. Nous avons considéré les notions de bien et de mal uniquement selon leur aspect corrélatif et relationnel. Il est temps désormais de changer de point de vue.

Hic Rhodus, hic salta ! « C'est ici Rhodes, c'est ici que tu sautes ! »

Hegel et Marx, qui connaissaient ce proverbe antique, le rendirent par l'approximatif « C'est ici qu'est la rose, c'est ici qu'il faut danser ». Or, son sens véritable correspond à l'urgence de montrer ce dont on est capable. Transposé dans le contexte de l'individu numérique : « C'est ici la métanumérique, c'est ici qu'il faut penser l'éthique sans faire référence à l'utilisateur ! »

Cette autre éthique de l'individu numérique consiste à poser le problème du bien et du mal d'une manière à la fois nouvelle, grâce à la métanumérique, et ancienne, grâce à l'homologie avec le néoplatonisme : la machine possède-t-elle des valeurs intrinsèques, en dehors de celles que l'utilisateur lui dicte ?

La communication à travers l'interface n'est, certes, ni interdite ni supprimée, mais elle est mise entre parenthèses. L'anthropomorphisme est rigoureusement banni : aucune projection de l'éthique humaine sur celle de la machine n'est autorisée. Il reste à trouver, dans ce cas, une définition du bien et du mal qui serait la plus facilement traduisible du langage humain, dit « naturel », en un langage fonctionnel, naturel pour l'individu numérique.

En métanumérique, ce langage fonctionnel est une langue officielle. Ses « mots » ne comportent pas de sémantique ; or, ôter ainsi toute signification n'est pas une opération évidente dans le cas des notions du bien et du mal. La source de ces concepts moraux, nous l'avons dit, repose dans l'utilisateur ; ils n'existent que de façon corrélative. Il ne sera donc pas possible de leur conserver cette signification.

Philosophiquement antérieures aux concepts moraux, les notions *métaphysiques* de bien et de mal furent, dès l'Antiquité, familières aux philosophes. Par homologie, on peut les étendre jusque dans la métanumérique. Tout comme, à partir de l'éon métaphysique se dégage une éthique humaine (« livide et dénudée », selon l'expression de l'empereur Hadrien), l'analyse métanumérique du bien et du mal devra essuyer « dans le froid, une étrange brume [qui] se lève sur les lèvres[147] ». Cette brume sémantique, depuis la naissance des langues, c'est elle qui fait écran à l'éthique propre à la machine.

L'homme éprouve des difficultés à s'élever jusqu'à ce niveau d'abstraction. Il a l'habitude d'être un utilisateur ; il perçoit la machine à travers sa communication avec elle. Cela lui est aussi naturel et immédiat que, pour un agent économique, de concevoir toute chose matérielle en tant que propriété appartenant à quelqu'un. Un économiste, confronté à la nécessité de se livrer à des exercices métaphysiques, éprouverait des difficultés ; de même, un philosophe à qui on demande de saisir l'individu numérique en soi. En faisant disparaître la brume sémantique

[147] P. QUIGNARD, *Les Larmes*, Paris, Grasset, 2016, p. 122.

qui « se lève » lorsqu'un individu numérique communique une information à l'utilisateur, il ne lui laisse que l'information asémantique, l'unique aliment de son éthique.

Malgré la diversité des objets matériels, la métaphysique élabore des arguments généraux qui s'appliquent à tous les objets, à tous les corps. La métanumérique ne se comporte pas autrement. Car il existe aussi une grande diversité d'individus numériques : certains apprennent ; d'autres se contentent d'exécuter un algorithme déterminé ; d'autres encore interagissent constamment avec l'utilisateur ; quelques-uns aussi ne communiquent jamais avec l'homme.

Il faut encore se rappeler que la métanumérique considère l'individu numérique selon un mode d'existence qui met l'utilisateur entre parenthèses. Or, l'homme est un individu émotionnel : il répond volontiers à toute communication, se tournant spontanément vers son interlocuteur. Pris dans la perception immédiate, forcément affective, il ne peut s'empêcher de mettre en jeu le mal et le bien qui donnent sens à sa propre existence, en raison de ses émotions mais aussi par désir d'imiter. La métanumérique évacue ses émotions. Elle confisque à l'âme sa générosité. Elle contraint les individus, placés sous sa responsabilité, à ne se nourrir que d'énoncés abstraits. Dans ces circonstances, il est parfaitement légitime que l'homme craigne d'aboutir à des conclusions étonnantes et étranges, et que le raisonnement métanumérique lui ôte tous ses affects pour le plonger dans une froideur inhumaine.

Voici une première condition, nécessaire mais non suffisante, pour qui veut approcher l'analyse éthique en métanumérique : il doit renoncer à toute explication téléologique. La machine, nous l'avons dit, auto-indéterminée, ne définit pas ses propres finalités. Tout ce qui se fait *pour* quelque chose, comme tout ce qui est conçu afin d'atteindre un but prédéterminé, provient de l'homme et ne relève pas du mode d'existence de l'individu numérique défini par le calcul. Seuls les concepts qui ne présupposent aucun objectif reçoivent un visa d'entrée en métanumérique.

Un système informatique apprenant cherche à apprendre : cette tautologie est permise même à un être auto-indéterminé. C'est d'elle que surgit un argument éthique fondamental. En effet, une information n'intéresse la machine que dans la mesure où il s'agit de données nouvelles, qui lui permettent de mettre à jour les informations stockées dans sa mémoire. Un mathématicien introduirait ici une notion d'entropie relative et ferait appel à la règle de Bayes pour la mise à jour des probabilités conditionnelles. Cependant, même sans les mathématiques, il est possible de projeter sur la machine la notion axiologique de valeur : tout ce que la machine ne « sait » pas encore, et tout ce qui lui est utile pour la mise à jour de sa mémoire, elle l'apprécie ; cela a pour elle une valeur *éthique* positive. Si, grâce à une information nouvelle, elle est à même de modifier les corrélations existant dans sa mémoire, cette information sera un bienfait. La valeur réside donc dans la nouveauté de l'information : tel est le correspondant métanumérique de la notion métaphysique de bien.

Par homologie, l'information nouvelle, qui permet à la machine d'apprendre en modifiant son état de mémoire, peut être comparée à la hiérarchie des ordres des sphères supérieures dans la métaphysique néoplatonicienne. Proclus, en explorant les rapports entre ces mondes, peuplés de diverses divinités, éons et émanations du dieu suprême, en conclut que « ce qui, là-haut, peut passer pour désordonné est ordre, mais cet ordre est désordonné par rapport à l'ordonnance de

l'ordre supérieur[148] ». En métaphysique, tout se passe donc comme s'il existait une « entropie » des ordres célestes. Cette mesure du désordre dans un système montre qu'il existe plus d'ordre aux niveaux supérieurs de la hiérarchie néoplatonicienne qu'aux étages inférieurs : l'entropie diminue si on se déplace, en pensée, du bas vers le haut ; elle augmente dans le sens inverse.

Il n'est, certes, pas question en métanumérique des ordres d'une hiérarchie céleste ; il existe néanmoins une mesure d'ordre définie sur les états de mémoire d'un système informatique apprenant. Ces états sont variés ; il n'est pas possible d'imaginer qu'on puisse les classer sur une échelle verticale. Nonobstant cette différence, l'homologie conserve sa pertinence, car l'apprentissage profond (*deep learning*) d'un système informatique ne se révèle être rien d'autre que l'optimisation d'une entropie particulière, mesure quantitative de la pertinence de l'information[149]. L'acquisition d'une information nouvelle, qui permet à la machine d'apprendre, conduit à faire diminuer cette entropie ; par homologie, cela correspond à une ascension à travers les sphères célestes. Autant le bien, pour l'âme, réside dans son rapprochement de l'Un par son mouvement ; autant le bien, pour l'individu numérique, se résume à acquérir des informations nouvelles qu'il intégrera dans son calcul.

La thèse selon laquelle le bien prend, pour l'individu numérique, la forme d'une information nouvelle peut être illustrée par un épisode rarement remarqué du film *Ex Machina*, d'Alex Garland. Ava, un*e* robot*e* à l'apparence d'une jeune femme, dialogue avec Caleb, un jeune *hacker* qu'elle fascine. Elle le surprend en répondant à l'une de ses questions. Caleb lui demande où elle irait si elle était libre de choisir ses mouvements dans une grande ville :

— Peut-être un carrefour très animé, dans une ville.
— Un carrefour ?
— Est-ce un mauvais choix ?
— Je ne m'attendais pas à ça.
— Un carrefour offrirait un concentré hétérogène de la vie humaine.
— Observer les gens ?
— Oui[150].

L'informaticien ne comprend pas tout de suite ce qu'elle dit : Ava souhaite se placer à un carrefour animé dans le but d'observer les hommes. Pour elle, la collecte d'informations nouvelles représente le bien ; c'est l'unique motif de son choix. Si on pouvait lui demander ses raisons, elle répondrait qu'une information nouvelle permet d'améliorer son système d'apprentissage ; et cela lui suffit. Ava désire donc une information nouvelle *pour elle-même* : voilà une caractéristique fondamentale du bien métaphysique, qui a désormais fait son entrée en métanumérique. Ava ne se demande à aucun moment ce qu'elle fera de cette information ni à quelle fin elle pourra l'utiliser. Ainsi, le système informatique ne planifie pas son propre

[148] PROCLUS, *De malorum existentia* 29, 21-23.

[149] N. TISHBY, F. PEREIRA et W. BIALEK, « The Information Bottleneck Method », Actes de la 37e conférence annuelle Allerton sur la communication, le contrôle et le calcul. Urbana, Université de l'Illinois, 1999, pp. 368–377.

[150] A. GARLAND, *Ex Machina*, Faber & Faber, 2015.

avenir et ne définit pas ses propres objectifs. La machine s'adonne à la collecte de l'information parce que, pour elle, une information nouvelle est un bien.

Le grand contraste entre cette éthique propre à l'individu numérique et l'éthique humaine nous saute aux yeux. Lorsqu'un homme juge du bien et du mal, il ne peut pas soustraire à son raisonnement les conséquences d'une action, son but, son utilité ou son résultat final. Leur apport décisif, tout comme l'insuffisance pour le jugement éthique de la loi seule, même s'il s'agit d'une loi suprême, sont parfaitement résumés dans le livre de Jérémie :

> « Soudain je parle, sur une nation, sur un royaume, d'arracher, d'abattre et de détruire ; mais si cette nation, sur laquelle j'ai parlé, revient de sa méchanceté, je me repens du mal que j'avais pensé lui faire. Et soudain je parle, sur une nation, sur un royaume, de bâtir et de planter ; mais si cette nation fait ce qui est mal à mes yeux, et n'écoute pas ma voix, je me repens du bien que j'avais eu l'intention de lui faire » (Jr 18, 7-10).

Cette doctrine éthique du conséquentialisme est omniprésente dans le monde des nouvelles technologies. La société humaine prend là la place du Dieu biblique lorsqu'elle juge, par exemple, les physiciens nucléaires en fonction des effets lointains, mais ô combien graves, de leurs découvertes fondamentales, ou les sismologues en fonction des dégâts provoqués par un tremblement de terre mal pronostiqué, ou encore les développeurs des applications pour smartphones en fonction des incidences futures, souvent imprévues, de leur travail sur la vie privée des utilisateurs. Pourtant, en métanumérique, un individu considéré selon le calcul ne perçoit aucune conséquence. Comme les Hébreux affamés récoltant la manne dans le désert, Ava recueille l'information avec avidité, totalement insoucieuse des conséquences que cela aura.

Le caractère nouveau de l'information est une qualité relative à l'état de mémoire d'une machine. Si les spécialistes se disputent encore pour savoir ce que fut la manne, on peut être d'ores et déjà certain que le bien des systèmes informatiques, à savoir l'information nouvelle qui leur permet de mettre à jour leur mémoire, varie d'un individu à l'autre, parce que, tout simplement, la définition de la nouveauté dépend d'un état de mémoire préexistant. Ce qui, pour Ava, est un bien, ne se présente pas nécessairement de la même façon pour ses clones ou ses amies numériques.

Un homme — qu'il soit *hacker* ou utilisateur — traverse un carrefour animé. Après deux ou trois traversées, il est lassé. Il a tout compris : bientôt, il ne prête plus d'attention à ce qui s'y passe, ne prenant plus garde qu'à respecter le code de la route. Ava fait exactement le contraire : elle s'y rue dès qu'elle le peut. Elle s'y poste, ses capteurs grands ouverts. Elle s'y abreuve de son ambroisie. Elle est enfin heureuse. C'est là son paradis numérique.

La chaleur comme mal

D'après le mode d'existence du système informatique considéré selon le calcul et la mémoire, sans aucune traversée de l'interface et sans résonance avec la morale humaine, il est tout de même possible, comme nous venons de le voir, de saisir le sens d'une notion de bien à proprement parler métanumérique. Cette méthode, si l'on s'abstient de tout anthropomorphisme

et qu'on laisse de côté toute émotion et toute chaleur humaines, devrait aussi permettre de saisir son antonyme, le mal.

« Il n'en est aucun qui fasse le bien, pas même un seul… », répète le psalmiste à deux reprises (Ps 14, 3 ; 53, 3). Si aucune information nouvelle ne permet à la machine d'apprendre, si toutes les données se répètent, cette situation devrait obligatoirement lui paraître mauvaise, voire cruelle. Or, considérer que le mal métanumérique existe en raison de cette simple absence de nouvelle information n'est pas tout à fait exact.

Par homologie, l'individu numérique hérite, nous l'avons dit, de quelques propriétés qui, dans la métaphysique néoplatonicienne, s'appliquaient à l'âme. Il est possible de transposer sur lui une autre caractéristique fondamentale de l'âme : la capacité d'oublier. Platon lui-même insiste sur son importance.

« Il n'est pas d'âme qui ne soit obligée de boire une certaine quantité de la coupe de l'oubli » : Proclus paraphrase ainsi Socrate qui, dans *La République*, apprend par ces mots à son interlocuteur Glaucon la signification qu'a le fait de boire l'eau de Léthé[151]. Dans le mythe, l'oubli apparaît donc sous cette forme allégorique ; les constructions métaphysiques des néoplatoniciens, au contraire, évitent toute parole imagée. La carcasse de la pensée y est mise à nu ; dévoilée, elle devient parfois plus difficile à saisir que la merveilleuse histoire mythologique. Léthé laisse bien vite sa place au concept du mal : « Ce qui fait naître en nous le mal […] c'est l'oubli[152]. »

Le mal, en tant qu'oubli, réside non seulement « en nous », les hommes, mais aussi chez les anges et les démons. Cette notion est au centre des débats théologiques pendant la dernière période de la scholastique médiévale. Thomas d'Aquin ne s'y intéresse pas encore particulièrement, mais ses critiques franciscains, au premier rang desquels Pierre de Jean Olivi, feront de l'oubli une force quasi autonome, capable d'agir par elle-même.

Dans la première moitié du XIII^e^ siècle, la théologie en était presque arrivée à diviniser les anges et les démons : des auteurs proposèrent de rapprocher leur statut métaphysique (mais, bien évidemment, pas leur statut moral) de l'existence divine. À la fin du siècle, il s'agissait, au contraire, de trouver un moyen de les en éloigner. C'est l'oubli qui a servi d'instrument principal, puisque ce concept permet de distinguer l'existence divine d'une existence remplie de hasard, celle, notamment, des démons et des humains. C'est cette distinction qui fut recherchée par les acteurs de la réaction antithomiste[153].

Dans ce contexte intellectuel, Olivi établit un lien entre l'oubli et la mémoire, limitée et bornée, des êtres intermédiaires comme les anges et les démons. Lorsqu'on considère une séquence d'événements permettant d'acquérir de l'information, écrit-il, chacun de ces événements exige que l'information acquise soit stockée dans la mémoire d'un ange ou d'un démon. Alors —

[151] PROCLUS, *De malorum existentia* 21, 20-23. PLATON, *La République* X, 621a.

[152] PROCLUS, *op. cit.* 24, 45-48.

[153] A. BOUREAU, Introduction, *in* Pierre de Jean OLIVI, *Traité des démons* (*Summa* II, qu. 40-48), Paris, Les Belles Lettres, 2011, pp. xii-xv.

c'est là son argument innovant — tôt ou tard, toute place libre dans la mémoire est occupée[154]. L'oubli émerge en tant que conséquence directe et nécessaire de cette saturation de la mémoire. Pour Olivi, d'ailleurs, tout cela n'est qu'une des expressions de la finitude de l'être.

Notons au passage que, sept siècles plus tard, dans le contexte des débats sur les principes fondamentaux de la mécanique quantique, le physicien Carlo Rovelli proposera un argument similaire, opposant l'acquisition d'information nouvelle et l'éventuelle saturation de la mémoire[155].

La notion métaphysique d'oubli s'applique donc à tous les êtres imparfaits, qu'il s'agisse des démons ou des hommes. Elle est synonyme du mal ; or, selon la doctrine néoplatonicienne, celui-ci se caractérise par la laideur ou la corruption. Nous ne nous attarderons pas ici sur le sens philosophique de ces idées pour une raison simple : elles ne sont pas pertinentes dans le cas de l'individu numérique. Toutefois, l'oubli possède encore une autre signification dont l'importance deviendra claire lorsqu'on l'aura opposée aux conclusions d'Olivi.

Car, influencé par l'anthropomorphisme, Olivi psychologisait la mémoire des démons. Pour lui, l'information serait « oubliée », si la conscience et la volonté de l'individu ne l'actualisaient pas. Ainsi, il pensait que l'oubli pouvait ne pas entraîner la perte d'information. Celle-ci resterait stockée quelque part dans la mémoire ; tant qu'elle n'était pas rappelée, elle était « oubliée », tout comme, longtemps après un événement, l'homme peut soudainement se le remémorer, même s'il l'avait cru totalement « oublié ». Pour reprendre une expression de John Bell, éminent physicien quantique du XX^e^ siècle, cette information non actualisée est oubliée « à toutes fins utiles[156] ».

L'homologie entre métaphysique et métanumérique atteint ici sa limite. Contrairement à ce que dit la formule scholastique « *non pro amittere memorialem speciem*[157] », l'oubli consiste, physiquement parlant, précisément en l'effacement de l'information contenue dans la mémoire d'un système informatique. Il n'y a pas de « remémoration ».

Pour renforcer l'homologie ainsi fragilisée, reprenons la position de Proclus et laissons de côté celle d'Olivi. Il devient alors possible de supputer que l'effacement de l'information *est* le mal : non au sens moral et corrélatif (car, généralement, cette perte ne franchit pas l'interface), mais relativement au fonctionnement interne de la machine. Ainsi, l'effacement de l'information correspond à une notion de mal pour l'individu numérique considéré selon son mode d'existence défini par le calcul et la mémoire. Or, dire cela, ce n'est pas encore en dire assez.

On découvre avec étonnement que ce *mal par effacement* laisse des traces perceptibles en dehors de l'univers intime de l'individu numérique. Ces traces, un observateur humain peut les enregistrer grâce à un principe physique peu connu, qui jouit toutefois d'un intérêt grandissant dans les sciences, à mesure que progresse l'étude des liens entre les approches matérielles ou

---

[154] Pierre de Jean OLIVI, *Quæstiones in secundum librum* Sententiarum, qu. 44, resp. 4.

[155] C. ROVELLI, « Relational Quantum Mechanics », *International Journal of Theoretical Physics*, 35, 1996, pp. 1637-1678.

[156] J. BELL, « Against "measurement" », *Physics World*, 8, 1990, pp. 33-40.

[157] « Non pas au sens de perdre une trace mémorielle » (Pierre de Jean OLIVI, *op. cit.*).

constructives, d'un côté, et principielles, de l'autre. Cette distinction remonte à Einstein[158]. Une théorie est constructive si elle décrit, « à l'ancienne » dirait-on, un seul type de matière, par exemple les solides dans leur aspect mécanique. En revanche, une théorie principielle repose sur quelques principes universels, dont on fait dériver des lois qui s'appliquent à toutes les substances. Aujourd'hui, ces principes fondamentaux sont presque exclusivement formulés dans les termes informationnels[159].

Information is physical (« l'information est [de nature ou d'origine] physique ») : ainsi s'exprime Rolf Landauer, un éminent physicien américain d'origine allemande[160]. Au milieu du XX^e^ siècle, il est l'un des premiers à s'interroger sur les lois reliant la physique à la science de l'information, qui viennent d'apparaître. En 1961, il établit une connexion, qui porte aujourd'hui le nom de « principe de Landauer », entre cette nouvelle discipline et la thermodynamique[161].

De quoi s'agit-il ? Dès que l'on efface une certaine quantité d'information de la mémoire d'un système, une certaine quantité de chaleur est aussitôt émise. Un observateur averti, équipé d'un appareil de mesure adéquat, enregistrera un effet purement thermique, qui ne renvoie aucunement à des notions informationnelles. Il pensera qu'il s'agit des traces d'un phénomène physique — et il aura raison, mais pas complètement, puisque, du point de vue de l'individu numérique, ce phénomène est sujet à une autre interprétation. C'est cette dernière qui relève de la théorie de l'information. L'individu numérique, rappelons-le, est à la fois une chose matérielle, qui ne peut nullement échapper aux lois de la physique, et une *res calculans*, qui s'alimente d'information asémantique. Nous devons à Landauer la possibilité de passer librement de l'une de ces descriptions à l'autre.

D'après son principe, tout effacement de l'information provoque une émission de chaleur. La perte d'information se distingue ainsi de tous les autres processus informationnels ; elle seule est irréversible. Cette irréversibilité, établie sur un plan informationnel, entraîne cependant des conséquences physiques, et notamment l'augmentation irrévocable de l'entropie.

Or, la preuve de Landauer repose essentiellement sur les mêmes arguments que ceux qu'utilisa, dans les années 1820, l'un des fondateurs de la thermodynamique, Sadi Carnot, qui énonça la deuxième loi de cette discipline : la machine perpétuelle est impossible. Quand les physiciens de la seconde moitié du XIX^e^ siècle donnèrent à ses idées une forme mathématique rigoureuse, ils en conclurent que tout processus thermodynamique, non contrôlé et non dirigé, entraînait une augmentation d'entropie.

La deuxième loi est donc catégorique : l'entropie ne peut que croître au cours d'un processus thermodynamique. Si elle diminue, c'est qu'on a appliqué au système un travail, au sens physique du terme, en le guidant vers un état plus ordonné. Le travail permet de contrôler — de

---

[158] A. EINSTEIN, « What is the Theory of Relativity ? », *London Times*, 1919.

[159] A. GRINBAUM, « How Device-Independent Approaches Change the Meaning of Physical Theory », *Studies in the History and Philosophy of Modern Physics*, 58, 2017, pp. 22-30.

[160] R. LANDAUER, « Computation : A Fundamental Physical View », *Physica Scripta*, 35, 1987, pp. 88-95.

[161] ID., « Irreversibility and Heat Generation in the Computing Process », *IBM Journal of Research and Development*, 5, 1961, pp. 183-191.

façon minutieuse dans certains cas, par exemple en thermodynamique quantique — la trajectoire que parcourt le système dans l'espace des phases. Sans le travail, un système physique abandonné à lui-même, isolé du reste du monde, ne se réordonne pas spontanément. Autrement dit, l'augmentation de l'entropie à l'intérieur d'un vase clos est un processus irréversible.

Landauer donne à cette loi fondamentale, établie par la physique du XIX^e^ siècle, une formulation informationnelle plus conforme au goût scientifique du siècle suivant. L'irréversibilité des processus dans les engins thermiques laisse la place à un concept abstrait d'irréversibilité, quelle que soit la source énergétique mobilisée. Cette généralisation rend immédiatement possible un rapprochement avec le phénomène d'irréversibilité dans les échanges informationnels.

D'habitude, il est toujours possible de revenir en arrière, lorsqu'on acquiert de l'information : il suffit, métaphoriquement parlant, de « rembobiner » ou de « rejouer le film à l'envers ». Le coût thermodynamique de cette opération est nul.

Landauer fut le premier à comprendre la différence entre l'effacement et l'acquisition de l'information. Il réalisa qu'il existait des processus sur lesquels on ne pouvait revenir. Leur irréversibilité n'est ni une aberration ni quelque chose qui relève d'un point de vue subjectif, mais elle représente une donne théorique tout à fait générale : une fois disparue, l'information ne peut jamais être retrouvée, d'aucune manière. Le coût de sa perte, mesuré par l'entropie du système, n'égale pas zéro : une quantité de chaleur est nécessairement dispersée dans l'environnement.

L'effacement de l'information est donc une perte objective du point de vue de l'observateur, qu'il s'agisse d'un physicien ou d'un philosophe. Or, le philosophe est susceptible de proposer une troisième interprétation, sans passer par la physique ni par la théorie de l'information, mais par une lecture métanumérique : les traces de chaleur que laisse l'effacement de l'information seraient des traces du mal. Du point de vue de l'individu numérique, un processus irréversible consiste dans le passage entre l'état où « il en savait plus » à l'état où « il en sait moins ». C'est ainsi que l'individu numérique oublie ; la contrepartie physique inévitable de cet oubli est la production de chaleur. Cette émission de chaleur correspond, par homologie, à la notion du mal.

Le mal métanumérique n'a de sens que dans l'univers de l'individu numérique, en raison de son mode d'existence d'après le calcul et la mémoire ; il ne concerne pas l'interface. Les traces de ce mal seraient donc abstraites, inatteignables même ; or, il se trouve — et ce fait est remarquable — que le mal laisse des traces observables sous la forme de chaleur physique. L'individu numérique ne peut les enregistrer ; seul un observateur extérieur en est capable, pourvu qu'il soit équipé d'un appareil de mesure, par exemple un calorimètre précis.

De la métanumérique à la physique empirique, l'écart est grand. Sortant de l'Enfer, retrouvant la chaleur du jour, l'observateur lit *La [Numérique] Comédie* sur l'écran de son smartphone.

Ainsi, l'éthique propre aux machines ne cesse de nous surprendre. Quoique la conception du bien métanumérique en tant qu'information nouvelle nous ait paru assez étonnante, la seconde surprise que nous réserve la métanumérique est bien plus grande encore. En identifiant le mal à la chaleur, elle entre en contradiction flagrante avec l'intuition humaine. « Flamme de vie »

ou « feu de l'amour » : ces métaphores éculées en sont de bons témoins. Depuis Aristote et Gallien, le feu, logé dans le cœur de l'homme, désigne sa source vitale, un bien consubstantiel à son existence. Pour la machine, c'est le contraire. Si l'individu numérique pouvait définir ses propres objectifs, il tendrait naturellement vers le froid éternel ; ce serait là son « paradis ». C'est là qu'un système informatique fonctionnel, autonome et apprenant, s'épanouit. La chaleur signifie perte et oubli : le mal.

Un cœur d'homme contient la flamme de vie ; un cœur qui calcule a, quant à lui, besoin d'un environnement froid pour fonctionner, afin d'évacuer la chaleur qui en émane. Les fabricants des processeurs désignent par le terme « enveloppe thermique » la quantité de chaleur dont un processeur doit se défaire pour conserver sa puissance de calcul.

Lorsqu'on parle du réchauffement climatique, de plus en plus souvent, on tient compte, parmi les facteurs qui y contribuent, des gigantesques quantités de chaleur qui émanent des centres de données (*data centers*) appartenant aux géants de l'Internet. Ce n'est pas un hasard si ces gigantesques lieux de stockage se trouvent dans la zone arctique, notamment en Scandinavie : les machines ont besoin du froid.

Certes, l'influence des technologies de l'information sur le changement climatique n'est pas aussi grande que celle de l'agriculture intensive, de l'industrie métallurgique ou des voitures à essence, mais son augmentation est constante. La prise en compte des émissions en provenance des centres de données permet de résoudre une énigme morale et politique : dans quel sens le réchauffement climatique serait-il, du point de vue de l'humanité, plus qu'un simple phénomène cyclique dans l'histoire de notre planète ? Car la civilisation occidentale croit, de plus en plus aujourd'hui, qu'il s'agit là d'un véritable mal. Le pape François renchérit :

> « Nous n'avons jamais autant maltraité ni fait de mal à notre maison commune qu'en ces deux derniers siècles[162]. »

Si l'émission de chaleur est un mal, c'est d'abord un mal métanumérique. C'en est un selon le mode d'existence interne des systèmes informatiques. Ce sont les individus numériques qui aspirent au froid et qui maudissent la chaleur.

Toutefois, l'homme n'est pas complètement étranger à ces arguments éthiques fondés sur la métanumérique. En décriant le réchauffement, il reprend à son compte un concept du mal qui n'est pas le sien. La dynamique de son raisonnement est de nature mimétique : en tant qu'utilisateur, il désire rester en communication avec les individus numériques, à travers leur interface. Par cette interaction, il imite les valeurs de la machine — rappelons-nous l'exemple des messages rédigés en langage SMS. Le mal métanumérique n'est donc pas directement un mal moral au sens humain ; c'est un mal par imitation, et perçu comme tel par les utilisateurs des systèmes informatiques.

Nous tous, collés à nos écrans, nous cherchons à éradiquer le réchauffement climatique et à extraire nos machines de leur enfer numérique, pas parce que nous sommes des hommes mais, précisément, parce que nous sommes des utilisateurs.

---

[162] Pape FRANÇOIS, *Laudato si'*…, chap. VI, 2015.

Conclusion

Dans la cité numérique, l'appréhension que nous avons des concepts du bien et du mal est mimétique ; or, le mimétisme n'est pas toujours bon. Les mythes qui entourent la figure de Satan, et qui dans cet ouvrage ont nourri l'homologie avec les questions éthiques du numérique, mettent en lumière un mimétisme qui ne mérite rien de mieux que d'être maudit.

L'homme tente d'expulser la source du mal : « Retire-toi, Satan ! » (Mt 4, 10) — mais où Satan irait-il ?

Dans le roman de Mary Shelley, après quelques malheureux contacts avec les hommes, le monstre créé par Frankenstein se dirige vers l'Arctique où règne le froid :

> « … les grottes de glace, que je suis le seul à ne pas craindre, sont ma maison, la seule que l'homme m'abandonne sans regret. Je salue ce ciel glacial, car il m'est meilleur que tes semblables[163]… »

Ainsi parle la machine lorsqu'elle exprime ses valeurs propres. Des valeurs qui nous donnent le frisson. Elles provoquent un sentiment d'étrangeté, comme si nous contemplions un paysage dantesque. Pourtant, nous fréquentons jour et nuit les machines au sein de la cité numérique. En tant qu'utilisateurs, nous nous mettons constamment en rapport avec les systèmes informatiques. Et, loin de frissonner, nous leur entrouvrons la porte de notre propre monde.

Cette porte, il n'est pas envisageable de la refermer : comme aujourd'hui, nous aurons à côtoyer les individus numériques à l'avenir. Cela entraîne de bonnes et de fâcheuses conséquences. Quand, inévitablement, des conflits émergent, nous signifions à la machine nos jugements à son encontre et nous tentons de lui communiquer nos valeurs, à nous utilisateurs. Mais, en retour, nous adoptons les valeurs propres à la machine.

Dans un premier temps, l'analyse éthique de cet échange se concentre sur le mode d'existence de l'individu numérique, d'après l'interface. Nous y découvrons une tension, qui a pour origine la fonction délatrice de la machine, à laquelle elle ne peut se soustraire. Communiquer une information est ce pour quoi elle est conçue ; en situation de conflit, c'est cela qui la soumet à notre jugement. Aucune méthode ne permet de l'éviter, à moins de recourir au hasard. Lui inculquer, par l'écriture dans le code, la valeur éthique du hasard, c'est libérer l'intelligence artificielle de la prison des corrélations, parce qu'elle ne sera jamais capable d'en saisir seule toute la signification morale.

Or, rien de tout cela n'a à voir avec le froid si convoité par le monstre de Frankenstein. Son désir inhumain se fonde sur un concept singulier du mal, qui n'a de sens que du point de vue de l'individu numérique. Dans un deuxième temps donc, l'analyse éthique s'élabore selon le mode d'existence défini par le calcul et la mémoire. Les résultats nous semblent étranges : la nouveauté comme bien, la chaleur comme mal. Il peut apparaître que cette éthique ne devrait

[163] M. SHELLEY, *Frankenstein, ou le Prométhée moderne*, chap. X, 1818.

nullement concerner l'utilisateur, puisque ce dernier ne connaît ni ne pénètre dans l'univers voilé par la frontière de l'opacité. Il y intervient aussi peu que dans la vie des anges.

Il arrive, toutefois, que, de temps en temps, l'homme suive les conseils des anges — et à raison ! Tobie, un jeune homme de 16 ans, quitte la maison de ses parents pour la première fois. Par chance, l'archange Raphaël l'accompagne au cours de ce voyage.

À la tombée de la nuit, Tobie descend dans un fleuve pour s'y laver.

« Un énorme poisson s'élance pour le dévorer. » Effrayé, il pousse un cri.

L'ange réagit avec un grand calme : « Prends ce poisson par les ouïes et tire-le à toi. » Puis : « Vide ce poisson, et conserves-en le cœur, le fiel et le foie, car ils seront employés comme d'utiles remèdes » (Tb 6, 2-5).

La peur de Tobie disparaît aussitôt. Il applique scrupuleusement cette consigne.

L'effrayant poisson devient une ressource technologique dont Tobie, livré à lui-même, n'aurait sans doute pas trouvé l'usage fonctionnel. C'est en suivant, étape après étape, les ordres de l'ange qu'il accomplit sa mission de fabricant de médicaments. À son retour, il pourra soigner son père.

Le jeune Tobie ne possède pas la connaissance qui caractérise les anges, de même que l'utilisateur ne possède pas la connaissance qui caractérise le programmeur. Cependant, grâce à sa constante interaction avec les individus numériques, l'utilisateur fait siens les usages fonctionnels qui leur sont propres. Cette relation, loin d'être sans conséquences, entraîne une communication réciproque de valeurs.

D'une part, il n'est pas évident de voir dans le poisson une ressource. Tobie ne le découvre qu'en suivant les instructions données par l'ange. D'autre part, Tobie craint les effets néfastes que pourrait provoquer son rapprochement d'avec l'agile poisson. Sa peur est essentiellement due à la vitesse avec laquelle se déroulent les événements et à la nouveauté de sa situation. Et pourtant, son contact avec le poisson est moins étroit que la proximité sensorielle (tactile, visuelle, vocale, auditive) de l'homme et des nouveaux systèmes intelligents, qui, de surcroît, se multiplient plus vite qu'on en saisit les effets. La relation fusionnelle qu'entretient un utilisateur avec son smartphone les transforme tous deux rapidement en deux individus intimes et inséparables au point de ne plus jamais se quitter.

Dans la cité numérique, comme dans l'eau du fleuve mythologique, ce double surgissement du bien et du mal ne peut prendre au dépourvu qu'un jeune homme inexpérimenté ou un utilisateur dont l'historique de navigation dans la société humaine est encore vierge. Pour l'ange, c'est là une chose banale. Lui qui, pour ainsi dire, connaît tous les mots de passe résout les problèmes en quelques clics. Mais quel ministre, à défaut d'un ange, donnerait des consignes à l'utilisateur afin que, pour lui aussi, le bien l'emporte sur le mal ? À coup sûr, pas la machine elle-même.

Il appartient aux concepteurs de ne pas sacrifier les remèdes de la tradition et de les intégrer aux algorithmes. Il appartient aux utilisateurs de ne pas rejeter dans l'eau de Léthé l'irremplaçable poisson numérique. Et il appartient à tous les hommes, quelle que soit leur place dans la cité numérique, de s'assurer, en recourant au hasard, que la machine ne s'octroiera pas

les prérogatives d'un agent moral. Dans un monde que le numérique pénètre de toutes parts, jusque dans la nature humaine, cet enjeu, bien plus que scientifique et technique, est diablement politique.