Note de lecture : Grinbaum (Alexei), Les Robots et le mal, Paris, Desclée De Brouwer, 2018, 216 p.

Thierry Ménissier

Thierry Ménissier

Institut de Philosophie de Grenoble, Université Grenoble Alpes

Note de lecture

Grinbaum (Alexei), *Les Robots et le mal*, Paris, Desclée De Brouwer, 2018, 216 p.

(Janvier 2020)

I

Prochainement, dans le cadre de la chaire « éthique & IA » (IPhiG/MIA, Université Grenoble Alpes), nous recevons Alexei Grinbaum autour de son récent livre *Les Robots et le mal*, ouvrage que je veux présenter en quelques pages.

L'auteur entend se situer sur un terrain difficile, celui de la relation entre les formes de discursivité très différentes que sont la rationalité typique de la technologie informatique, numérique et robotique, la théologie monothéiste (juive et chrétienne principalement, et parmi celle-ci les doctrines catholiques et orthodoxes principalement) et la philosophie morale. Ces genres de discours sont différents, ils apparaissent disjoints et potentiellement contradictoires. Mais, ordinairement et jusqu'ici, on pourrait affirmer que le lien entre eux existe du fait que l'humain se trouve à la croisée de ces différentes formes de connaissance du monde, en tant que ce dernier est tout à la fois : concepteur et usager des programmes informatiques, lui-même croyant ou tout au moins culturellement affilié à une tradition religieuse, et agent moral devant être responsable de ses propres actions devant la loi civile. Etre humain, c'est implicitement exister simultanément sur ces trois plans, et cela consiste même, en particulier lorsqu'on se trouve occasionnellement devant des cas de conscience, à faire dialoguer les genres de discours pour déterminer une décision satisfaisante (et en ce cas leur antagonisme potentiel peut devenir une ressource pour la délibération).

Le monde dont on entrevoit aujourd'hui l'aube apparaît différent, sous l'effet du développement généralisé des technologies informatiques, numériques et robotiques : par exemple, dans le transport sous la forme du véhicule autonome ou des drones de livraison, dans les questions de sécurité et de guerre avec les armes autonomes, et dans le monde des services en général avec les « agents conversationnels », autant de cas où des machines apparaissent susceptibles de *mal se comporter*. Il se produit comme une mise en tension entre les trois types de discours, du fait qu'on peut désormais se trouver dans l'obligation de déterminer la responsabilité de machines plus ou moins autonomes mais pas encore responsables, là où, hier, le droit se concentrait plus simplement sur l'imputation, à savoir, sur la relation entre d'une part les intentions et les actes des personnes humaines, et de l'autre leurs effets sur la société.

Plus précisément, l'ouvrage repose sur le constat d'une faiblesse dans la manière contemporaine d'aborder les problèmes. Cette faiblesse regarde le lien qu'on établit aujourd'hui entre les problèmes posés par la nécessaire maîtrise des machines (ou l'absence de maîtrise, qui engendre des accidents) et la régulation éthique. Cette dernière, telle qu'elle est habituellement conçue, regarde la calculabilité des opérations techniques et repose sur une valeur telle que la transparence. Or, l'une et l'autre s'avèrent insuffisantes : la seconde est souvent impossible à réaliser, la première ne dit pas l'essentiel de ce qui se passe avec un ordinateur. En effet, celui-

ci ne fait pas que *calculer* ; comme son nom l'indique, il met en ordre, il priorise, range et arrange le monde, par conséquent l'organise, l'ordonne et par-là finalement contribue déjà à l'instituer. Dans ces conditions, un ordinateur calcule si l'on veut, mais il procède un peu comme un créateur le ferait à l'égard de sa création, tout en étant un créateur non totalement doué d'intentions ou qui n'est pas l'auteur de ses propres intentions primordiales. L'ordinateur, créateur créé par l'humain, exécute les programmes prévus mais ne fait pas qu'appliquer des décisions toutes faites, en un sens il en invente déjà certaines. Puisque nous sommes d'ores et déjà quotidiennement ou presque confrontés de tels êtres, une approche éthique en de tout autres termes s'impose, d'urgence même vu le retard pris. Cette approche en de nouveaux termes, c'est celle de la métaphysique. Mais si les êtres dont il convient de traiter sont nouveaux, faut-il inventer une nouvelle métaphysique ?

Pour répondre à cette question, l'auteur propose de se pencher de manière attentive sur la manière de procéder typique du discours théologique. Il qualifie sa démarche d'« homologique » : il ne s'agit nullement de considérer l'informatique du point de vue de la théologie simplement entendue comme discours évaluatif, mais de trouver à penser quelque chose des difficultés rencontrées par la première par le moyen des catégories, termes, mythes et images inventés par la seconde. Ainsi, tout en invitant son lecteur à faire de même, entreprend-il d'identifier des thèmes issus de la tradition monothéiste qui peuvent être pertinents pour éclairer à nouveaux frais les problèmes aujourd'hui posés par l'informatique, le numérique et la robotique (dans un post suivant je me concentrerai sur l'analyse de certains arguments identifiés comme tels par l'auteur).

C'est ainsi que la notion de mal prend, par l'intermédiaire d'une lecture savante de Bible et de sa tradition, une dimension substantielle – le mal dont Grinbaum entreprend de penser la pertinence valable pour évaluer l'informatique et la robotique est bien plus « consistant » sur le plan éthique que les valeurs du « bon » ou du « mauvais » traditionnellement proposées par la *computer ethics* contemporaine pour déterminer si tel ou tel programme est éthiquement valable ou non. Il est permis de saisir le haut niveau de questionnement de l'auteur en évoquant les passages où, tout en concédant que les machines ne se substituent pas à nous en tant qu'agents moraux, il s'agit pour lui de déterminer « la question du mal pour l'individu numérique » et plus généralement celle du sens du mal pour l'IA. En d'autres termes, et l'on saisit ici l'ambition et l'originalité de la démarche, il ne s'agit pas de dire de l'humain, qui est pourtant un nouveau Prométhée capable de développer des technologies presque surhumaines, qu'il *fait le mal en voulant le bien*, ni des machines créées qu'elles *font le mal parce qu'elles sont imparfaites.* Il s'agit pour Grinbaum à la fois d'affirmer que dans la conception et l'usage de l'informatique, du numérique et de la robotique *il y a du mal*, de définir cette catégorie dans sa spécificité lorsqu'on la rapporte à ce type de savoir, et enfin d'envisager y remédier.

Evidemment, en poussant fort loin la comparaison entre le désir humain de maîtriser le monde par la connaissance calculatoire et la puissance contemporaine des informaticiens, le mythe de Faust s'avère fondateur pour la réflexion, ne serait-ce que par l'évocation du risque que prennent ces derniers. Si bien que la figure de Satan domine l'ouvrage, tant par la référence constante qui y est faite, dès l'introduction, que par les évocations des ouvrages qui, dans la civilisation chrétienne, l'ont de loin en loin nourrie (la tradition de Marlowe à Thomas Mann en passant par Goethe bien entendu, mais également *Le Maitre et Marguerite* de Mikhaïl Boulgakov, ou encore *Je vois Satan tomber comme l'éclair* de René Girard). Grinbaum rappelle que Satan, de quelque façon qu'on le nomme (Méphistophélès, l'Ennemi, le Malin, le Diable, le Démon, le Tentateur, etc.) constitue, pour la tradition biblique, une fonction fondamentale.

Nul ne peut dire qu'il a affronté la question du sens éthique de ce qu'il fait s'il n'a vécu l'épreuve imposée par Satan. Cette étape doit aujourd'hui est franchie par les programmeurs informatiques voués à « ordonner le monde », elle le sera immanquablement demain par les individus artificiels autonomes qu'ils sont en train de créer.

Après le droit canonique, la science juridique et administrative, puis la statistique, la finance et la science de gestion ont, au fil des siècles, prétendu valoir comme des discours fondateurs de la normativité sociale ultime, en entreprenant de régir les actions des humains par la connaissance de la Loi (celle de Dieu, celle du Pouvoir, qu'il soit royal ou impérial, républicain, financier et managérial). De nos jours, forte de ses performances techniques, l'informatique s'inscrit dans une comparable ambition, en leur imposant la Loi de l'information, elle-même issue des Faits eux-mêmes – encore que les faits purs sont bien difficiles à atteindre, car toute « donnée » est une « construite ». Même si elle n'affiche pas ouvertement cette ambition, elle revendique à son tour de valoir comme un discours instituant, un indiscutable foyer de normativité fondamentale. On peut le remarquer par bien des aspects, par exemple grâce à cette observation que l'informaticien se plait à la présence du juriste. Aux informaticiens, tous aujourd'hui implicitement placés dans une telle posture (et, pour certains, en revendiquant déjà la supériorité), l'ouvrage de Grinbaum rappelle que, si l'informatique avait jamais eu le dessein de se constituer explicitement en une telle doctrine tout en échappant aux apories de l'expérience vécue, aux ambiguïtés constitutives de l'existence humaine et aux problèmes indépassables de la nature des choses, c'est peine perdue. Au contraire, les grands ordonnateurs de nos nouveaux Golem, actuels maîtres de toutes choses ou de presque toutes, se sentiront à la lecture de cet ouvrage incontestablement moins seuls, mais, dans le même temps, ils apprécieront tout ce que la riche tradition de la philosophie morale héritée du monothéisme peut leur apporter en termes de réflexivité, c'est-à-dire de pensée non technique de leur propre action.

II

À cette nouvelle discipline canonique qu'est l'informatique, Alexei Grinbaum rappelle dans son ouvrage *Les Robots et le mal* que le monde sur lequel se déploie le pouvoir de ses algorithmes connaît des formes fondamentales de pensée qui recèlent de trésors d'intelligence pertinents pour les situations qu'elle a à affronter. La théologie monothéiste, de par les thèmes et les arguments qu'elle promeut en voulant justifier la révélation divine, les images et les qu'elle déploie dans des imaginaires cohérents, les termes mêmes qu'elle a utilisés au fil de son histoire, constitue pour l'informatique un riche répertoire pour sa propre inspiration. Moyennant la méthode appropriée pour adapter les deux types de discours (méthode dite « homologique »), ce répertoire s'avère utile car dans le fond les deux discours traitent, sinon des mêmes problèmes, du moins évoquent des situations comparables. Si dans la démarche homologique, certaines limites apparaissent parfois, dues à la particularité des systèmes technologiques contemporains par rapport aux concepts inventés par la théologie et si, de ce fait, il existe des cas où la nouveauté technique engendre un vide pour l'éthique et un trouble dans la compréhension métaphysique, l'intelligence théologique appliquée à l'informatique et à la robotique semble bel et bien féconde. Elle promet non pas de *dépasser* le plan empirique des questions posées par l'usage massif des machines intelligentes, mais *d'éclairer* ce plan par la méditation approfondie des relations de l'humain à son action via l'emploi d'outils théoriques appropriés car forgés au niveau qui convient. Ce niveau, c'est le plus haut possible pour la réflexion humaine : envisagé du point de vue de la révélation, c'est celui de la théologie, et, du

point de vue de la philosophie, celui de la métaphysique. Ce qu'il conviendra ensuite de déterminer, c'est en quoi le fait de débattre à un niveau aussi élevé d'abstraction apporte une plus-value aux chantiers de l'éthique appliquée à l'IA, c'est-à-dire, préciser dans quelle mesure les instruments élaborés dans les plus hautes sphères apportent un air nouveau et bienvenu aux problèmes rencontrés par l'action humaine.

En se donnant un tel dessein, Grinbaum procède à une sorte d'« épaississement » des concepts de bien et de mal, car il arrache la réflexion d'éthique appliquée à l'IA au conventionnalisme des valeurs qui accompagne nécessairement l'utilitarisme dominant la *computer ethics*. Rapportée aux notions théologiques, la question se reformule ainsi : puisque l'on parle du rapport entre le mal et les robots, en quoi s'avère pertinente l'homologie entre d'une part les figures prises par ce dernier dans la double tradition de pensée morale et, de l'autre, la conception contemporaine des algorithmes ?

Pour répondre à cette question et saisir la réponse qu'y apporte l'auteur, il convient de faire un rapide détour par la caractérisation du mal (figures du diable comprises), qui a été variée au fil de l'histoire de la théologie monothéiste et de la philosophie occidentale. Pour simplifier l'exposé, ainsi que l'a écrit Leibniz au début du XVIIIème siècle dans ses *Essais de théodicée sur la bonté de Dieu, la liberté de l'homme et l'origine du mal* (Livre I, § 21), le mal apparaît à la fois comme physique, moral et métaphysique.

Le mal *physique* comprend toutes les formes sensibles de la vulnérabilité humaine, il se manifeste toutes les fois que nous pouvons dire « j'ai mal » (aux dents, à la gorge, etc.). Ces formes ne sont pas toutes bénignes : par exemple, la maladie vécue constitue en effet une expérience fondamentale pour l'existence humaine, et, comme l'a écrit Marcel Conche, la souffrance des enfants, c'est-à-dire le mal enduré par les êtres innocents, peut valoir comme un argument en faveur de l'existence du « mal absolu » (*Orientation philosophique*, 1974, chap. 1).

Le mal *moral* comprend les formes passives et actives liées à l'engagement de l'humain dans son action, à savoir, les cas où notre volonté s'avère faible devant les tâches à accomplir ou les devoirs à remplir, où notre ambition ou notre orgueil sont disproportionnés, où nous commettons des fautes, et où de la méchanceté se révèle (nous agissons en sachant que nous faisons du mal). Ces formes ont été thématisées dans la tradition occidentale, qu'elle soit théologique ou métaphysique, sous la forme de la faiblesse de la créature en regard de l'omniscience et de l'omnipotence du Créateur : le mal moral trouve son origine dans le manque d'être de l'humain par rapport à Dieu, tant en matière de connaissance qu'en termes de volonté (un des fondateurs de cette formulation est Augustin d'Hippone qui dans ses *Confessions* dialogue avec l'Eternel à propos des fautes qu'il a commises durant son existence).

Le mal *métaphysique*, enfin, comprend quant à lui les formes plus élaborées qui, fort mystérieuses, laissent à penser qu'il y a davantage dans le mal qu'un défaut des créatures vis-à-vis du créateur. Cette dimension pointe par exemple la perversité humaine, qui peut être exprimée de manière mythologique à travers l'épreuve de la tentation ayant conduit à la Chute de l'humanité et à la perte de l'Eden (Genèse, 3), ou de manière philosophique par les concepts de « mal radical » (Kant, *La Religion dans les limites de la simple raison*, 1793) ou de « banalité du mal » (Arendt, *Eichmann à Jérusalem*, 1963). Elle évoque également le soupçon, à vrai dire irrépressible pour la conscience, que le bien n'est ou ne vaut rien, ou qu'il n'existe ni pour l'humain ni pour la nature telle qu'il est voué à la connaître. Le Livre de Job de l'Ancien

Testament a de manière puissamment suggestive évoqué cette possibilité, Leibniz a entrepris de la réfuter dans ses *Essais de théodicée*, et les existentialismes contemporains l'ont en revanche exploitée. A ce soupçon correspond la conscience aiguë du mal, tant de son irréductible persistance que de son infinie variété.

Dans la tradition monothéiste, les multiples figures de Satan expriment cette conscience. Il est évoqué comme un être puissant dont l'action est constante et invisible, qui tantôt inspire la volonté des humains afin qu'ils commettent des actes interdits par Dieu, tantôt opacifie la connaissance qu'ils pourraient avoir de l'enseignement divin et leur ment sur les fins qu'ils doivent poursuivre, tantôt enfin les met intérieurement dans le doute et les divise entre eux (*diabolos*, celui qui divise, est le nom grec de Satan). Dans tous les cas, son action est efficace du fait de l'imperfection de la connaissance humaine, si bien que de nombreuses élaborations théoriques du mal moral trouvent dans l'hypothèse du mal métaphysique leur condition de possibilité.

Que signifie mettre ce paysage théorique en relation avec les machines intelligentes contemporaines ? Plusieurs réponses peuvent être apportées à cette question, je vais successivement en mentionner trois que Grinbaum examine tour à tour.

D'abord, traditionnellement, les machines peuvent apparaitre diaboliques car elles ne sont pas directement les œuvres de Dieu, mais celles de son imparfaite créature, l'humain. Si bien qu'elles peuvent être dites diaboliques pour deux types de raisons : d'une part, elles sont artificielles donc non issues de la création divine originelle ; de l'autre, étant encore plus imparfaites que ce dernier, elles sont susceptibles de créer des catastrophes. C'est pourquoi les mythes anciens de Prométhée et du Golem, ou plus adéquatement pour aujourd'hui celui de la créature du Docteur Frankenstein (le sous-titre de l'ouvrage fondateur de Mary Shelley est « *The Modern Prometheus* ») constituent la base de la perception intuitive ou spontanée du problème moral posé par les machines. Le mythe fournit ici une sorte de connaissance culturelle spontanée de la situation, et offre une représentation des risques pris par l'essor d'êtres à la fois artificiels et animés. Pourtant, en dépit de son évidence première, il convient de dépasser ce mythe, car, sous-tendu qu'il est par la peur manifestée par les humains à l'égard des êtres artificiels qu'eux-mêmes engendrent, il opacifie le traitement du problème qu'il rend sensible. Si les machines intelligentes étaient vectrices du mal seulement à cause de notre propre peur à leur égard et de notre manque de compétence dans la conception, ce serait en effet peu de choses, ou bien peu dire.

Ensuite, elles peuvent apparaître diaboliques du fait que, improprement appelées « intelligentes », elles reproduisent les décisions humaines qui ne sont ni claires quant à leurs intentions ni morales quant à leurs finalités. On pourrait donc dire que les machines empruntent la voie satanique du « mauvais mimétisme », selon l'argument développé par René Girard dans le chapitre 3 de son ouvrage *Je vois Satan tomber comme l'éclair*, justement intitulé « Satan ». Mimétique, l'IA l'est en effet parce qu'elle nous invite à reproduire des actions pour le pire et non pour le meilleur, et c'est bien le cas en deux sens différents. Premièrement, elle paraît nous conduire au mauvais mimétisme parce qu'elle est caractérisée par la propriété de suivre ses propres fins de manière absolue, sans pouvoir modifier son programme. Grinbaum cite à ce propos un passage de Thomas d'Aquin : « Il revient [au diable] en vertu de sa nature propre de demeurer invariablement attaché à l'objet sur lequel porte sa volonté » (*De Malo*, 16, 5), et le commente ainsi : « 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 » (p. 181). Et elle l'est deuxièmement parce que, fonctionnant en mode fermé ou peu ouvert, elle invite les humains à un pareil mimétisme, de manière dégradante pour eux. Elle les incite en effet à reproduire sans cesse les mêmes actions, les enferme dans l'étroitesse de leur bulle numérique, les réduit à satisfaire leurs besoins élémentaires ou à demander de la jouissance primaire, elle leur ment en les conduisant à simplifier à outrance la complexité à la fois déconcertante et merveilleuse du réel, et par-là les voit amoindrir leurs capacités de résolution des difficultés et d'invention dans une société de plus en plus assistée et automatisée. Et cependant, au final, on pourrait dire que ce plan apparait lui aussi insuffisant et dépassable, car le caractère « diabolique » des machines ne tiendrait ici qu'à la limite actuelle de leur programmation ou qu'à leur faible degré d'ouverture aux apprentissages, et là qu'au manque de volonté des usagers humains.

Enfin, si les machines intelligentes peuvent être de manière pertinente rapportées à la notion philosophique de mal, et s'il est pertinent de confronter l'informatique aux notions et images théologiques, c'est du fait de leur fonctionnement même : dans la mesure où ce dernier tel qu'il est d'ores et déjà et sera *de plus en plus autonome*, c'est-à-dire fondé sur des facultés d'apprentissage certes programmées par l'homme mais lui échappant irrémédiablement. Sur cet aspect, la réflexion de Grinbaum prend une tournure intéressante et très originale de par le fait que la valeur véritablement éthique de l'IA est localisée par l'auteur précisément dans la possibilité d'« ouvrir » le système artificiel apprenant à l'aléatoire, qui combine l'accueil de la variété et de l'imprévu. Certes, d'un côté, le programme informatique peut être dit « satanique par conception » (p. 110), car il est normalement délateur, et aucune éthique de la transparence ne pourra jamais contrer efficacement ce travers ; mais d'un autre côté il possède une caractéristique qui l'exonère de cette tare initiale, il s'agit de sa capacité à saisir l'ouverture. Ainsi se comprend l'interprétation qu'il propose d'un passage de l'Ancien Testament tiré du *Livre de Josué* signalant la pertinence du tirage au sort pour réaliser la volonté de Yahvé : lorsque Dieu commande à Josué de « lancer les dés » afin de démasquer qui, parmi le peuple hébreu, a enfreint ses commandements, il convient de comprendre que le programme le plus éthique, celui qui souscrit le mieux à la valeur du bien, c'est celui qui est ouvert à l'aléatoire (chap. IV, « La valeur éthique du hasard », p. 111 sq.).

Il faut prendre la mesure du déplacement opéré par Grinbaum vis-à-vis de la conception habituelle de l'éthique de l'IA : prudentielle, celle-ci entend dé-biaiser les algorithmes et les rendre transparents. Selon l'auteur, « pour des raisons éthiques, le programmeur doit se donner pour mission d'inscrire dans le code le hasard explicite d'un tirage au sort », tout en assumant cette posture, aux antipodes de l'image reçue (et attendue) de l'informaticien dont la maîtrise lui permet de contrôler son programme (p. 154). Cette tâche est rendue nécessaire par la nature même de l'individu numérique, « système informatique autonome et apprenant » dont le chapitre V entreprend d'établir le profil métaphysique. La connaissance homologique permet en effet de statuer sur la nature profonde de cet individu et d'en saisir le caractère hybride : en s'inspirant du néoplatonisme (Plotin, Proclus), Grinbaum suggère que cet individu est en partie « matière », en partie « âme », et qu'il peut être qualifié d'« auto-indéterminé » du point de vue matériel, ce qui conditionne son statut d'apprenante. Dans cette perspective, l'auteur en appelle à la création d'une nouvelle discipline, qu'il nomme la métanumérique (dans un projet évoquant les *Prolégomènes à toute métaphysique future* de Kant, 1783) visant à « s'appuyer sur la pensée abstraite afin d'explorer les modes d'existence dans la cité numérique, par-delà l'existence communicative et corrélative » (p. 182). L'angle de vue particulier de cette nouvelle discipline offre l'espoir de décaler fondamentalement le propos tenu sur l'IA par des êtres humains qui ne

peuvent la comprendre à partir du langage ordinaire : il ne s'agirait plus de partir du point de vue de l'usager, considéré comme étroitement anthropologique et inapproprié à la nature authentique de l'être technique, mais de celui de la machine, en évacuant toute explication de type sémantique et téléologique, ainsi que les émotions humaines.

L'analyse métanumérique tend donc à abolir la relation entre la machine et l'usager, via l'interface, considérée comme « affaiblissante ». Or, parce qu'elle se veut réellement adéquate à son sujet, cette analyse vise à réinventer l'éthique de l'IA avec des catégories qui n'amoindrissent pas sa nature. A ce titre, elle délivre deux nouvelles valeurs de bien et de mal, pertinentes pour un individu numérique émancipé de l'interface avec l'usager. Le bien, c'est la nouveauté ou le renouvellement permanent de l'information ; le mal, comme l'information, saisie dans ses composants ultimes, est quelque chose de physique, c'est la chaleur produite par l'oubli, par l'effacement des données.

Je voudrais pour terminer souligner un point qui m'apparaît curieux et discutable, et émettre une critique à l'égard d'une démarche qui ne me convainc pas sur un aspect important.

Le point curieux et discutable réside dans ce qui peut sembler un véritable paradoxe. En effet, à rebours de la conclusion visant à « dé-sémantiser » l'éthique de l'IA, l'ensemble de l'ouvrage, bien sûr parce qu'il est écrit dans une langue sobre et précise mais surtout parce qu'il emprunte aux meilleures traditions philosophiques, théologiques et juridique, avait méthodiquement eu recours à la sémantique. Et on lui est reconnaissant d'avoir éclairé l'entendement des usagers. On comprend que l'ambition de l'auteur est de réformer l'éthique en la mettant au niveau de son sujet (dans les deux acceptations de ce terme). A cet égard, son ouvrage creuse la différence entre le programmateur et l'usager (cf. p. 57-63 qui consacre la puissance du premier sur la société : « 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 »). Mais ce faisant, il s'agit de rejoindre le point de vue de Dieu. Si la métanumérique, telle que Grinbaum en propose le cadre, apparaît comme une nouvelle doctrine canonique valable pour notre temps, c'est qu'elle se donne aussi pour le discours parfait de la connaissance suprême : puisque le Code est Dieu, la programmation vaut aujourd'hui pour ce que valait hier la Révélation.

Pour autant, compte tenu des effets du développement de l'IA dans la société, ou plus exactement des conséquences de ce développement sur des usagers toujours davantage soumis à une forme de marketing qui tend à obscurcir leurs facultés éthique et politique d'appréciation tout en les insérant dans des relations de plus en plus étroites avec les machines (relations dans lesquelles ils sont émotionnellement pris), on ne saurait exonérer les informaticiens du nouvel humanisme qu'ils doivent au contraire fonder. Si, en tant que philosophe, nous respectons et admirons les prouesses des mathématiciens auteurs du Code, nous ne pensons pas que la sémantique obscurcisse l'intelligence, bien que la mauvaise sémantique y contribue incontestablement, qu'elle soit par exemple ou trop catastrophiste ou naïvement enthousiaste à l'égard de l'IA ainsi qu'on le voit en ce moment par la publication d'essais tonitruants. Ce constat ne rend pas superflu le travail de sémantisation, mais au contraire oblige à se montrer plus précis et rigoureux avec les concepts, les arguments, les images et même les émotions des humains. Tel est le travail de pédagogie (c'est pourquoi j'emploie le terme d'humanisme) auquel doivent se livrer les détenteurs du savoir aujourd'hui canonique, sauf à s'enfermer dans

de nouveaux ordres monastiques coupés de la cité numérique dans laquelle ils entendent pourtant jouer un rôle, et même, à en écouter certains, préserver les valeurs de la démocratie.

La critique concerne la tradition de pensée qui, au moment de caractériser l'individu numérique, se trouve convoquée par l'auteur. La métaphysique occidentale est issue d'une double origine : la source religieuse de la Bible, la source philosophique de la pensée grecque. Souvent ces deux sources ont divergé, mais parfois elles ont convergé. Et tel est le cas dans la tradition que choisit Grinbaum puisqu'il construit sa conception métanumérique de l'IA à partir de la pensée néoplatonicienne. Celle-ci, issue des *Ennéades* de Plotin (traités d'ontologie et de théologie composés entre 254 et 270 de notre ère), se présente comme une hénologie, à savoir une doctrine de l'unicité sinon de l'univocité absolue, puisqu'elle procède de l'acte de contemplation ou « vision » du principe premier dont toutes choses existantes découlent. Un tel choix a des conséquences importantes, notamment parce qu'il tend à enfermer la réflexion dans une ontologie qu'on peut trouver trop « mystique » dans son mode d'appréhension de la vérité, particulièrement rigide dans sa définition de l'être premier et des ordres qui en découlent, ou encore très imperméable à ce qui fait la contingence et la variété des faits empiriques.

Il est permis de penser que la science métanumérique s'enrichirait d'une confrontation avec d'autres façons de pratiquer la métaphysique, dont je veux évoquer pour finir deux formes, car elles comprennent l'une et l'autre des concepts prometteurs pour poursuivre autrement la refondation dont l'éthique de l'IA a aujourd'hui besoin. Premièrement, la manière classique qui, de Malebranche à Leibniz, a conceptualisé les relations entre Dieu, le mal et l'humain à partir du dialogue entre l'intelligence, la volonté et la grâce – or, ce concept de *grâce* n'est-il pas prometteur pour une interprétation « homologique » de l'IA capable de saisir certains des effets produits par les apprentissages artificiels ? Et deuxièmement, la manière sceptique, qui, de Sextus Empiricus à Hume, a posé les conditions d'une analyse très fine des relations entre la pensée et les phénomènes, comme autant de nuances, toujours particulières et irréductibles les uns aux autres – ce concept de *nuance* n'est-il pas suggestif pour une conception renouvelée des formes du bien et du mal en informatique ?