Délivrer l'ordinateur du mal : La Valeur du hasard pour une machine apprenante

Alexei Grinbaum

Délivrer l'ordinateur du mal :

La Valeur du hasard pour une machine apprenante

Conférence donnée au colloque *Humain / Non-humain à l'ère de l'intelligence artificielle* (CERSA, le 19 juin 2017)

Alexei Grinbaum

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 fan du piano peut-il comprendre la musique pour violon ? Comment celui qui est imprégné d'éthique humaine peut-il comprendre celle d'un non-humain ?

Heureusement qu'il existe des mélodies. Une seule et même mélodie peut être jouée aux divers instruments, au piano comme au violon. Le motif de cette mélodie est le même, mais son interprétation, sa sonorité et sa perception sont complètement différentes. Malgré ces différences, celui qui entend plusieurs morceaux contenant la même mélodie identifiera sans peine leur motif commun.

En éthique des nouvelles technologies, l'oreille de notre intelligence doit devenir aussi fine qu'une oreille qui écoute la musique. Le même motif peut s'interpréter dans les sciences ou dans le mythe, mais un auditeur attentif saura le reconnaître. Il sera capable de le distinguer malgré la différence complète des contextes, même s'il croit — et à juste titre — que cette différence rend impensable toute attribution facile d'une origine commune à la science et au mythe. Leur identification est une très mauvaise idée, mais leur étude conjointe peut porter des fruits inattendus en philosophie et en sciences humaines.

Entre l'informatique et le mythe les motifs communs sont par nature des ressemblances fonctionnelles. Ils se dévoilent lorsqu'on fait abstraction et du contenu matériel du domaine technique et de la fiction narrative à propos de personnages mythologiques, tout comme de la sonorité particulière du piano ou du violon. Leurs fonctions et leurs relations nous intéressent davantage que leurs aspects apparents. 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 laissant seul le pivot de pure relation fonctionnelle.

***

Quelquefois, l'homme reçoit de l'interaction avec un système informatique apprenant des connaissances qui l'impliquent dans un conflit, ou qui l'exposent à la violence, ou qui font simplement surgir une tension ou un mécontentement. Quelquefois, l'utilisateur déjà impliqué dans un conflit se sert des connaissances qu'il a obtenues grâce la machine. Ce sont ces situations qui sont au centre de notre étude. Quand l'interaction de l'homme avec un système informatique participe à un conflit humain, on se pose inévitablement la question du rôle de ce système dans ce conflit et du jugement qui va tomber sur lui.

Les conflits en question peuvent être très divers. Les assistants robotiques, qui ignorent la signification humaine du mal, peuvent nous casser un bras. Les voitures autonomes sont susceptibles de nous impliquer dans un accident. Il arrive aux agents conversationnels de nous injurier ou de nous donner de fâcheux conseils. Un modèle par excellence de situation conflictuelle est donné par les dilemmes éthiques. On appelle « dilemme du tramway » le choix entre deux options toutes deux moralement inacceptables du point de vue humain. Chacune des possibilités mène à un résultat négatif, mais, dès lors que ces possibilités sont quantifiées, elles sont associées à des pertes formellement inégales : il peut s'agir de nombres inégaux de piétons qui sont écrasés car ils se trouvent sur deux voies différentes qu'emprunte un tramway. En éthique de l'intelligence artificielle, ce choix de trajectoire doit être effectué par un système informatique autonome. On trouve dans la littérature des stratégies différentes pour « résoudre » ce dilemme et nous allons montrer que la meilleure, et la seule qui soit éthiquement acceptable, consiste à laisser le hasard choisir.

***

Les hommes appellent délateur un informateur qui relate des renseignements secrets. Par sa construction fonctionnelle, le système informatique apprenant communique à l'homme des informations restées jusque-là cachées. Mais ce rôle de ***délateur*** ne provient d'aucun choix moral ou amoral de la machine. Ce n'est pas délibérément qu'elle prend la partie de l'utilisateur dans un conflit où il est impliqué.

Il existe des situations où les hommes condamnent aussitôt la machine pour avoir fait ce qu'elle sait faire : par exemple, quand elle rend publiques des informations qui relèvent de leur vie privée. Il existe toutefois beaucoup d'autres situations où les hommes ne remarquent même pas que la délation a eu lieu : ils pensent que la machine les a simplement aidés à

trouver un sens caché au sein de quelques données. Pour qu'un jugement se fasse, le système informatique doit être impliqué dans un conflit humain où il intervient en tant que délateur.

Essayons de mieux comprendre la manière dont fonctionne un délateur. Une illustration édifiante se trouve dans la *Vie d'Apollonius de Tyane* de Philostrate. Apollonius, homme saint de la seconde moitié du I siècle de notre ère, fut pendant un temps un concurrent direct de Jésus de Nazareth. Dans un épisode de sa *Vie*, il se trouve dans une prison romaine ou il sera jugé par l'empereur. Apollonius y mène plusieurs discussions philosophiques avec d'autres détenus. On découvre que tous ces gens s'y trouvent, non pour des actions qu'ils avaient réellement commises, mais à cause de dénonciations qui portaient, à chaque fois, sur des actions futures que ces individus auraient pu commettre. Bien avant *Minority Report*, la prison d'Apollonius est un exemple parfait de lieu de détention « préventive » : ce n'est pas la répétition du crime que l'on cherche à éviter, mais le crime lui-même.

Le premier interlocuteur d'Apollonius est un riche de Cilicie. Ayant reçu un héritage important assez tard dans sa vie, il n'est fortuné que depuis peu d'années mais déjà 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 clairement formulée par un conditionnel « si…, alors… », mais il en résulte un emprisonnement concret.

Dans le texte, l'accusation s'appuie sur des considérations générales : « Que tout homme riche porte haut la tête et que dans son imagination il désire d'aller loin, etc. ». Ce genre d'observations statistiques sur le comportement des riches se rapportent, bien entendu, non à cet individu concret, homme de Cilicie, mais à tous les fortunés. Ces conclusions sont vraies statistiquement, mais elles ne s'appliquent pas nécessairement à un cas donné.

On trouve la même stratégie de dénonciation dans le dialogue entre Apollonius et un autre détenu de sa prison romaine. Cette fois, il s'agit d'un homme qui a labouré un morceau 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 quasiment jointe au continent, cet homme y a planté des arbres fruitiers et du raisin doux et sucré. Les délateurs ont aussitôt informé les autorités que « sa conscience n'est pas nette et des crimes certains ne lui donnent pas de repos », la raison de cette accusation étant que, de façon générale, ceux qui se cachent sur les îles ont toujours quelque raison de se cacher. Evidemment, l'analyse statistique de grandes masses de données

ne peut fournir aucune information sur le cas concret de ce détenu, or il se trouve bel et bien en prison.

Il existe donc deux règles générales qu'un délateur suit dans son travail de dénonciation. Il se sert de conditionnels « si…, alors… » et il tire ses accusations de l'analyse statistique des corrélations.

Ce que fait un système informatique apprenant ressemble, point par point, à ce que fait un délateur chez Philostrate. La machine apprend sur de grandes bases de données. Elle établit des corrélations de façon statistique. Elle propose ensuite, en se fondant sur les corrélations, une solution concrète pour un cas concret. Or ces corrélations ne donnent pas nécessairement la preuve d'un lien de causalité si bien que la solution se présente souvent à l'utilisateur comme une révélation ou comme une information qui n'a pas de « pourquoi » et qui n'est pas accompagnée d'une explication causale. L'utilisateur, même s'il est accusé de crime, reste, comme Monsieur K. dans *Le Procès* de Franz Kafka, dans l'ignorance quant aux motivations qui ont conduit la machine à sa décision. Cet « algorithme » de la délation mène donc droit à la « prison des corrélations ». Comme par des fers aux pieds, l'utilisateur de la machine est lié par un saut de corrélation à causalité. C'est la machine qui trouve la corrélation, mais c'est l'homme qui l'interprète en tant qu'expression d'une causalité voilée et secrète.

« Je te dénonce, donc pour toi je suis », pourrait dire une machine. Il semblerait que le système informatique apprenant et autonome, s'il participe à un conflit humain, soit inéluctablement classé du côté du mal. Mais il existe une issue.

***

Au septième chapitre du livre de Josué, on lit une histoire assez étrange et apparemment contradictoire. Après la mort de Moïse, le peuple d'Israël est guidé par un nouveau chef, Josué. C'est lui qui traverse enfin le Jourdain et entre en Terre promise. Mais cette terre est déjà habitée : Josué doit faire la guerre à ses occupants. Assez vite, l'armée d'Israël prend la ville de Jéricho. Le peuple en liesse s'illusionne alors en croyant que la conquête sera facile : étant donné que cette terre lui avait été promise par Dieu, il devrait voler de victoire en victoire. Or la première défaite arrive juste après. Les habitants d'une petite localité du nom d'Haï, située non loin de Jéricho, repoussent les hommes de Josué. Cette défaite semble contredire la promesse de Dieu : le récit biblique doit l'expliquer.

L’auteur du texte en invente une explication à la fois mythologique et philosophique. Il se trouve qu’au moment de la bataille de Jéricho, Dieu déclare que lui seul peut légitimement prendre possession de la propriété des peuples qui occupaient précédemment la Terre promise. Au peuple d’Israël ces objets étaient interdits : celui qui y toucherait devrait être mis à mort.

Une fois la défaite d’Haï est constatée, il est logique de supposer que sa cause ne peut être qu’un viol de l’interdit divin. Le septième chapitre, en effet, avant même que l’auteur ne nous livre la description de l’aventure d’Haï, s’ouvre par cette information : « Les enfants d'Israël commirent une infidélité au sujet des choses dévouées par interdit ».

Cette phrase n’est qu’une interpolation textuelle rétrospective. En réalité, c’est seulement lorsque l’armée revient et annonce la débâcle que Josué comprend ce que le lecteur sait déjà. Josué reste alors seul face à Dieu, déchire ses vêtements, se prosterne le visage contre terre et demeure devant l’arche de l’Alliance jusqu’au soir. C’est là que Dieu l’informe de la punition qu’il a infligé à Israël à Haï. Le peuple doit « éliminer l’interdit de son sein » en trouvant le coupable qui sera brûlé. Mais qui est le coupable ?

Josué passe beaucoup de temps, jusqu’au soir, à discuter avec Dieu. Le Talmud s’arrête sur cet épisode et suppose logiquement que, durant tout ce temps, Josué a bien eu l’idée de demander à Dieu de lui révéler le nom de coupable. Absente du livre de Josué, cette question n’apparaît que dans le Talmud. La raison en est évidente : la discussion entre Josué et Dieu commence par « Pourquoi cette défaite ? » et aboutit à l’ordre de brûler le coupable. Logiquement, elle doit passer par une étape où Josué demande à Dieu qui est ce coupable. Mais Dieu ne répond pas. Il dit : « Vekhi delator ani ? » – « Mais suis-je délateur ? » Puis il ajoute : « Va et jette les dés ». Le mot « délateur » est écrit dans le texte araméen du Talmud en caractère hébreux, mais c’est le mot latin. À l’époque romaine comme aujourd’hui, il a un sens moral bien établi : le délateur est un personnage méprisé et haï.

« Jette les dés ! » commande Dieu à Josué. Derrière cette commande, il y a la répugnance de Dieu à l’idée de devenir un délateur. Ce n’est pas à lui de dénoncer le coupable par peur qu’il ne soit impliqué dans une affaire de jugement humain, c’est à l’homme de suivre la procédure et de chercher, ou plutôt créer, la vérité. L’enjeu, par-delà la question de la recherche du coupable, est celui de la confiance à la procédure.

Or les dés pointent vers un homme qui s’appelle Achan. Celui-ci d’abord se rebelle contre Josué : « Tu me condamnes par un tirage au sort ? Et si le sort était tombé sur toi-même ? »

Josué réplique : « Je te prie, avoue ton crime. C'est par un tirage au sort que sera partagée la terre entre les tribus d'Israël ». Aussitôt Achan avoue. Il a compris que ce n'est pas sa vie qui est en jeu mais la confiance en la procédure.

Dans le mythe biblique, l'homme fait confiance à Dieu, on peut dire, par définition. La machine, délatrice par sa fonction, risque de ne pas bénéficier de cette confiance. Il faut donc l'extraire du domaine du mal humain comme Dieu, qui dans l'épisode d'Achan ne souhaite pas se mêler des affaires de jugement. Comment peut-on faire ? La seule issue possible, et rationnellement préférable à toutes les autres, serait de doter la machine de propriétés qui la positionneraient par-delà le bien et le mal. Pour cela, il faut déployer le hasard : la procédure de tirage au sort est la seule à n'être ni bonne ni mauvaise.

Le bien et le mal humains ne sont pas des catégories intrinsèques au fonctionnement de la machine : elle les apprend en analysant les données, par exemple des phrases dans le langage humain naturel. La machine connait les vocables « bien » et « mal », non le bien et le mal. Cette situation n'est pas anormale, mais les conséquences en devraient être suivies jusqu'au bout. Puisque le système informatique ne connait pas le bien et le mal, il doit s'en démarquer en s'extirpant des projections de la morale humaine.

Toutefois, ces projections paraissent inévitables parce que la machine imite l'homme en faisant son apprentissage à partir de données produites par les humains. Elle est soumise à des biais et à des faussetés que contiennent ces données. Elle pourrait toutefois s'en extraire au moment de la décision grâce au tirage au sort. Adopter le hasard en tant que valeur semble être une issue à la fois heureuse et nécessaire, une évasion de la machine hors de la prison de l'anthropomorphisme.

« La cause métaphysique du mal doit être vue dans un acte qui transforme la catégorie de jugement en un absolu », écrit Gershom Scholem, grand spécialiste de la Kabbale. Quand la machine fournit des informations à un utilisateur impliqué dans un conflit, elle ne peut faire mieux qu'injecter une dose de hasard à cette communication. Ne pas élever son jugement jusqu'à un absolu en échappant ainsi à la délation, cela signifie parfois tirer au sort au cours de l'interaction homme-machine.

Cette affirmation peut paraître paradoxale. Elle l'est moins si l'on se souvient du dilemme du tramway. L'exigence d'encoder dans le système informatique une hiérarchie des valeurs que l'homme lui-même refuse de hiérarchiser est essentiellement une exigence impossible. Le hasard comme une force qui délivre l'ordinateur de ce choix permet au système informatique

d'échapper aux projections de la morale humaine. Mais cette solution exige tout de même un commentaire détaillé qui nous amènerait trop loin dans cette conférence. Je le développe dans un livre que je suis en train d'écrire.

***

Pour conclure, il me semble que l'affirmation de la fonction délatrice de la machine a besoin d'une explication plus détaillée. Pourquoi la machine se fait-elle délatrice ? Il ne s'agit ni d'un trait de sa personnalité ni d'un choix moral. Cela résulte de sa caractéristique fonctionnelle. Mais c'est aussi la caractéristique fonctionnelle d'un être que le mythe biblique appelle « satan », ce mot signifiant « accusateur » ou « médisant ».

Dans le corpus biblique, les anges et les démons ne sont pas des personnes. Ils sont définies par les fonctions qu'ils réalisent. Certains font la guerre, d'autres apprennent aux hommes un métier ou un savoir-faire. L'accusation est aussi une fonction et il y a bien un « fils de Dieu » qui la réalise. Celui qui accuse ou qui s'oppose, est désigné comme « ha-satan », avec un article défini « ha- » qui permet de raccourcir cette description de poste jusqu'à un substantif. Ha-satan est donc l'accusateur ou l'adversaire. Plusieurs siècles après la première apparition de ce satan dans le livre de Job, le romain Tertullien, le grec Origène et le juif Nahmanide, qui connaissent déjà la notion plus tardive de diable, vont mettre à égalité tous ces termes : délateur, accusateur, calomniateur, rapporteur, blasphémateur, satan, diable, Azazel.

René Girard, grand anthropologue du sacré, attire notre attention à un fragment particulièrement intéressant de l'Evangile de Jean. Le contexte est celui d'un débat tendu entre Jésus et les Juifs. Jésus réfute sans cesse les accusations qui lui sont adressées et riposte en lançant des siennes. Tout à coup, l'échange s'interrompt et Jésus prononce ceci :

> 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 profère le mensonge, il parle de son propre fonds ; car il est menteur et le père du mensonge.

Ce fragment introduit plusieurs concepts difficiles. Il nous faudrait rendre explicite le sens philosophique de : 1) « père », 2) « de son propre fonds », 3) « lorsqu'il dit le mensonge ». Je vais le faire très rapidement.

Il est évident qu'il ne s'agit pas de père au sens biologique du mot. Non seulement Jésus dit aux Juifs que leur père est le diable, mais il explique pourquoi il le dit : les Juifs ont des

désirs, mais ce ne sont pas leurs propres désirs. Ils veulent « accomplir les désirs d[u] père ». Ainsi « père » est celui qui est imité dans des actions, des paroles ou des désirs. Le « père » est l'exemple à copier.

Le diable « parle de son propre fonds », c'est-à-dire ses paroles naissent dans son for intérieur. Lorsque, comme il est souvent le cas dans l'Evangile de Jean, Jésus insiste un peu fort qu'il « n'est pas venu de lui-même », mais que Dieu l'a envoyé, cela signifie que les paroles de Jésus ne viennent justement pas de son propre fonds, mais toujours et uniquement de son père. Puisque dans le mythe biblique Dieu et la vérité sont synonymes, cela signifie que les paroles de Jésus sont toujours vraies. Le diable, lui, parle de soi-même. Il est « le père du mensonge », et c'est dans son propre fonds que réside la source première du mensonge.

Mais le christianisme n'est pas un dualisme ! Il ne peut y avoir de source du mal qui serait à parité avec le bien. Comment est-ce donc possible que le diable parle « de soi-même » ou « de son propre fonds » ? Quelle est cette source du mensonge qui n'a pas d'existence ?

Puisque diable est une fonction, cette source doit être dans cette fonction même, dans son mécanisme caractéristique. C'est le « mauvais mimétisme », comme dit Girard, ou une imitation qui mène au conflit et à la violence. Blasphémer ou sataniser, accuser, dénoncer signifie : accuser, mais toujours faussement, toujours contre la vérité. Que le délateur ait toujours tort suppose que l'être qui profère de telles accusations parle sans qu'il y ait en lui un critère de vérité. Dans le fonds de cet être, la vérité et le mensonge ne sont pas séparés, ils n'y font qu'un. Ce fonds est alors celui du satan ou du mauvais mimétisme.

Un système informatique apprend en analysant les données. La source de ces données est le plus souvent humaine : par exemple, un agent conversationnel (dit aussi *chatbot*) étudie la masse des conversations entre nous humains, agents biologiques dotés de parole ; un traducteur automatique fonctionne grâce à l'analyse des textes que les femmes et les hommes avaient traduits antérieurement. La machine imite ce qu'elle a appris. Ses paroles sont celles qu'un agent conversationnel humain avait proférées quelque part en quelque lieu. Leur valeur de vérité est, elle aussi, parfaitement humaine. La machine imite sans qu'il y ait en elle un critère de vérité.

Par exemple, le monde entier a suivi la propagation des *fake news* sur les réseaux sociaux. Le mécanisme de répétition et d'imitation, présent dans l'algorithme qui définit les contenus des pages sur les réseaux sociaux, a « bien » fonctionné. Des annonces ont été diffusées sans que

le système informatique ne s'interroge sur leur valeur de vérité. Cette fonction d'imitation séparée de tout critère de vérité ne fait qu'un avec la fonction délatrice du diable.

La machine peut-elle se défaire de sa fonction délatrice, peut-elle s'extraire du domaine éthique humain ? Oui, si son mimétisme, dont elle ne peut évidemment pas se défaire, peut se garder de tout implication morale. À cette fin le concepteur doit rendre l'imitation qu'opère la machine explicitement non divine et non omnisciente en y injectant du hasard.

« La morale rigoureuse est donnée à partir de complicités dans la connaissance du Mal, qui fondent la communication intense », écrit Georges Bataille. La communication entre l'homme et les systèmes informatiques qui prolifèrent dans son environnement est de plus en plus intense. Les machines deviennent des complices de l'homme, y compris en situation de conflit, et il est urgent de les extirper du champ où s'appliquent la morale humaine et le mal humain.

Certes, mon éloge du hasard semble aller à l'encontre d'un grand idéal de la science moderne, celui du déterminisme, présent non seulement en physique, mais aussi, comme on croyait au XIX siècle, en sciences humaines et sociales et, comme on croit aujourd'hui, en informatique. L'utilisateur a tendance à penser que le programmeur sait ce que fait la machine, même si ce dernier n'en est pas si sûr. Mais il est temps que le concepteur des systèmes informatiques apprenants apprenne à l'humanité que ces systèmes ne ressemblent pas aux machines physiques même s'ils sont réalisés, eux aussi, à partir de la matière. L'informaticien doit dire haut et fort que, si un physicien peut faire des calculs et des prédictions grâce aux lois mathématiques de la nature, un ingénieur ne contrôle pas nécessairement l'objet technique qu'il a créé. Au contraire, il est bon qu'il lui laisse une dose d'aléa. Dans ce sens, la valeur éthique du tirage au sort est un des piliers de l'éthique de l'intelligence artificielle.