Justice et intelligence artificielle

Éric Delassus

Dans la lettre au Marquis de Newcastle du 23 novembre 1646, Descartes affirme que ce qui nous permet de faire la distinction entre un automate (c'est-à-dire une machine autonome) et un être humain doué de pensée, c'est la capacité que possède le second de parler et de s'exprimer « à propos », c'est-à-dire de tenir un discours qui soit en phase avec celui d'un autre sujet conscient, de tenir un propos dont le sens entre en résonance avec la signification de celui tenu par une autre personne[1].

Au XX^{e} siècle, Alan Turing, l'un des pères de l'informatique moderne, considère que l'on pourra considérer que l'intelligence d'une machine aura atteint un niveau comparable à l'intelligence humaine, lorsqu'un être humain dialoguant avec une machine ne sera pas en capacité de savoir s'il a affaire à un programme d'IA ou à l'un de ses semblables.

Il semblerait qu'aujourd'hui le progrès des nouvelles technologies invalide la thèse cartésienne et que le test de Turing confirme que la confusion est possible entre une IA et une IH. Les possibilités de l'IA sont tellement étendue désormais qu'un programme de ce type est capable d'apprendre (deep-learning, machine-learning) et de se corriger, voire de prendre des décisions, comme c'est le cas dans les domaines médicaux (diagnostic, prescription) ou juridique (avec entre autres ce que certains nomment la justice prédictive). On est donc en droit de s'interroger quant à savoir si l'IA ne va pas finir par remplacer l'IH et si nous n'allons pas devoir, à plus ou moins long terme, nous soumettre à sa logique et à son mode de fonctionnement.

Face aux innovations technologiques, deux attitudes sont fréquentes : la technophilie et la technophobie. La première consiste à se laisser fasciner par la puissance de la technique et à s'imaginer qu'elle va pouvoir résoudre tous les problèmes de l'humanité. Certains ne vont-ils pas aujourd'hui jusqu'à s'imaginer que la technologie

---

[1] « *Enfin il n'y a aucune de nos actions extérieures, qui puisse assurer ceux qui les examinent, que notre corps n'est pas seulement une machine qui se remue de soi-même, mais qu'il y a aussi en lui une âme qui a des pensées, excepté les paroles, ou autres signes faits à propos des sujets qui se présentent, sans se rapporter à aucune passion. »,* Descartes, *"Lettre au marquis de Newcastle du 23 novembre 1646.*Descartes, "Lettre au marquis de Newcastle du 23 novembre 1646"

va nous permettre de bientôt vaincre la mort ? La seconde attitude, la technophobie, est son parfait inverse, puisqu'elle consiste à s'imaginer que la technique va finir par nous envahir et nous dominer pour faire des hommes les esclaves des machines et nous conduire à une sorte de meilleurs des mondes digne des fictions les plus pessimistes de type Matrix.

Si ces deux attitudes sont plus fantasmatiques que réalises, elles doivent cependant nous interpeller et nous inviter à réfléchir sur la nature de notre rapport à la technique et plus particulièrement aujourd'hui aux nouvelles technologies qui ont de plus en plus tendance à venir supplanter l'action des hommes, sous prétexte qu'elles permettent d'éviter tous les biais liés au facteur humain. Cependant, évacuer l'humain de l'organisation de nos existences, qu'elles soient individuelles ou sociales, n'est-ce pas le risque du système technique dans lequel nous sommes plongés.

Aussi, la question se pose de savoir quelles conséquences pourrait avoir le recours à l'intelligence artificielle dans des domaines comme le droit et la justice. La question qui se pose alors est celle que soulèvent Antoine Garappon et Jean Lassègue dans leur livre *Justice digitale*[2] : accepterions-nous d'être jugés par des algorithmes ?

En effet, la question centrale ici est certainement celle du jugement. Or, qu'est-ce que juger sinon appliquer une règle ayant un certain caractère de généralité à un cas singulier. Autrement dit, appliquer ce qui a été édicté pour tous les cas à un cas qui n'a pas son pareil. C'est là d'ailleurs toute la difficulté de juger, car comme le précise Kant, le jugement ne s'enseigne pas, il ne peut que s'exercer[3]. On ne peut, en effet, enseigner la règle qui permet d'appliquer toutes les autres, car cela nous conduirait à cette aporie logique qu'est la régression à l'infini étant donné que l'on pourrait toujours rechercher la règle qui permet d'appliquer la règle qui permet d'appliquer toutes les autres et ainsi à l'infini. Il faut donc si l'on veut apprendre à juger exercer son jugement et c'est à force d'exercice que celui-ci s'affine, se précise et devient

[2] Antoine Garrapon et Jean Lassègue, *Justice digitale*, PUF, Paris, 2018.

[3] « Le jugement est un don particulier qui ne peut pas du tout être appris, mais seulement exercé. Aussi le jugement est-il la marque spécifique de ce qu'on nomme le bon sens (Mutterwitzes) et au manque de quoi aucun enseignement ne peut suppléer ; car, bien qu'une école puisse présenter à un entendement borné une provision de règles, et greffer, pour ainsi dire, sur lui des connaissances étrangères, il faut que l'élève possède par lui-même le pouvoir de se servir de ces règles exactement, et il n'y a pas de règle que l'on puisse lui prescrire à ce sujet et qui soit capable de le garantir contre l'abus qu'il en peut faire quand un tel don naturel manque*. C'est pourquoi un médecin, un juge ou un homme d'Etat peuvent avoir dans la tête beaucoup de belles règles de pathologie, de jurisprudence ou de politique, à un degré capable de les rendre de savants professeurs en ces matières, et pourtant se tromper facilement dans l'application de ces règles, soit parce qu'ils manquent de jugement naturel, sans manquer cependant d'entendement et que, s'ils voient bien le général in abstracto, ils sont incapables de distinguer si un cas y est contenu in concreto, soit parce qu'ils n'ont pas été assez exercés à ce jugement par des exemples et des affaires réelles. Aussi l'unique et grande utilité des exemples est-elle qu'ils aiguisent le jugement. », (Kant, *Critique de la raison pure*, Traduction Tremesaygues et Pacaud, Puf, p. 148.149.

progressivement de plus en plus pertinent. Néanmoins, il n'est jamais parfait et peut toujours contenir de manière aléatoire une marge d'erreur plus ou moins importante.
On pourrait donc en conclure que si le jugement est porté par une application d'intelligence artificielle le risque d'erreur s'en trouvera limité, voire aura totalement disparu. Ainsi, les facteurs affectifs, émotionnels, idéologiques qui peuvent interférer dans un jugement humain, ne viendront pas parasiter le jugement de la machine, puisque l'on suppose que cette dernière ne ressent rien.
C'est apparemment ce qui se passe avec la justice prédictive qui, si je suis bien renseigné n'est pas encore autorisée en France, mais commence à l'être outre-atlantique. L'exemple le plus souvent cité est celui la décision que doit prendre un juge de mettre un prévenu en détention provisoire ou de le laisser en liberté. Si c'est une personne humaine qui prend la décision, il est fort probable que, selon le juge, la décision ne soit pas la même. En revanche, si l'on recourt à un programme d'intelligence artificielle dédié à ce type de tâches, il suffira qu'il étudie statistiquement la dangerosité de la personne concernée et les risques que celle-ci ne se présente pas devant ses juges lors de son procès pour qu'en découle une décision vierge de tout biais lié au facteur humain. On pourrait donc croire que la décision prise par l'application d'IA est plus « objective ». Mais est-ce vraiment le cas ? Et même si cela était possible, est-ce souhaitable ?
Il semblerait que les logiciels d'IA présentent aussi certains biais tout à fait comparables à ceux qui peuvent être rencontrés dans les décisions humaines. C'est ce que laisse entendre une étude de 2018 concernant le logiciel *Compass* utilisé aux États-Unis pour déterminer les risques de récidives. Apparemment, les décisions prises à l'aide de ce logiciel seraient aussi fiables que celles qui seraient prises par des non-experts répondant à un questionnaire en ligne[4]. D'autres étude auraient mis en évidence un biais raciste de ce logiciel, comme si la programmation avait été contaminée par les préjugés du programmeur. L'algorithme aurait en quelque sorte hérité des défauts du jugement humain. Néanmoins, le cœur du problème ne se situe peut-être pas à ce niveau. On pourra toujours rétorquer que l'application peut être corrigée, voire qu'elle peut se corriger elle-même grâce au deep-learning ou au machine-learning. Il n'y aurait donc qu'à parfaire l'application d'IA pour régler le problème.

[4] Science Advances (http://dx.doi.org/10.1126/sciadv.aao5580).

Cependant, la difficulté est peut-être d'une autre nature. On nous présente souvent l'IA comme une sorte d'optimisation de l'intelligence humaine, sous-entend par là que ces deux intelligences fonctionneraient de la même façon. Au point d'ailleurs que certains vont jusqu'à comparer le fonctionnement de notre cerveau à celui d'un ordinateur. Or, il n'est pas vraiment certain qu'il en aille ainsi.

Comme nous l'avons souligné précédemment, juger consiste à appliquer une règle générale à un cas singulier. Il faut insister ici sur le terme singulier qui est à distinguer du particulier. On entend par particulier ce qui caractérise les parties d'un tout. Ainsi, dans un ensemble, tous les éléments sont particuliers, ce qui n'empêche pas qu'ils puissent tous être parfaitement identiques. En revanche, le singulier désigne ce qui n'a pas son pareil, il faut donc pour juger le singulier une part de créativité et d'inventivité dans le jugement dont peut difficilement faire preuve une machine. On pourrait, certes, nous rétorquer que cette prise en compte de la singularité du cas est effectuée par la machine puisque toutes les données le concernant sont traitées par l'algorithme qui doit prendre la décision. Cependant, la manière dont sont analysées ces données n'a que peu de rapport avec la manière dont fonctionne l'esprit humain dans la prise de décision. Un être humain ne décide pas simplement à partir de statistiques. Il se réfère également à son intuition et à ce que lui inspire la relation justement singulière qu'il entretient avec la ou les personnes qu'il a en face de lui. Il ne faut donc pas confondre ici le risque et l'incertitude. Le risque renvoie à une notion statistique objective tandis que l'incertitude est, quant à elle, plus subjective puisqu'elle renvoie à notre ignorance face à l'avenir. Or si je peux connaître le risque de récidive d'un condamné, je ne suis pas pour autant certain qu'il passera nécessairement à l'acte. Il faut donc introduire dans la décision qui sera prise à son égard une part d'intuition. En d'autres termes, la décision en ce domaine ne pourra pas se limiter à un pur est simple calcul, elle consistera en un véritable jugement, c'est-à-dire, selon la définition qu'en donne l'économiste Franck Knight dans son livre *Risk, Uncertainty and Profit*, à une appréciation subjective et circonstanciée. Et le fait que cette appréciation soit subjective ne doit pas ici être considérée comme un défaut, car c'est cette subjectivité qui fait toute l'humanité de la décision.

Lorsque l'on rentre dans une application d'I.A. toutes les données concernant une personne - afin de déterminer, par exemple, son risque de récidive - on réduit cette personne à un objet soumis à un déterminisme des plus simples. Cette personne serait quasiment déterminée à récidiver comme l'eau est déterminée à bouillir lorsque

l'on élève sa température à plus de 100°. Qu'il y ait un certain déterminisme présidant au comportement des êtres humains, ce n'est pas cela qui doit être remis en question. Il s'agit plutôt de savoir si ce déterminisme peut être mesuré par un algorithme, eu égard à sa complexité due principalement au fait que l'on affaire à des sujets conscients.

En effet, la relation entre le ou les juges et le justiciable n'est pas une relation de sujet à objet, elle se tisse entre des sujets, entre des êtres conscients, sensibles et interagissant. C'est donc dans un cadre intersubjectif que cette relation s'inscrit, ce qui n'a rien à voir avec la relation homme/machine. On pourra toujours dire que la machine et l'algorithme ne font qu'apporter une indication au juge et qu'il n'est pas nécessaire qu'il soit contraint de s'y soumettre, il n'empêche que le recours à la technologie interfère ici dans la relation. Ce n'est pas parce qu'une personne est statistiquement jugée dangereuse qu'elle l'est nécessairement. La décision du juge peut très bien déterminer le comportement de cette personne statistiquement à risque et l'inciter à modifier son comportement par souci d'être fidèle à la confiance qui lui est faite. Bien entendu, il peut aussi trahir cette confiance et ne pas respecter ses engagements, mais toute décision en la matière ne comporte-t-elle pas toujours un risque ? On pourrait alors rétorquer que le recours à l'I.A. permet justement d'éliminer toute forme de risque. Mais ce souci d'éliminer toute forme de risque de la décision n'est-il pas illusoire ? D'autant qu'il ne faut pas confondre le risque au sens statistique et le risque au sens moral, c'est-à-dire le pari que l'on peut faire sur autrui.

En se fiant uniquement à une décision fondée essentiellement sur des statistiques, ne prend-on pas également le risque ne pas permettre à une personne de s'amender et peut-être même d'augmenter sa dangerosité ? La relation entre le juge et le justiciable, comme toute relation humaine d'ailleurs, comme tout rapport à l'altérité, ne peut reposer que sur la confiance ou la défiance. On ne connaît jamais l'autre, on ne peut que croire ou ne pas croire en lui. C'est la présence ou non de la foi en l'autre qui oriente les décisions que nous prenons à son égard et qui engage notre responsabilité. Or, il est fort douteux que l'on puisse être fidèle à une machine ou que l'on puisse lui inspirer ou non confiance.

C'est donc en ce sens qu'intégrer la justice dans le système technicien en recourant à l'I.A. risque fort de la déshumaniser.

De plus, cette déshumanisation risque d'être encore plus profonde dans la mesure où, si on donne trop d'importance à l'I.A. dans l'acte de dire le droit, on modifie

radicalement le procédé par lequel celui-ci va être rédigé et donc codé. En effet, et c'est la première critique que font Garapon et Lassègue à la justice digitale. Cette critique souligne en effet que le code graphique par lequel le droit va se dire et s'écrire ne sera plus maîtrisé par le juristes eux-mêmes. Selon ces auteurs la révolution que fait subir au droit l'I.A. est une révolution graphique. Ce n'est plus simplement avec le langage ordinaire et l'alphabet que seront formulées les décisions de justice, mais à partir d'un code que seuls les informaticiens maîtrisent. Et encore ! Avec le deep-learning et le machine-learning, il est fort probable qu'au bout d'un certain temps, on ne sache plus trop ce qui se passe dans la machine. À cela,il faut ajouter le fait que si l'algorithme est propriétaire personne d'autre que la société qui dispose de la licence ne peut le modifier, ce qui signifie donc que les juristes qui l'utilisent n'ont pas la main sur le code par lequel est dit le droit.

Cela dit, il ne suffit pas de diagnostiquer les problèmes que pose l'introduction de l'I.A. dans le domaine juridique, il faut aussi s'interroger sur la meilleure attitude à adopter face à une évolution contre laquelle il est difficile de s'opposer. La solution ne se situe certainement pas dans une posture radicalement technophobe qui relèverait plus du déni de réalité que d'un réel esprit critique. L'I.A. est là et on ne plus faire sans elle, on ne peut pas se contenter de la rejeter ou de faire comme si elle n'existait pas. On ne peut pas plus désinventer le feu que l'I.A. Il faut donc nous efforcer, à la lumière d'une réflexion critique sur son usage, de concevoir une utilisation de l'I.A. qui tout en permettant de corriger certains biais humains ne conduirait par vers d'autres biais propres à l'utilisation de l'I.A. Nous touchons là au cœur du problème que pose à l'être humain son intégration dans un système technicien dont il est l'initiateur, mais qu'il ne contrôle pas. Si, en effet, le progrès technologique résulte de notre maîtrise de la nature et de ses lois, nous ne sommes pas toujours en mesure de maîtriser cette maîtrise et d'anticiper sur les conséquences qu'elles peut entraîner.
Il est donc crucial que les juristes, mais on pourrait dire la même choses des médecins, des enseignants et de tous ceux qui vont avoir à collaborer de manière de plus en plus étroite avec l'I.A., interrogent leur rapport à l'I.A. pour éviter que celle-ci ait pour conséquence une évacuation quasi totale de la dimension humaine de leur pratique.
Nous avons pu souligner qu'en cherchant à évacuer les biais humains dans l'exercice de la justice, on en arrivait à dériver vers d'autres biais propres à l'usage de L'IA.

Peut-être faut-il en conclure que jamais on ne pourra évacuer le risque des décisions humaines et donc des décisions de justice qui reposent sur une responsabilité qui ne peut être délégué à aucun dispositif technique.

Comme l'écrivent Antoine Garapon et Jean Lassègue :

> L'humanité suppose un possible, quand la technique se présente comme une certitude ; la justice doit être défendue comme un risque à une époque qui n'a de cesse de vouloir les réduire5.

[5] Antoine Garrapon et Jean Lassègue, *Justice digitale*, *op.cit.,* p. 348.