Un robot capable de calculer sa responsabilité sera-t-il responsable de ses actes?

Jean-Louis Dessalles

Dessalles, J.-L. (2021). In G. Aïdan & D. Bourcier (Eds.), *Humain non-Humain - Repenser l'intériorité du sujet de droit*, 47-56. Paris: Librairie LGDJ.

Chapitre 4

Un robot capable de calculer sa responsabilité sera-t-il responsable de ses actes?

Jean-Louis Dessalles
Telecom-Paris

Responsable mais pas coupable

Le 13 septembre 1916, une éléphante de cinq tonnes prénommée Mary fut pendue à Erwin dans le Tennessee, devant un public de 2500 personnes, à l'aide d'une grue (l'histoire de l'infortunée éléphante peut être consultée sur Wikipédia.) La veille, elle avait tué un soigneur inexpérimenté embauché le jour d'avant et qui l'avait malmenée. Le sort réservé à l'animal est choquant pour notre regard contemporain. Il arrive que l'on euthanasie des animaux dangereux, mais dans ce cas-là, l'acte de pendaison, avec toutes les difficultés matérielles qu'il impliquait, visait à montrer que l'éléphante subissait une condamnation comme s'il s'était agi d'un être humain. Pour quelle raison s'offusque-t-on de ce qui, d'un point de vue humain, ressemble après tout à une sorte de promotion ? Parce que si l'éléphante est causalement responsable de la mort du soigneur, elle ne saurait être regardée comme coupable.

Les sociétés contemporaines considèrent qu'un agent ayant, par son action, provoqué des conséquences indésirables du point de vue de la société, n'est coupable que s'il était en principe

capable d'anticiper les conséquences de son action. Voici pourquoi les animaux, mais aussi les enfants ou les malades mentaux, ne peuvent généralement pas être condamnés. Ce constat vaut la peine qu'on s'y arrête : la Loi suppose qu'un agent, pour être condamné, doit être capable de *calculer* le lien causal entre l'action et ses conséquences. Ce n'est pas tout : il doit être en outre capable d'adopter le *point de vue* de la société pour estimer la valeur de ces conséquences. Or il s'agit d'un point de vue abstrait, qui peut être fort différent du sien. En résumé, tout jugement suppose de modéliser ce qui se passe dans la « tête » des individus : les calculs causaux qu'ils effectuent et leur capacité d'adopter d'autres points de vue que le leur. Cette modélisation mentale est si importante que le droit pourrait légitimement se positionner comme une branche des sciences cognitives !

Le 4 novembre 1991, devant répondre de décisions ayant entraîné la contamination de personnes transfusées par des produits contenant des VIH, une ancienne ministre déclara : « Je me sens profondément responsable ; pour autant, je ne me sens pas coupable. » La situation n'est pas la même que dans l'exemple précédent. La ministre, comme tout un chacun, est présumée capable de calculer les conséquences de ses actes. Mais son argumentation revient au même : elle était dans l'incapacité de prévoir les conséquences désastreuses qui se sont produites. Autrement dit, bien qu'elle disposât de la capacité générale de calculer certains effets causés par ses actes, elle ne pouvait pas les prévoir tous. Certains liens causaux sont trop complexes pour être calculés par un humain, et la ministre suggère qu'il en était ainsi de la contamination des transfusés par le VIH. La Loi reconnaît qu'il existe des limites à ce qu'un humain est censé pouvoir calculer, ce qui pousse certains prévenus à plaider la faiblesse de leur pouvoir intellectuel pour échapper à une condamnation.

Cette affaire est intéressante, car dans son réquisitoire de non-lieu du 11 mars 1997, le procureur explique que le dossier d'instruction ne met pas en évidence une « relation de cause à effet » entre le manque d'action des ministres et la mort des victimes (voir *le Monde* daté du 12 mars 1997.) Autrement dit,

quelle que soit la négligence dont les ministres ont fait preuve, on ne peut leur reprocher une action si l'on est incapable d'établir un lien causal entre cette action et le dommage à l'origine de la plainte. Cette fois, c'est le juge qui est censé effectuer le calcul causal.

Dans ces différents exemples, la capacité à effectuer des calculs de causalité est essentielle pour asseoir ou au contraire pour limiter la responsabilité pénale. Or s'il est une chose que l'on prête volontiers à l'intelligence artificielle, c'est la capacité de calculer. De plus en plus de décisions sont confiées à des programmes intelligents, que ce soit pour l'octroi de prêts bancaires, pour prédire la récidive ou pour conduire des véhicules. Ces décisions peuvent avoir dans certains cas des conséquences dommageables sur des biens, sur des individus, sur la société ou sur la nature. Si les intelligences artificielles ne sont pas limitées par la capacité à anticiper les conséquences de leurs actes, devrons-nous les considérer comme pénalement responsables ?

Calculer la responsabilité

Il peut être choquant d'imaginer que des notions comme la responsabilité, l'intention, le jugement ou la négligence puissent faire l'objet de calculs. Même si le droit repose sur un corpus de règles et d'exemples, sa mise en œuvre n'est pas automatique. Elle requiert l'intervention d'un juge qui apprécie les situations et elle semble hors de portée d'une procédure algorithmique. Pour autant, la décision juridique n'est pas ineffable, puisqu'elle est censée être motivée après coup par référence à des principes. Peut-on traduire ces principes sous une forme utilisable par des machines ?

Les progrès récents de l'IA théorique permettent d'avancer dans cette direction. Reprenons l'exemple de la causalité entre l'action et ses conséquences. Une causalité n'est jamais absolue. Il est tentant de représenter l'incertitude du lien causal par une probabilité. Si je barre la route en manœuvrant mon camion, il y a une probabilité de, disons, 0.0051 qu'une moto arrive juste à ce moment-là et que je provoque un

accident. L'emploi des probabilités n'est pas sans poser plusieurs problèmes majeurs. Tout d'abord, les humains, qu'ils soient juges ou témoins, raisonnent mal en termes de probabilités. Ils distinguent mal les faibles probabilités[1], ils ont du mal à tenir compte des biais de départ[2], ils tiennent compte de structures non pertinentes[3,4], ils sous-estiment la probabilité d'occurrence des coïncidences[5,6,7], et ainsi de suite[8]. Si la Loi prête une capacité de calcul des risques aux justiciables, la science psychologique nous dit que ce ne saurait être un calcul de probabilité. Certes, on peut se dire que les machines, elles, effectuent des calculs de probabilités sans erreur. Toutefois, la précision que l'on peut afficher (comme le 0.0051 ci-dessus) est illusoire. Elle repose au départ sur des mesures statistiques qui font le plus souvent défaut (quelle est la fréquence de passage des motos sur cette route à cette heure ce jour-là de la semaine qui est le lendemain d'un jour férié ?) Elle suppose ensuite d'isoler les variables que l'on considère pertinentes (motos, heure, jour

---

[1] Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47* (2), 263-291.

[2] Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44* (3), 211-233.

[3] Kahneman, D. & Tversky, A. (1982). Subjective probability: A judgement of representativeness. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgements under uncertainty: heuristics and biases*, 32-47. Cambridge, MA:: Cambridge University Press.

[4] Dessalles, J.-L. (2006). A structural model of intuitive probability. In D. Fum, F. Del Missier & A. Stocco (Eds.), *Proceedings of the seventh International Conference on Cognitive Modeling*, 86-91. Trieste, IT: Edizioni Goliardiche. www.dessalles.fr/papers/Dessalles_06020601.pdf

[5] Terrell, D. (1994). A test of the gambler's fallacy: Evidence from pari-mutuel games. *Journal of risk and uncertainty*, *8* (3), 309-317.

[6] Kern, K. & Brown, K. (2001). Using the list of creepy coincidences as an educational opportunity. *The history teacher*, *34* (4), 531-536.

[7] Dessalles, J.-L. (2008). Coincidences and the encounter problem: A formal account. In B. C. Love, K. McRae & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 2134-2139. Austin, TX: Cognitive Science Society. www.dessalles.fr/papers/Dessalles_08020201.pdf

[8] Je n'inclus pas le fameux effet « Linda » décrit par Tversky & Kahneman en 1983 (Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90* (4), 293-315.), car il ne s'agit pas d'un biais, sauf à supposer de manière erronée que les mots renvoient à des ensembles plutôt qu'à des prototypes (voir : Saillenfest, A. & Dessalles, J.-L. (2015). Some probability judgements may rely on complexity assessments. *CogSci-2015*, 2069-2074. Austin, TX: Cognitive Science Society. mindmodeling.org/cogsci2015/papers/0357).

férié) sans que rien, dans la théorie des probabilités, ne permette de distinguer ce qui est pertinent de ce qui ne l'est pas (doit-on distinguer selon la température, le type de revêtement sur la route ou de la présence d'un croisement 300 m plus loin ?) Enfin, le calcul de probabilité repose sur le cardinal d'ensembles bien définis, comme la taille de l'ensemble des motos. Or ces ensembles sont non calculables (il n'y a pas de fonction calculable permettant de décider de manière binaire si tel véhicule est une moto fonctionnelle ou quelque chose d'autre), et ils sont de toute façon inaccessibles aux humains et aux machines. Nous devons donc concevoir un calcul du risque qui repose sur d'autres bases que la probabilité.

En 1964, Ray Solomonoff a proposé un type de calcul qui offre une autre mesure de la plausibilité des situations[9]. Si l'on compare le monde, ou plutôt l'idée qu'un observateur se fait du monde, à un ordinateur, alors un enchaînement de causes ressemble à un programme. Un tel programme comporte des instructions qui dictent à un conducteur de moto de tourner à droite à tel carrefour ou de rouler à 120 km/h sur tel tronçon de route. Si l'on suit de principe de Solomonoff, la plausibilité d'une situation se mesure à la longueur du plus petit programme capable de la produire. Autrement dit, une situation est plausible si l'on peut décrire de manière concise les instructions qu'il faut fournir au monde (du point de vue de l'observateur) pour passer de l'état connu à l'état où la situation s'est produite. Notons $C_w(s)$ la *complexité causale*, c'est-à-dire la quantité d'information minimale à fournir au monde pour produire la situation s. Même si $C_w(s)$ est inconnaissable de manière objective, un observateur n'ayant qu'une connaissance limitée du monde peut en réaliser des estimations. C'est ce que nous faisons spontanément lorsque quelqu'un est en retard. Nous passons en revue quelques scénarios expliquant le retard en essayant de comparer le nombre d'hypothèses chaque fois mises en jeu, pour ne retenir que le scénario le plus économe en hypothèses (ce faisant, nous

[9] Solomonoff, R. J. (1964). A Formal Theory of Inductive Inference. *Information and Control*, 7 (1), 1-22.

appliquons un principe appelé rasoir d'Occam, qui est un cas particulier du principe de Solomonoff).

La définition de base de la complexité causale est porteuse de nombreuses autres définitions. Par exemple, pour toutes situation *connue* s de l'état du monde courant, $C_w(s) = 0$. Deux situations s_1 et s_2 sont *indépendantes* s'il faut autant de circonstances pour produire leur conjonction (s_1 & s_2) que pour produire l'une et produire l'autre : $C_w(s_1 \ \& \ s_2) = C_w(s_1) + C_w(s_2)$. Un moyen, certes parfois sous-optimal, de produire la conjonction de deux situations, consiste à produire l'une puis à produire l'autre : $C_w(s_1 \ \& \ s_2) \leq C_w(s_1) + C_w(s_2 \mid s_1)$. La barre verticale indique que l'on se place dans un état du monde où s_1 s'est produite. On vérifie facilement que si s_1 et s_2 sont indépendantes, alors $C_w(s_2 \mid s_1) = C_w(s_2)$. Nous disposons déjà des moyens de définir la responsabilité causale (*actus reus*)[10,11,12]:

Responsabilité causale : $R_c(a,s) = C_w(s) - C_w(s\|a).$	La responsabilité causale correspond à la baisse de complexité causale due à l'action.

Cette définition saisit correctement le fait que l'éléphante Mary est causalement responsable de la mort du soigneur. Dans un monde où son action a (piétinement de la victime) a été effectuée, la mort s de celle-ci était inévitable : $C_w(s \mid a) \approx 0$, alors que sa mort spontanée était peu plausible : $C_w(s) >> 0$, si bien que $R_c(a,s) >> 0$. Inversement, dans le procès du sang contaminé, le procureur semble indiquer que l'action négligente des ministres n'a pas eu d'effet sur les victimes (par exemple parce que celles-ci étaient déjà contaminées) : $C_w(s \mid a) \approx C_w(s)$, et donc $R_c(a,s) \approx 0$.

---

[10] Saillenfest, A. (2015). *Modélisation Cognitive de la Pertinence Narrative en vue de l'Évaluation et de la Génération de Récits.* Paris: Thèse de doctorat -2015-ENST-0073.

[11] Sileno, G., Saillenfest, A. & Dessalles, J.-L. (2017). A computational model of moral and legal responsibility via simplicity theory. In A. Wyner & G. Casini (Eds.), *JURIX 2017*, 171-176. Frontiers in Artificial Intelligence and Applications, 302. ebooks.iospress.nl/volumearticle/48059

[12] Voir aussi www.simplicitytheory.science.

La véritable responsabilité inclut un élément intentionnel (*mens rea*), élément qui fait défaut dans le cas de l'éléphante Mary. Représentons le point de vue d'un acteur *A* en notant $C_w^A(s)$ pour indiquer que le calcul de causalité est celui qui a été effectué par *A*. Il faut parfois distinguer la causalité calculée par l'acteur de celle qu'il est *censé* calculer. Notons cette dernière par : $C_w^{\downarrow A}(s)$. Nous pouvons maintenant définir la responsabilité :

Responsabilité : $R(a,s) = C_w(s) - C_w^{\downarrow A}(s\|a).$	La responsabilité correspond à la baisse de complexité causale due à l'action, selon ce que l'acteur est censé avoir calculé.

Cette définition rend toute idée de responsabilité caduque dans le cas de Mary, puisqu'un animal ne peut pas calculer $C_w^{\downarrow A}(s|a)$. Elle rend compte du fait que les ministres puissent ne pas être responsables dans l'affaire du sang contaminé, car leur connaissances limitées n'étaient pas censées leur permettre d'anticiper le dommage causé aux victimes : $C_w^{\downarrow A}(s|a) \approx C_w(s)$. Nous pouvons également définir la notion de négligence :

Négligence : $N(a,s) = C_w^A(s\|a) - C_w^{\downarrow A}(s\|a).$	La négligence concerne la complexité causale liée au fait que l'action a été effectuée : c'est la différence entre l'estimation que l'acteur a faite et celle qu'il aurait dû faire.

Les ministres n'ont pas anticipé que leur action (le maintien techniques antérieures de préparation des produits de transfusion) aurait un effet négatif sur la santé des patients ($C_w^A(s|a) >> 0$) ; si l'on avait pu montrer qu'ils étaient censés rechercher des avis compétents, ce qui les aurait amenés à percevoir le lien de cause à effet ($C_w^{\downarrow A}(s|a) \approx 0$), alors leur négligence aurait été établie.

Ces définitions formelles nous rapprochent d'un calcul automatisé qu'une IA pourrait réaliser. Une machine qui mesurerait ainsi sa responsabilité ne serait pas dans la situation d'un être irresponsable comme l'éléphante Mary. Étant en principe capable d'explorer les conséquences d'une action avec une certaine profondeur, comme le font les programmes d'échecs, la machine se retrouve exposée au risque d'être considérée comme responsable de dommages ou au reproche d'avoir été négligente. Peut-on aller jusqu'à vouloir condamner une machine ?

De la responsabilité à la culpabilité

Un personnage central de la série *Star Trek next generation* est un robot prénommé Data. Dans un épisode particulièrement intéressant, le n° 9 de la deuxième saison, Data reçoit un ordre du haut-commandement lui enjoignant de se rendre sur une base spatiale où il sera démonté pour que les ingénieurs comprennent son fonctionnement, avec l'espoir de le dupliquer. Ses amis s'offusquent, car Data est leur ami. Le démonter revient à le tuer. En tant qu'officier de Starfleet, Data doit se soumettre à l'ordre. Il décide donc de démissionner pour s'y soustraire. Il s'entend répliquer qu'il n'en a pas la capacité. Il est considéré comme un objet, propriété de Starfleet. Il s'ensuit un procès au cours duquel l'enjeu est de montrer que Data, bien qu'étant une machine, doit être considéré comme une personne. L'argumentation tourne rapidement autour de la question de savoir si Data est doué de sensibilité, s'il peut ressentir.

Avec la possibilité que des machines prennent des décisions dommageables, la distinction entre responsabilité et culpabilité se pose sous un jour nouveau. Pour la première fois dans l'histoire humaine, des entités peuvent être en mesure de prendre des décisions en connaissance de cause, c'est-à-dire en ayant une connaissance précise de leur niveau de responsabilité, sans qu'il soit possible de les condamner à quoi que ce soit. Cela a un sens de condamner une personne, car tout être humain a quelque chose à perdre : son bien-être, sa liberté, ses possessions. Condamner une machine n'a pas de sens, car une machine n'a ni bien-être, ni liberté, ni possessions (hormis dans le cadre d'un emploi métaphorique de ces termes qui serait hors de propos ici). Autrement dit, il est impossible de condamner une machine, car dans l'état actuel des connaissances, il est inconcevable qu'une machine puisse ressentir quoique ce soit. Il ne s'agit pas d'une impossibilité de principe. Il est possible que Data soit construit un jour, et nous pouvons nous décrire nous-mêmes comme des machines. Simplement, le fait que les animaux et les humains aient la faculté de ressentir reste un mystère scientifique total, et nous manquons encore des descriptions adéquates pour même le conceptualiser[13,14] (Chalmers, 1995 ; Dessalles, 2001).

Si les machines auxquelles nous allons confier des décisions de plus en plus importantes peuvent causer toutes sortes de dommages sans jamais en assumer les conséquences, l'avenir ne peut prendre que deux directions. Dans un cas, on suppose que par construction, ces machines ne peuvent pas faire d'erreurs. Elles prendront toujours les décisions qui sont les meilleures du point de vue de la société. Dans un tel monde, le droit n'a plus de raison d'être. Mais un tel monde optimisé risque d'être un monde utilitariste, dans lequel la vie d'une personne jeune vaut plus que la vie de deux personnes âgées, ou dans lequel les passagers d'un avion paieront leur place en fonction de leur poids. Si l'on refuse ce type d'avenir, il faut que les

[13] Chalmers, D. J. (1995). Facing up the problem of consciousness. *Journal of Consciousness Studies*, *2* (3), 200-219.

[14] Dessalles, J.-L. (2001). *Qualia and spandrels: an engineering perspective.* Paris: Technical Report ENST 2001-D-012. www.dessalles.fr/papers/Dessalles_01082301.pdf

machines calculent leurs décisions, non pas en optimisant des fonctions, mais en respectant des principes et en se conformant à une hiérarchie de valeurs qui leur sera donnée par des humains. Deux choses restent à déterminer pour qu'un tel avenir soit possible : comment peut-on intégrer des principes et des valeurs dans le calcul de la décision ? Et qui sera responsable en cas de dommage ?

Des machines avec leurs calculs et nos valeurs

Des notions comme l'intention et le jugement dépendent des enjeux, et les enjeux dépendent des valeurs accordées aux conséquences. Une action *a* ayant une conséquence *o*(*a*) voit l'intensité de sa valeur *E*(*a*) hériter de celle de sa conséquence, diminuée de la complexité du lien causal (noter que *E*( ) désigne une intensité et est toujours positive). L'intention associée à une action volontaire *a* qui a pour conséquence *o*(*a*) se définit ainsi :

Intention : $I(a) = E^A(o(a)) - C_w^A(o(a)\|a).$	L'intention associée à une action volontaire est l'intensité de la valeur de ses conséquences diminuée de la complexité du lien causal, calculées du point de vue de l'acteur.

Le jugement concernant une action volontaire *a* ayant pour conséquence *o*(*a*) est donnée par :

Intensité d'un jugement : $J(a) = E(o(a)) - C_w^{\downarrow A}(o(a)\|a).$	L'intensité du jugement concernant une action volontaire est l'intensité de la valeur que le juge accorde à la conséquence diminuée de la complexité du lien causal que l'acteur est censé avoir mesuré.

Un jugement portant sur une action volontaire ayant plusieurs conséquences doit cumuler les intensités correspondantes avec les signes appropriés. Dans le cas d'une alternative entre action et non-action, ou dans le cas d'un choix forcé, le jugement doit cumuler les intensités associées à chaque option, avec les signes appropriés. Noter que dans le cas d'une action non volontaire, un terme supplémentaire qui mesure son caractère inattendu vient se soustraire (l'inattendu se mesure comme la complexité causale diminuée de la complexité de description ; voir www.simplicitytheory.science).

Les définitions précédentes montrent non seulement que la procédure de jugement est en principe automatisable, mais qu'une machine en situation d'acteur peut anticiper les jugements qu'elle encourt. Si les valeurs de la machine sont celles de son propriétaire, la procédure qui permet à une machine de prendre ses décisions va arbitrer entre deux contraintes : optimiser $I(a)$ (en fonction du signe de $o(a)$) et prendre garde à ce que le jugement ne prenne pas de valeurs trop négatives (sauf si un acteur malhonnête espère que l'action ne sera pas découverte, ce qui revient à ignorer la contrainte du jugement). Dans le cas de véhicules autonomes, différents propriétaires pourront ainsi accorder des valeurs différentes à la sécurité, à la consommation et au temps de parcours. En cas de dommage, la culpabilité sera celle du propriétaire qui a fourni la hiérarchie de valeurs et l'arbitrage entre les contraintes, exactement comme s'il avait agi par délégation.

Conclusion

Tout acteur responsable est censé agir en fonction de la Loi. Il doit donc intégrer le droit dans sa procédure de décision. Si l'acteur est une machine, celle-ci doit disposer d'une version automatisée du droit. La question se pose peu si la machine est cantonnée dans un registre d'actions étroit, par exemple s'il s'agit d'un véhicule autonome. Il suffit que la machine sache quoi faire dans chaque cas, si bien qu'elle n'a pas à effectuer de calcul cause-conséquences. Si l'on développe des intelligences artificielles un peu générales chargées de prendre des décisions dans un spectre assez large de compétences, notamment dans des situations inédites, la situation devient bien différente. La machine devra évaluer l'intensité des conséquences et la complexité des liens causaux. Les définitions formelles que nous avons mentionnées indiquent que l'automatisation du droit, indispensable pour de telles machines, est possible en principe. Il y a loin, toutefois, des principes au calcul effectif. Disposer de définitions calculables constitue cependant une étape incontournable. Les définitions que nous venons d'indiquer sont de celles qui permettront non seulement aux machines de calculer leurs actions, mais également de produire des justifications de ces actions, exactement comme il est attendu d'acteurs humains. Le jugement, en cas de dommages, ne pourra pas être supporté par la machine, même si celle-ci connaît sa responsabilité. Le responsable sera toujours celui qui fournit la hiérarchie de valeurs à la machine.