Transparence des algorithmes : quelles réponses juridiques et techniques ?

Daniel Le Métayer
Inria
http://www.inrialpes.fr/planete/people/lemetayer/dlemetayer-fr.htm

On s'intéresse ici à une catégorie particulière d'algorithmes, ceux qui sont utilisés pour l'aide à la décision ou dans des traitements qui ont des incidences sur les comportements individuels, qui ont donc un effet normatif. A titre d'exemples, on peut citer :

- Les algorithmes de classement, qui établissent des priorités, des recommandations : on pense évidemment aux algorithmes de présentation des résultats des moteurs de recherche, mais aussi à ceux qui sont utilisés pour classer les candidats à un poste ou les rues d'une ville que la police doit surveiller en priorité (police prédictive : PredPol), etc.
- Les algorithmes de catégorisation, de classification, de profilage comme ceux qu'on met en œuvre pour détecter des profils de potentiels terroristes, de fraudeurs, de clients, de personnes non solvables ou à cibler dans une campagne électorale.

Ces algorithmes présentent quelques caractéristiques communes :

- ils ont une incidence importante sur les vies des personnes concernées (soit leurs vie quotidienne, soit à des moments spécifiques, généralement déterminants: candidature à un poste, demande de prêt, demande de visa, etc.),
- ils ne sont ni neutres (au sens où ils mettent en œuvre des critères de priorité, de catégorisation, etc.), ni forcément corrects (faux positifs, faux négatifs), et
- leur fonctionnement est généralement opaque (certains utilisateurs ignorent même parfois leur existence).

Cette combinaison de caractéristiques plutôt inquiétante permet d'imaginer toutes sortes de dérives : traitements injustes, discriminations, manipulation, etc.

Face à cela, on peut se poser au moins deux types d'interrogations: est-ce qu'on peut distinguer des types d'utilisation, des circonstances, où il faudrait limiter l'usage de ces algorithmes : par exemple jusqu'où souhaitons-nous aller en terme d'individualisation des traitements en matière d'assurances (doit-on passer par pertes et profits le principe de mutualisation ?), de consommation (doit-on pouvoir faire payer n'importe quel bien à la tête, ou plutôt au profil, du client ?). Est-il acceptable que des décisions importantes puissent être prises sur la base d'algorithmes complètement opaques pour le décideur (et qu'en est-il des responsabilités si la décision s'avère mauvaise) ?

Puisqu'un des problèmes provient de l'opacité de ces algorithmes, une question essentielle est celle de la transparence. Tout d'abord, la transparence qui nous intéresse ici ne peut pas se réduire à la simple mise à disposition des codes source des logiciels, que le néophyte (et même parfois l'expert) a peu de chances de comprendre. Ce n'est pas non plus forcément connaître leur mode opératoire dans ses moindres détails. L'important est de pouvoir comprendre certains aspects critiques du

fonctionnement d'un algorithme, notamment les informations qui sont utilisées et leur impact sur le résultat final (favorable, défavorable, dans quelle mesure ?).

Différentes méthodes peuvent être appliquées pour améliorer la compréhension des algorithmes, notamment la rétro-ingénierie de logiciels. Cependant, celle-ci a des limites. D'une part légales, puisqu'elle est parfois interdite par les auteurs des logiciels, mais surtout techniques : en effet, le procédé demande beaucoup d'effort, la réussite n'est pas certaine et de plus certains des algorithmes s'y prêtent mal, notamment les algorithmes qui reposent sur l'apprentissage. La complexité de ces algorithmes provient généralement de la taille des données analysées plus que de celle du code lui-même ; de plus, leur fonctionnement évolue au cours du temps. Par ailleurs certains algorithmes, ou leur paramétrage, peuvent aussi être modifiés régulièrement par les acteurs qui les contrôlent (on sait que c'est le cas pour les moteurs de recherche).

L'inconvénient principal de la rétro-ingénierie est qu'il s'agit une démarche non-collaborative, a posteriori. L'idéal serait en fait de promouvoir une démarche de transparence et de responsabilité par construction – ce qu'on appelle parfois « accountability by design » de la même façon qu'on parle de « privacy by design » – incorporer ces valeurs, ces exigences dès la phase de conception d'un système.

Pour conclure, il faut aussi se poser les questions de ce qu'on peut légitimement exiger en matière de transparence et des limites de cette transparence. Ces limites peuvent être liées à des impératifs de protection de la propriété intellectuelle : quand cet argument est-il recevable ? Est-ce que c'est nécessairement un obstacle à la transparence ? Une autre limite est liée au phénomène de contournement (la connaissance du fonctionnement de l'algorithme permet de s'y adapter) avec les mêmes questions : quand cet argument est-il recevable ? Quand l'exigence de transparence doit-elle l'emporter ? Et, finalement, dans les situations où l'objectif de transparence serait tout à fait hors de portée, peut-on accepter l'utilisation d'un algorithme comme outil d'aide à la décision quand les enjeux sont importants pour les personnes concernées (emploi, prêt, visa, décision de justice, etc.) ?

Le législateur s'est déjà saisi de la question de la transparence des algorithmes, notamment à l'occasion de la loi pour une République numérique qui impose de nouvelles obligations aux administrations et aux plateformes. Le nouveau Règlement européen sur la protection des données personnelles comporte également des dispositions visant à améliorer la transparence mais celles-ci sont très limitées et d'une effectivité douteuse. Comment améliorer la transparence des algorithmes ? Comment règlementer en la matière ? Comment améliorer leur production, leur compréhensibilité ? Nous sommes en présence de défis majeurs pour les années à venir, des défis qui ne peuvent être abordés que dans une démarche interdisciplinaire tant les aspects juridiques et techniques sont entremêlés en la matière.