Algorithmes d'aide à la décision : au-delà de la transparence, la redevabilité

Daniel Le Métayer, Inria

En collaboration avec Sonia Desmoulin-Canselier, CNRS, Univ. Nantes

Plan

1. Enjeux: pourquoi est-ce important ?
2. Terminologie: de quoi parle-t-on ?
3. Ressources du droit: de quoi dispose-t-on ?
4. Ressources de la technique: de quoi dispose-t-on ?
5. Défis: que faire ?

Omniprésence des algorithmes

- Importance croissante des algorithmes d'aide à la décision dans tous les secteurs d'activité (information, publicité, recrutements, assurances, médecine, justice, enseignement, police, etc.)
- Impacts majeurs sur les citoyens, les acteurs économiques, les états (bénéfices et risques)
- Frontière poreuse entre aide à la décision et décision

Enjeux pour les personnes

Equité:

- Discrimination, traitements défavorables,
- Stigmatisation

Autonomie, développement individuel, dignité:

- Censure
- Effet « bulle filtrante »
- Manipulation, pratiques déloyales
- Atteinte à la vie privée, à la liberté de circuler librement
- Atteinte à la présomption d'innocence, au droit à un procès équitable

Plan

1. Enjeux: pourquoi est-ce important ?
2. Terminologie: de quoi parle-t-on ?
3. Ressources du droit: de quoi dispose-t-on ?
4. Ressources de la technique: de quoi dispose-t-on ?
5. Défis: que faire ?

Quels remèdes ?

- Transparence: montrer
- Explicabilité: expliquer
- Intelligibilité: comprendre
- Redevabilité: rendre compte, justifier
- Loyauté: tenir sa parole, « assurer de bonne foi le service » (Conseil d'Etat)
- Equité: absence de discrimination, de biais

Quels remèdes ?

- Transparence: montrer
- Explicabilité: expliquer
- Redevabilité: rendre compte, justifier

Relations tripartites: responsable, destinataire, contenu

Transparence

- Montrer le code seul ne suffit pas (documents de conception, données d'apprentissage, paramètres, etc.)
- Montrer ne signifie pas forcément rendre public (auditeurs, autorités de contrôle, etc.)
- Transparence n'implique ni explicabilité, ni intelligibilité, ni redevabilité, ni loyauté
- … mais peut y contribuer

Explications

- Différents destinataires: concepteurs, utilisateurs professionnels, personnes affectées, auditeurs
- Différents objectifs: vérification, amélioration, compréhension, contestation, aide à la décision
- Différentes formes: explications locales ou globales, opérationnelles ou fonctionnelles

Pourquoi expliquer ?

Différentes objectifs selon les destinataires:

- Concepteurs: améliorer, justifier, rendre compte
- Utilisateurs professionnels: comprendre, prendre des décisions, rendre compte, etc.
- Personnes affectées: comprendre, contester, prendre des décisions, etc.
- Auditeurs, organismes de certification: valider

Différents types d'explication

Différents types d'explication selon les objectifs:

- Explications globales (logique générale de l'algorithme) ou locales (explication de cas particuliers)
- Explications fonctionnelles (liens entrées-résultats, incertitude) ou opérationnelles
- Explication des traitements et/ou des collectes de données personnelles

Qu'est-ce qu'une bonne explication ?

Critères multiples et parfois en tension:

- Intelligibilité (simplicité)
- Fidélité
- Précision
- Complétude
- Cohérence
- Généralité (et caractérisation des limites)

Limites

- Des explications trompeuses peuvent contribuer à aggraver les risques
- La transparence et l'explicabilité ne contribuent pas forcément à améliorer la confiance (phénomène d'« aversion algorithmique »)
- Situations où les connaissances sur l'algorithmes peuvent permettre de les leurrer pour en tirer avantage
- La transparence et l'explicabilité ne doivent pas être utilisés comme des moyens de s'exonérer de ses responsabilités

Redevabilité

- A party A is accountable to a party B with respect to its conduct C, if A has an obligation to provide B with some justification for C, and may face some form of sanction if B finds A's justification to be inadequate.
- In the context of algorithmic decision-making, an accountable decision-maker must provide its decision-subjects with reasons and explanations for the design and operation of its automated decision-making system.

Binns R. (2017), Algorithmic accountability and public reason, Philosophy & Technology, May 2017

Redevabilité

En anglais, on utilise le terme «accountability», issu du monde anglo-saxon où il est d'usage courant et où il existe un vaste consensus sur le sens à lui donner – bien qu'il soit difficile d'en définir avec précision le sens dans la pratique. Globalement, on peut toutefois dire qu'il met l'accent sur la manière dont la responsabilité est assumée et sur la manière de le vérifier. On ne peut inspirer une confiance suffisante que s'il est démontré que la responsabilité est efficacement assumée dans la pratique.

WP29, Avis 3/2010 sur le principe de responsabilité, Commission Européenne

Plan

1. Enjeux: pourquoi est-ce important ?
2. Terminologie: de quoi parle-t-on ?
3. Ressources du droit: de quoi dispose-t-on ?
4. Ressources de la technique: de quoi dispose-t-on ?
5. Défis: que faire ?

Obligations en matière d'explications RGPD

Règlement européen sur les données personnelles
Collecte et droit d'accès (art. 13, 14, 15)

Obligation de fournir des informations concernant

« l'existence d'une prise de décision automatisée, y compris un profilage, visée à l'article 22, paragraphes 1 et 4, et, au moins en pareils cas, des informations utiles concernant la logique sous-jacente, ainsi que l'importance et les conséquences prévues de ce traitement pour la personne concernée »

Obligations en matière d'explications RGPD

Règlement européen sur les données personnelles
Décision individuelle automatisée (art. 22)

- La personne concernée a le droit de ne pas faire l'objet d'une décision fondée exclusivement sur un traitement automatisé produisant des effets juridiques la concernant ou l'affectant de manière significative de façon similaire.
- Nombreuses exceptions (contrat, consentement, etc.)

Obligations en matière d'explications RGPD

Règlement européen sur les données personnelles
Décision individuelle automatisée (art. 22)

> « Le responsable du traitement met en œuvre des mesures appropriées pour la sauvegarde des droits et libertés et des intérêts légitimes de la personne concernée, au moins du droit de la personne concernée d'obtenir une intervention humaine de la part du responsable du traitement, d'exprimer son point de vue et de contester la décision »

Le mot explication n'apparaît pas dans l'article mais dans le Considérant (71): « le droit d'obtenir une intervention humaine, d'exprimer son point de vue, d'obtenir une explication quant à la décision prise à l'issue de ce type d'évaluation et de contester la décision »

Obligations en matière d'explications RGPD

Limites du règlement européen sur les données personnelles en matière d'explications:

- Dispositions sujettes à interprétation
- Restriction majeures :
  - Décisions fondées exclusivement sur un traitement automatisé
  - Décisions produisant des effets juridiques ou affectant de manière significative de façon similaire le sujet
  - Limitation au traitement des données personnelles

Loi pour une République numérique

Nouvelles obligations pour les administrations:

- Art. 4: … une décision individuelle prise sur le fondement d'un traitement algorithmique comporte une mention explicite en informant l'intéressé. Les règles définissant ce traitement ainsi que les principales caractéristiques de sa mise en œuvre sont communiquées par l'administration à l'intéressé s'il en fait la demande
- Art. 6: …les administrations … publient en ligne les règles définissant les principaux traitements algorithmiques utilisés dans l'accomplissement de leurs missions lorsqu'ils fondent des décisions individuelles

Loi pour une République numérique

Décret du 16 mars 2017 (Art. R. 311-3-1-2):

L'administration communique à la personne faisant l'objet d'une décision individuelle prise sur le fondement d'un traitement algorithmique, à la demande de celle-ci, sous une forme intelligible et sous réserve de ne pas porter atteinte à des secrets protégés par la loi, les informations suivantes :

1° Le degré et le mode de contribution du traitement algorithmique à la prise de décision ;

2° Les données traitées et leurs sources ;

3° Les paramètres de traitement et, le cas échéant, leur pondération, appliqués à la situation de l'intéressé ;

4° Les opérations effectuées par le traitement ;

Loi pour une République numérique

Tout opérateur de plateforme en ligne est tenu de délivrer au consommateur une information loyale, claire et transparente sur :

1. Les conditions générales d'utilisation du service d'intermédiation qu'il propose et sur les modalités de référencement, de classement et de déréférencement des contenus, des biens ou des services auxquels ce service permet d'accéder ;
2. L'existence d'une relation contractuelle, d'un lien capitalistique ou d'une rémunération à son profit, dès lors qu'ils influencent le classement ou le référencement des contenus, des biens ou des services proposés ou mis en ligne ;

…

Premières conclusions sur les instruments juridiques

- Absence de réflexion générale sur l'explication des algorithmes d'aide à la décision
- Faiblesse du cadre juridique actuel, dispositions parcellaires et sujettes à interprétation
- Ressources dans les principes juridiques (obligation du médecin de prendre une décision en conscience et de pouvoir en répondre, obligation du juge de motiver sa décision, débat contradictoire, etc.)

Plan

1. Enjeux: pourquoi est-ce important ?
2. Terminologie: de quoi parle-t-on ?
3. Ressources du droit: de quoi dispose-t-on ?
4. Ressources de la technique: de quoi dispose-t-on ?
5. Défis: que faire ?

Mode opératoire

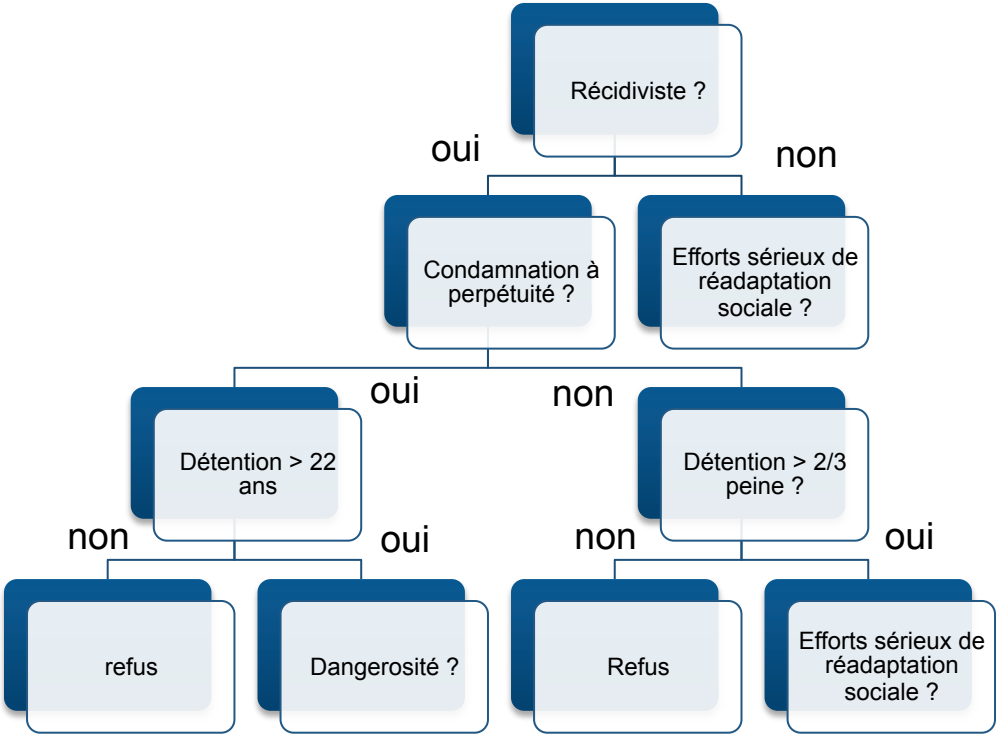
Action a posteriori (ex post) :

- Mode non-collaboratif : analyse en « boîte noire » quand le texte de l'algorithme n'est pas disponible (test, forme de rétro-ingénierie)
- Mode collaboratif : analyse en « boîte blanche » quand le texte de l'algorithme est disponible

Action a priori (ex ante) :

- Mode constructif: explications générées avec le résultat nominal

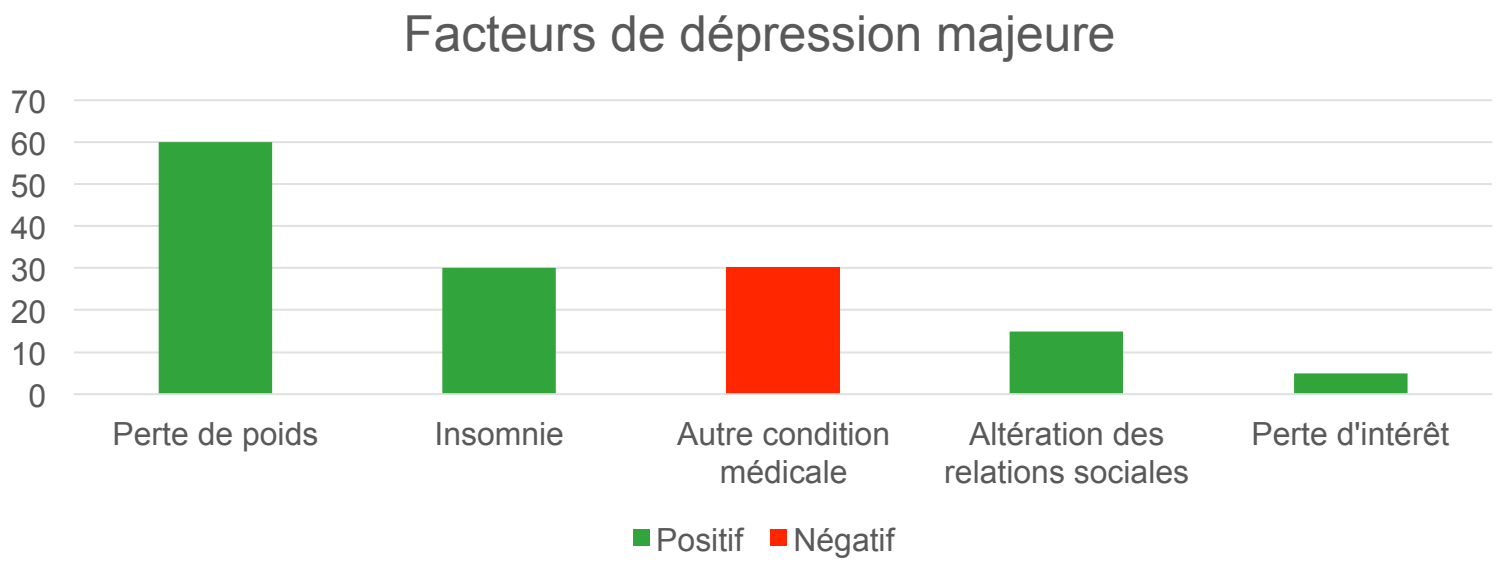
Explication globale: arbre de décision



Explication globale: table de décision

Perte d'intérêt	Perte de poids	Inso-mnie	Altération des relations sociales	Autre condition médicale	Dépression majeure
oui	non	non	oui	non	3 %
oui	non	oui	oui	non	6 %
oui	oui	non	oui	non	10 %
oui	oui	oui	oui	oui	2 %
oui	oui	oui	oui	non	20 %

Explication globale: histogramme



Explication locale: LIME

LIME: Local Interpretable Model-agnostic Explanations (université de Washington, USA)

Why should I trust you ? Explaining the predictions of any classifier, M.T. Ribeiro, S. Singh, C. Guestrin, KDD 2016.

Explication locale: contrefactuelles

Exemples:

- Quel serait le profil le plus proche du mien qui se verrait attribuer l'université de Lyon ?
- Toutes choses égales par ailleurs, quel taux d'endettement me permettrait d'obtenir ce prêt ?

Explication locale: contrefactuelles

Avantages:

- Recherches en sciences cognitives: valeur des explications par contraste
- Intérêts multiples: comprendre, contester, améliorer, etc..

Limites:

- Explications locales
- Insuffisant pour valider un algorithme

Plan

1. Enjeux: pourquoi est-ce important ?
2. Terminologie: de quoi parle-t-on ?
3. Ressources du droit: de quoi dispose-t-on ?
4. Ressources de la technique: de quoi dispose-t-on ?
5. Défis: que faire ?

Conclusion: défis techniques

Multiples défis à relever sur le plan technique :

- Explications par construction
- Concilier des objectifs en tension (intelligibilité, précision, fidélité, etc.)
- Evaluer la qualité des explications (intelligibilité, fidélité, etc.)
- Au-delà des explications: mode interactif, test d'hypothèses, etc.

Interdisciplinarité: collaborations nécessaires avec cogniticiens, psychologues, éthiciens, juristes, etc.

Conclusion : questions juridiques, éthiques, politiques

- Quelles formes de « redevabilité » exiger et vis à vis de quels acteurs?
- Quels facteurs devraient-on considérer comme inacceptables pour quelles décisions et pour quelles aides à la décision?
- Quels moyens de contrôle ? Faut-il instaurer des processus de certification spécifiques, une ou des autorité(s) de contrôle, un comité d'éthique, un corps d'experts assermentés (rapport Villani) ?
- Comment préserver la part d'humanité dans la décision ?