Synthèse et propositions sur la responsabilité sociale des algorithmes

Conseil Scientifique de l'institut CNRS INS2I

Fiche de synthèse, Septembre 2018

1 Motivations

De procédures formelles pour résoudre de problèmes complexes de "décision" existent depuis de siècles. Plusieurs de ces procédures sont utilisées par ailleurs aujourd'hui dans des "machines" qui gèrent de manière automatique ou semi-automatique de processus complexes : gestion de la production, gestion de réseaux de communication et/ou de distribution, gestion du trafic, allocation de fréquences en télécommunication mobile, conduite automatique (trains, métro etc.). L'utilisation de telles procédures (soit sous forme de logiciels ou simplement comme procédures d'allocation de ressources) n'avait soulevé que peu de débats de nature éthique ou encore sur leur impact potentiellement négatif sur les personnes et la société. [1]

Nous assistons aujourd'hui à une diffusion d'études, de discussions et d'articles de presse centrés sur l'impact social et économique de la diffusion "massive" de l'utilisation de telles procédures et des algorithmes associés. Si d'un coté nous assistons à une vision non scientifique de cette thématique nous ne pouvons pas non plus ignorer son existence. Pour cette raison le Conseil Scientifique de l'INS2I a promu l'organisation de quatre interventions le 25/09/2017 :

— Serge Abiteboul (ENS, INRIA) "Responsabilité sociale des algorithmes",
— Jérôme Lang (LAMSADE, Dauphine) "Conception de mécanismes et responsabilité sociale des algorithmes",

1. Notons cependant le débat en France dans les années 70 autour de la loi informatique et liberté de 1978 aui anticipait certaines conséquences négatives de l'utilisation généralisée des bases de données contenant des données personnelles.

— Benjamin Nguyen (LIFO, Orleans) “Responsabilité sociale des algorithmes : quelques enjeux”,
— François Pellegrini (LABRI, Bordeaux et membre de la CNIL) “Enjeux sociétaux des traitements informatiques”.

Le document que suit est essentiellement fondé sur les présentations de ces collègues et la discussion qui s’ensuivait.

2 Historique et positionnement

Le concept d’algorithme est ancien. Toute procédure qui peut être décrite dans un langage formel et qui permet de résoudre un “problème” à travers l’allocation de ressources à des taches peut être définie comme un algorithme et l’histoire nous présente des exemples d’algorithmes depuis de siècles. Le caractère formel de la description de ce qu’un algorithme fait a permis d’imaginer (et de réaliser) de machines qui exécutent les algorithmes : de manière simplifiée nous sommes en train de décrire la naissance de la discipline Informatique comme le champ scientifique qui s’occupe de cette typologie de procédures. Ce qu’il faut retenir est :

1. l’existence des algorithmes avant la naissance de l’Informatique et
2. l’existence des logiciels que permettent d’utiliser les algorithmes de manière massive et efficace.

Le concept de responsabilité a également été étudié depuis bien avant la naissance de l’Informatique. Toute action rend l’agent qui l’entreprend responsable des conséquence qu’elle entraîne. Il doit donc analyser les conséquences potentielles d’une action avant son exécution, de manière à prévenir, maîtriser, réglementer ou interdire. Toute action intentionnelle nécessite une décision : nous trouvons ainsi le lien entre la responsabilité d’une action et le processus de décision qui l’a causé. Aujourd’hui un tel processus peut être géré (ou est géré) par un algorithme.

Cependant le concept de responsabilité reste principalement associé aux personnes et groupes de personnes ainsi qu’à des organisations. Du point de vue juridique (et de manière intuitive) une responsabilité ne peut être associée qu’à une personne physique ou morale et non pas à une machine, ou encore à une procédure ou à des règles. Prenons l’exemple des système de conduite automatique comme le métro VAL : la responsabilité en cas d’accident revient à la société de construction et/ou la société d’exploitation et non pas au VAL lui même. De ce point de vue le terme “Responsabilité sociale des algorithmes” est un paradoxe :

les algorithmes ne peuvent pas être responsables des conséquences de leur utilisation.

Le fait est que malgré ce principe élémentaire nous assistons à une diffusion de préoccupations autour de l'utilisation des algorithmes dans la prise de décisions impactant notre vie quotidienne et dans des systèmes qui gèrent l'organisation de nos sociétés et de l'économie. Pourquoi ? Dans la suite nous énumérons quelques causes possibles :

1. L'ampleur de la diffusion de procédures automatiques (ou semi-automatiques) de décision.
2. Le recours à des procédures qui apprennent de manière autonome et sans qu'on puisse contrôler la manière par laquelle l'apprentissage intervient, entraînant un comportement non prévisible de l'algorithme.
3. La concentration de la production de logiciels entre les très grandes entreprises d'édition de logiciels avec une industrialisation de l'implémentation informatique d'algorithmes (en dehors des *open source*) qui les rend encore plus opaques.
4. L'utilisation de données "privées", "sensibles" ou assimilables à des communs (*commons*) sans un retour bien identifié aux fournisseurs de données et sans toujours une autorisation claire, ainsi que l'absence de connaissances sur l'étendue des utilisations.
5. Le constat que dans plusieurs cas les procédures automatiques de décision ont produit de conséquences biaisées, inattendues, parfois aberrantes avec des impacts à long terme potentiellement non discutés dans la société.
6. L'absence d'une limite claire sur le transfert d'autonomie décisionnelle jugé acceptable.
7. Le monopole *de facto* des grands collectionneurs de données (les GAFA), engendrant un sentiment d'absence d'alternatives et d'impuissance.

3 Thématiques

Essayons donc de structurer le problème et de comprendre les points qui peuvent poser des questions à la fois dans la communauté scientifique et dans la société en général. Nous le ferons à travers des interrogations selon cinq dimensions : l'autonomie décisionnelle, la confiance aux décisions automatiques, la vérification des systèmes de décision automatique, la régulation, l'asymétrie de la connaissance.

1. Le premier problème (probablement le plus fondamental) est l'autonomie décisionnelle grandissante qu'on transfère à des "machines", à des artéfacts autonomes (AA) et la manière dont ce transfert est opéré. Rappelons cependant que ce transfert existe depuis long temps. Il faut comprendre ce qui le rend aujourd'hui un problème ou une menace potentielle.

1. Qui décide si un certain type de processus de décision peut être transféré à un Artéfact Autonome et rendu automatique ? À qui nous pouvons confier ce pouvoir de décision (dans le cas de décision publiques comme sur les modalités de l'Admission Post-Bac) ?
2. Supposons que il y a un accord sur l'automatisation d'un processus de décision (ou qu'une entreprise décide d'automatiser un processus de décision faisant partie de ses activités) ; qui décide la manière par laquelle cette automatisation sera rendue effective ?
3. L'interrogation précédente nous amène donc à nous interroger sur les propriétés que la solution et la procédure automatique doivent garantir. Qui décide les propriétés à respecter et comment concevoir des Atréfacts Autonomes qui les respectent ? Est-ce qu'on peut utiliser des outils de vérification de programmes pour démontrer les propriétés de la procédure ?
4. D'où viennent les données qui sont nécessaires pour faire marcher correctement la procédure de décision automatique ? Sont-ce des données privées ? Publiques ? Comment établir la propriété des données et les droits d'exploitation ? Comment garantir que la procédure n'introduit pas de biais ou discriminations ?
5. Qui est responsable pour une décision automatique prise par l'Artéfact Autonome ? Qui peut être tenu responsable en cas de litige devant la justice ?

La dernière demande soulève un problème d'ordre général : jusqu'à quel point sommes nous prêts à transférer de l'autonomie décisionnelle à des Artéfacts Autonomes ? Y a-t-il des processus de décision que nous ne souhaitons pas automatiser (même si c'est possible) ?

2. À quelles conditions sommes nous prêts à accepter une décision prise de manière automatique par un Artéfact Autonome ? Ceci soulève le problème de la confiance dans les algorithmes qui sont utilisés pour permettre à l'Artéfact Autonome de décider de manière autonome.

1. Étant donné un ensemble d'algorithmes utilisés par un Artéfact Autonome, sommes nous capable de construire une trace complète de l'activité des algorithmes ?

2. Si nous sommes capables de construire une telle trace, sommes nous capables de construire des explications (interprétables, simples et intuitives, utilisables) pour justifier les décisions prises à n'importe quelle partie prenante ?
3. Sommes nous capables de produire les "raisons ultimes" d'une décision qui a été prise de manière automatique ? Si c'est le cas pouvons-nous reproduire cette décision à partir des même données en entrée ?
4. Si nous utilisons des algorithmes qui apprennent à chaque exécution (et se modifient en conséquence) alors nous ne pouvons pas garantir la reproduction du résultat : nous avons un comportement non déterministe de l'algorithme. Quelles explications/justifications/raisons nous accepterions dans ce cas ?
5. L'utilisation d'un Artéfact Autonome qui automatise certaines décisions peut avoir de conséquences à long terme que nous ne pouvons pas prévoir aujourd'hui (des 'inconnus aujourd'hui inconnus'). À quelle horizon temporel faut-il vérifier les impacts à longue terme induites par l'utilisation de tels Artéfact Autonome ?

De manière générale les interrogations soulevées tournent toutes autour du problème de l'opposition possible à une décision prise de manière automatique, ou de sa défense dans le cas où une contrepartie considère que une telle décision soit opposable. Comment pouvons nous structurer une discussion argumentée autour de décisions prise de cette manière ? Au bout du compte une décision est "acceptable" si nous la retenons non opposable.

3. Clairement, une dimension de préoccupation dans la construction des Artéfacts Autonomes est le fait qu'il s'agit de logiciels complexes pour lesquels se pose le problème de la vérification formelle de leur comportement.

1. Supposons qu'on ait conçu un algorithme garantissant une certaine propriété (spécification) du résultat (voir le point 1 de la discussion ; par exemple un algorithme 'équitable'). Comment pouvons-nous garantir de manière formelle que le logiciel qui implémente un tel algorithme continue à respecter les bonnes propriétés de l'algorithme ? En d'autres termes : l'ingénierie des spécifications peut-elle garantir aujourd'hui le respect de cette typologie de demande ? Est-ce qu'on peut vérifier des Artéfacts Autonomes de la même manière qu'on vérifie aujourd'hui la spécification d'un programme contre son implémentation ?

2. Est ce que des logiciels qui garantissent de tels spécifications peuvent garantir également l'efficacité de l'algorithme ?
3. Est ce que nous pouvons garantir au même temps le respect des données privées et leur protection d'une part et de l'autre part la possibilité d'expliquer, tracer les décisions d'un Artéfact Autonome ? Où devons-nous placer le compromis nécessaire entre ces deux exigences ?
4. Si le recours à des logiciels libres peut garantir un contrôle plus efficace sur les caractéristiques et le comportement d'un Artéfact Autonome, comment intégrer l'utilisation des composantes cryptographiques et de sécurité en général qui nécessitent une certaine confidentialité du code ? Encore une fois comment trouver un compromis acceptable et quelles seront les conséquences sur la limite de conception d'un Artéfact Autonome ?

Plus généralement, le problème de la vulnérabilité et donc de la sécurité des Artéfact Autonome avec autonomie décisionnelle augmentée sera un sujet d'importance majeure dans l'immédiat (voir la discussion sur la sécurité du vote électronique).

4. Jusqu'à maintenant nous avons soulevé des questions qui concernent la conception d'algorithmes et de systèmes de décision automatique que nous avons appelé Artéfact Autonome. Vue l'impact potentiel de tels systèmes, une question qui se pose de manière naturelle est celle de la régulation de la conception, de la mise en œuvre et de l'utilisation de tels Artéfacts Autonomes.

1. Si nous ne pouvons pas prévoir l'impact d'un tel système (par la vérification de leur spécification), faut-il prévoir une phase de test (plus ou moins comme dans le cas de nouvelles molécules chimiques ou de nouveau aliments) avant d'autoriser l'utilisation d'un Artéfact Autonome avec autonomie décisionnelle augmentée ?
2. Faut-il donc envisager la création d'une autorité indépendante de certification et d'autorisation ?
3. Étant donné l'impact potentiel sur le société dans certaines contextes de l'utilisation de tels Artéfacts Autonomes, faut-il prévoir une déclaration explicite de la politique et de la raison pour laquelle l'utilisation d'un Artéfact Autonome est prévu ?
4. Étant donné l'importance économique de l'industrie informatique et de l'édition de logiciels, une telle régulation est-elle réaliste ? À quelles conditions ?

5. Étant donné le cycle de vie (souvent assez court) de tels Artéfacts Autonomes, est-il réaliste d'imaginer une régulation (qui a typiquement un cycle de vie bien plus long) ?

5. D'une manière générale, on peut constater une asymétrie massive de l'information : les GAFA savent tout sur nous tandis que nous ne savons que peu sur eux, à commencer par les algorithmes qu'ils utilisent. L'exemple le plus frappant en est le *PageRank* de Google dont le fonctionnement est tenu secret. En économie il est bien connu que l'asymétrie informationnelle peut entraîner des défaillances du marché. L'importance de cette asymétrie et le contexte technologique posent ici des questions nouvelles.

1. Comment pouvons-nous juger l'explication d'une décision d'un algorithme qui nous est fournie si le fonctionnement de l'algorithme n'est pas connu ? Pouvons nous faire confiance à un tel service ?
2. Les conditions d'utilisation de multiples services requièrent actuellement l'accord des utilisateurs dans le cadre du nouveau Règlement général sur la protection des données (RGPD), et presque tous les utilisateurs acceptent en cliquant sur "J'ai lu" sans en avoir pris connaissance : comment éviter une telle situation ? Comment s'assurer d'une utilisation responsable de services de ce type ?
3. Si l'utilisation d'un Atéfact Autonome engendre de-facto la stipulation d'un contrat, comment s'assurer que les parties prenants : "concepteur d'un Artéfact Autonome", "éditeur du logiciel", "utilisateur de l'Artéfact Autonome", "la société en général", soient conscients du contrat signé ?
4. Qui est l'autorité compétente en cas de litige dans le contexte de contrats de ce type ?

Nous laissons par ailleurs ouverts les dimensions éthiques de la manipulation de données et de la "morale" associable à de procédures de décision automatique : ces sujets demandent également une discussion et peut-être une régulation, cf. la discussion actuelle sur les 'machines morales' (*moral machines*). Nous partons effectivement de l'hypothèse que les raisons économiques qui poussent l'évolution et la création d'Artéfact Autonome dotés d'autonomie décisionnelle resteront impératives.

4 Ressources

La communauté Française a déjà plusieurs activités dans ce domaine. Dans la suite nous listons les initiatives dont nous avons connaissance.

— La CNIL a publié un rapport en novembre 2017 “Comment permettre à l’homme de garder la main ?” [2] suite à une série de débats publics dans le cadre de la mission éthique qui lui a été confiée par la loi pour une république numérique de 2016.

— Le GDR “Sécurité Informatique” [3] a un groupe de travail sur la protection de la vie privée dirigé par Benjamin Nguyen (LIFO, Université d’Orleans). Notons que ce sujet dépasse le cadre de ce GDR avec une dimension clairement interdisciplinaire.

— Le GDR “Intelligence Artificielle” a organisé en 2017 une Journée Explicabilité des systèmes d’IA [4] animée par Nicolas Maudet (LIP6, Université Pierre et Marie Curie) ; il est prévu de lancer un groupe de travail sur le sujet.

— Le GDR “Policy Analytics” [5] a organisé en 2017 un colloque sur la “Responsabilité Sociale des Algorithmes” [6].

— À l’Université Paris Est Créteil existe un groupe de travail interdisciplinaire “ Algorithme et citoyenneté” (ALGOCIT) [7] animé par Pierre Valarcher.

— À l’ISCC (CNRS et Université Pierre et Marie Curie) Melanie Dulong de Rosnay et Francesca Musiani travaillent sur ces thématiques. [8]

— Deux ouvrages de Serge Abiteboul (“Le temps des Algorithmes”, avec Gilles Dowek et “Terra Data”, avec Valérie Peugeot) et l’ouvrage de Francesca Musiani (“Nains sans géants”).

Au niveau international signalons :

— Un projet Européen dirigé par Dino Pedreschi (Université de Pise) sur la collecte responsable de données. [9]

2. https://www.cnil.fr/sites/default/files/atoms/files/cnil_rapport_garder_la_main_web.pdf
3. https://gdr-securite.irisa.fr
4. https://www.gdria.fr/journee_explicabilite/
5. http://www.gdr3720.fr
6. http://www.lamsade.dauphine.fr/sra2017
7. http://www.algo-cit.fr/
8. http://www.iscc.cnrs.fr
9. https://sobigdata.eu

— Un séminaire Dagstuhl sur la conception d'agents moraux. [10]
— La création de la communauté FAT (Fair, Accountable, Transparent Algorithms) essentiellement aux Ètats-Unis. [11]
— La création de l'Institut 3A (Autonomy, Agency, Assurance) à l'Australian National University, dirigé par Geneviève Bell (ancienne vice-président de INTEL). [12]

5 Recommandations

Nous ne pouvons pas ignorer que les "algorithmes", leur conception, leur analyse et leur utilisation représente un élément essentiel de la discipline Informatique et que leur étude du point de vue de leur impact dans la société et dans notre vie quotidienne doit être une préoccupation majeure pour tous les membres de notre communauté scientifique. La discussion de la section précédente montre que, bien que l'existence d'Artéfact Autonome dotés d'autonomie décisionnelle soit une réalité depuis très longtemps il est temps de passer à une phase de réflexion plus approfondie. De ce point de vue nous voyons deux types de recommandations : l'une en ce qui concerne les aspects scientifiques et l'autre les aspects sociétaux.

Nos recommandations s'adressent essentiellement au CNRS : l'INS2I de manière spécifique (vue que nous adressons avant tout la communauté Informatique), la Mission pour l'interdisciplinarité (vue la dimension fortement interdisciplinaire du sujet) et la direction du CNRS pour l'impact du sujet sur la société.

En ce qui concerne les *aspects scientifiques* il est probablement temps de susciter des actions de recherche orientées à construire des éléments de réponse aux demandes posées. Actions qui peuvent être prises :

— lancer un appel à PEPS (via le INS2I ou en coopération avec la MI) qui peuvent adresser une ou plusieurs de demandes posées dans ce document ;
— fair de ce sujet une année thématique du CNRS (INS2I et/ou MI) avec l'implication des GDR "Sécurité Informatique" , "IA", "RO", "MADICS", "Policy Analytics" dans l'organisation de journées de discussion ; dans ce cadre (ou de manière indépendante) organiser un colloque national des chercheurs intéressés par la thématique ;

10. http://www.dagstuhl.de/en/program/calendar/semhp/?semnr=16222
11. http://www.fatconference.org
12. https://3ainstitute.cecs.anu.edu.au/

— proposer (direction du CNRS) le sujet à l'ANR et ou le Ministère (PIA AI) comme programme de recherche interdisciplinaire (et motiver la communauté scientifique à participer) ;
— proposer (direction du CNRS) également le sujet à l'attention de l'UE en ce qui concerne les priorités du 9ème PCRD ;
— construire (direction du CNRS/DRI) des liens avec des actions similaires au niveau international, comme l'Institut australien 3A et la communauté FAT déjà mentionnés plus haut, le DIMACS [13] et les financements NSF aux Ètats-Unis, ou le projet Algorithm Watch [14] en Allemagne ; dans ce cadre proposer la création d'un réseau Européen (Action COST ?) sur la thématique.
— discuter (dans le cadre de l'alliance ALLISTENE) avec l'INRIA autour de la plateforme de test et d'audit des algorithmes pour créer une initiative nationale.

En ce qui concerne les *aspects sociétaux* nous proposons des actions dans deux sens : un vers le CNRS et sa communauté scientifique et l'autre vers la société en général. L'objectif de ces actions est d'attirer l'attention de la société en général ainsi que de composants spécifiques à l'importance du problème et de la nécessite d'investir de ressources et de l'intelligence pour s'occuper. Des actions que le INS2I, la MI et la direction du CNRS peuvent entreprendre incluent :

— un positionnement du CNRS sur l'ensemble des problèmes cités à travers un document succinct qui précise l'investissement du CNRS ;
— la publication d'un insert spécial dans le Journal du CNRS ;
— prendre l'initiative pour solliciter une discussion à l'Assemblée nationale sur la base d'un document CNRS/INRIA/Ministère ;
— se concerter avec d'autres organismes de recherche en Europe pour solliciter le Parlement Européen (créer un document commun ?) ;
— proposer (via la Direction de la Valorisation) la création d'une table ronde des entreprises qui produisent ou ont un recours massive à des Artéfacts Autonomes avec autonomie décisionnelle augmentée, en vue d'une réflexion sur l'auto-régulation ;
— proposer (via la Direction de la Valorisation) un brain-trust avec des organismes professionnels sur l'avenir de certaines professions et de certaines activités économiques qui sont les premières à être touchées par cette évolution.

13. http://dimacs.rutgers.edu
14. https://algorithmwatch.org/en