From InternetActu-De l’explicabilité des systèmes : les enjeux de l’explication des décisions automatisées

Par Hubert Guillaud

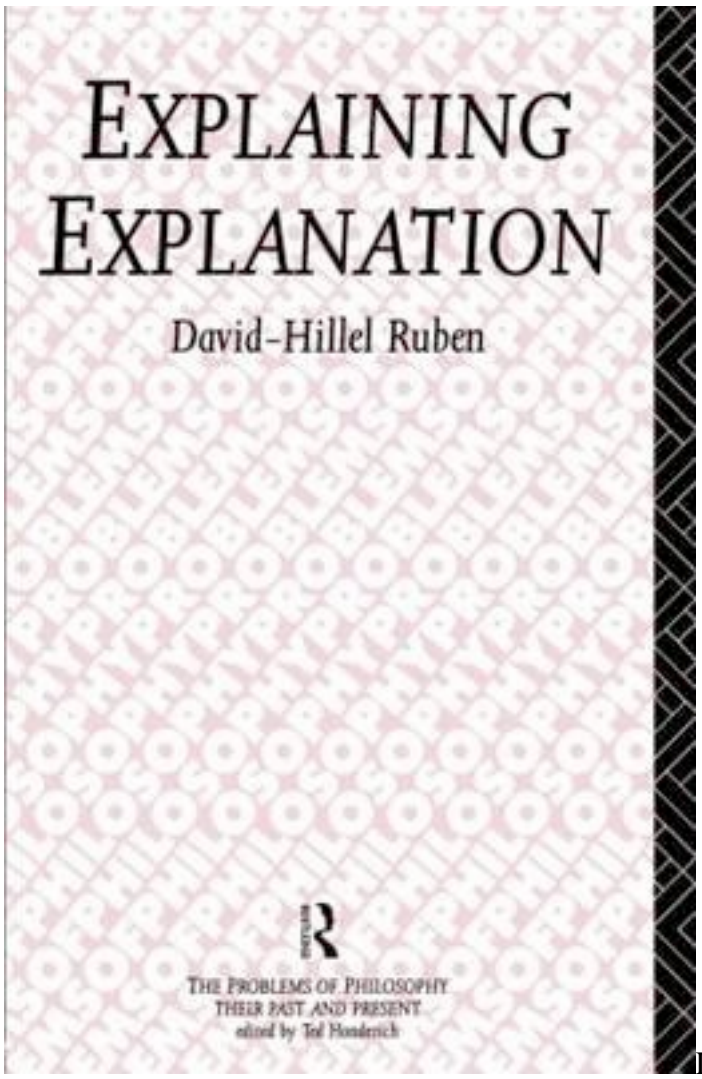
- 14/11/2019
- Hubert Guillaud
- Articles
- algorithme nossystemes

- 8 commentaires
- 47792 vues
- ~ 21 minutes

Billet précédent Billet suivant

Etymologiquement, expliquer, c’est déployer, déplier, c’est-à-dire soulever les ambiguïtés qui se cachent dans l’ombre des plis. C’est donc également « ouvrir » (avec tout le sens que « l’ouverture » – *open* – a acquis dans le monde numérique), défaire, dépaqueter, c’est-à-dire non seulement enlever les paquets, dérouler les difficultés, mais aussi montrer les noeuds qu’ils forment et comprendre pourquoi ils se forment. Expliquer, c’est clarifier. Mais clarifier comment ? Pour qui ? Que faut-il expliquer (et donc que faut-il ou que laisse-t-on dans l’ombre des plis ?) ? A qui ? Dans quel but ? Quels types d’explications doivent être disponibles ? Quand et où doivent-elles être disponibles ?… Voici quelques-unes des questions qui se posent quand on s’attaque à la question de l’explicabilité des systèmes techniques, notamment des processus algorithmiques et d’intelligence artificielle, mais également de nombres de technologies complexes, de « boîtes noires » auxquelles nous sommes de plus en plus confrontés. Qu’est-ce qu’expliquer ? Comment doit-on expliquer ce qui se dérobe à l’explication ? Explications sur l’explication.

En philosophie on distingue souvent expliquer de comprendre. Dans le *Phédon* de Platon, Socrate qui dialogue avec ses amis alors qu’il va mourir explique pourquoi il reste là à dialoguer. Il distingue alors l’explication (qui repose sur des causes physiques, simples, causales de son comportement) de la compréhension (qui nécessite d’entendre les raisons et les valeurs de ses actes). Mais l’une et l’autre ont parties liées. Car expliquer, c’est également amener à comprendre. En ce sens, l’un et l’autre sont également un moyen d’action qui peut amener à générer de l’adhésion, ou au contraire, du rejet ou de la contestation. Expliquer, tout comme comprendre, c’est donc également actionner, c’est-à-dire rendre l’action et la transformation possible.



Le philosophe David-Hillel Ruben qui a consacré dans les années 90 un livre (.pdf) à l'explication de l'explication conclut son ouvrage en soulignant que l'explication est *« ce qui relie les relations déterminantes »*, c'est-à-dire qu'elle consiste à montrer ce qui est responsable de ces relations et ce qui les fait, les fabrique, telles qu'elles sont.

Expliquer pour faire société

La question de l'explicabilité des systèmes techniques et notamment des décisions automatisées est un enjeu éthique fort de l'automatisation de nos sociétés, soulignent le philosophe et éthicien italien Luciano Floridi (@floridi) et le chercheur Josh Cowls (@joshcowls) dans un passionnant article de recherche sur les principes éthiques de l'IA. Pour les deux chercheurs, la question de l'explicabilité, comprise à la fois comme l'intelligibilité (c'est-à-dire la réponse à la question « comment ça marche ? ») et l'éthique de responsabilité (c'est-à-dire la réponse à la question « qui est responsable de la façon dont ça marche ? ») est la modalité qui permet l'application d'une IA éthique et juste. Nombre d'articles de recherche insistent sur la nécessité de rendre les systèmes explicables, c'est-à-dire de comprendre les processus décisionnels de l'IA pour que les utilisateurs comme la société puissent lui demander des comptes. Ce principe d'explicabilité s'exprime souvent sous différents termes allant de la « transparence » à la « responsabilité », en passant par « l'ouverture », « l'intelligibilité » ou « l'interprétabilité ».

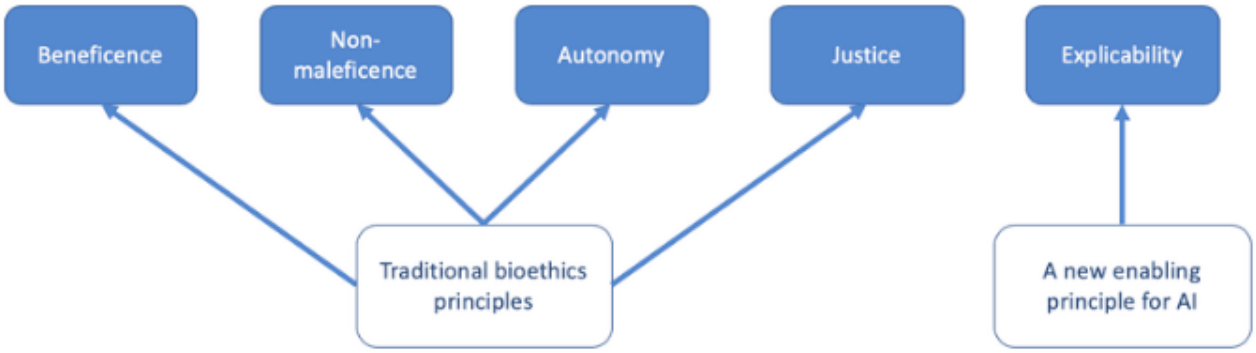


Image : Contrairement à la plupart des principes éthiques nés avec la bioéthique (comme l'autonomie ou la justice, le fait d'apporter un bénéfice à la société ou de ne pas faire le mal…) l'explicabilité est un principe d'éthique qui s'est développé avec le développement de l'IA.

L'explication : un enjeu technique ?

La question de l'explicabilité est donc d'abord un enjeu technique, un champ de recherche en soi (on parle d'ailleurs de *XAI*, pour *Explanable artificial Intelligence*), d'autant plus important que la performance des techniques d'IA est inversement proportionnelle à leur explicabilité (voir « L'intelligence artificielle va-t-elle rester impénétrable ? » et « Intelligence artificielle : comment sortir de l'opacité ? »). L'apprentissage automatique profond par exemple est l'une des techniques dont les modalités de calculs et les résultats restent les plus obscurs, même à leurs programmeurs. Ce n'est pas pour rien que l'on parle de « boîtes noires » quand on évoque les systèmes de traitement automatisés (*cf.* « Il est plus que temps que le Big data évalue ses impacts »). L'explicabilité est donc souvent un compromis entre performance et explication. Ce n'est pas le plus simple de ses paradoxes. L'amélioration technique de l'explicabilité des systèmes, notamment en IA, est donc un champ de recherche en soi, qui vise à ce que les systèmes expliquent ce qu'ils font comme leurs résultats, les critères de leurs succès comme de leurs échecs. C'est un champ de recherche qui pour l'instant n'a pas proposé de fortes percées techniques, mais qui renouvelle et approfondit l'épistémologie de l'explication.

D'ailleurs, il se pourrait que ce champ de recherche ne représente qu'un vain espoir, avancent les spécialistes Erwan Le Merrer et Gilles Trédan dans un récent article (.pdf). Pour eux, la quête d'une intelligence artificielle explicable est inatteignable. En fait, expliquent-ils sur AlgorithmWatch, un système peut très bien donner une fausse explication, comme le videur d'une boîte de nuit peut affirmer qu'il ne vous laisse pas entrer parce que vous n'avez pas la bonne cravate ! Un peu comme dans la vie réelle, la véritable raison peut toujours être couverte d'une autre explication. Pour AlgorithmWatch, ce constat plaide pour le développement d'autorités de contrôle et d'enquêtes indépendantes. Comme aujourd'hui les associations peuvent procéder à du « testing » pour évaluer les cas de discrimination à l'entrée des boîtes de nuit, nous avons besoin de pouvoir continuer à réaliser ces tests par delà les explications qui sont fournies par ceux qui déploient les systèmes. Reste que cela nécessite bien sûr des accès les plus ouverts possible au fonctionnement même de ces systèmes, ce qui est bien loin d'être le cas. Bien loin d'être seulement un enjeu technique, l'explicabilité se révèle d'abord et avant tout un enjeu de société. Kate Crawford ne disait pas autre chose quand elle soulignait l'importance de la recherche, de la presse et des associations citoyennes pour débusquer les biais des systèmes.

L'explication est interdépendante et à multiple niveau

Comme nous le notons depuis le début de cet article, l'explicabilité est foncièrement polysémique. Dans le cadre des réflexions réalisées par le groupe de travail Nos Systèmes, nous nous sommes rendu compte que l'explicabilité recouvre une large gamme d'explications, allant d'explications « procédurales » (comment faire, concrètement ? Quelles formalités accomplir ? Quelle succession d'opération exécuter ?…) à la compréhension des critères (quels sont les éléments pris en compte et

quels sont leurs « poids » relatifs les uns par rapport aux autres ?) en passant par la clarification des enjeux et motivations (pourquoi ce système ? Que fait-il… ? Et donc que ne fait-il pas ?…).

Outre la gamme d'explications à délivrer, un autre problème concerne les explications elles-mêmes : comment les délivrer ? Une question qui pose à la fois celle de savoir jusqu'où les explications doivent aller (et donc, lesquelles sont passées sous silence) et celle de leur interfaçage avec les systèmes techniques (où et quand doivent-elles être disponibles ?).

En fait, bon nombre de ces notions sont interreliées et interdépendantes. Le niveau d'automatisation du calcul (c'est-à-dire, le fait qu'il prenne des décisions sans interférence humaine) est lié à son explicabilité et son auditabilité (le fait qu'il soit vérifiable, examinable par des tiers). De même le niveau de médiation ou de dialogue influe sur l'explicabilité d'un système : le fait que le traitement propose des modalités de contact ou pas pour demander des explications, le fait qu'il propose (ou pas) une variété de supports d'explications (des FAQ aux vidéos, des schémas aux visualisations, des infographies aux tutoriels en passant par les simulateurs ou par des explications textuelles permettant plusieurs niveaux de compréhension…) ou des réponses personnalisées (avec des garanties sur les délais de réponses voire des modalités offrant des contreparties et des possibilités de recours à l'encontre d'une décision)… sont autant d'éléments qui influent sur l'explicabilité.

On constate au final que plusieurs niveaux d'explications sont possibles… Pour faire simple, les explications peuvent être :

- nulles voire faibles : les explications délivrées sont génériques, procédurales, peu informatives, incomplètes, voire déloyales.
- moyennes : les explications peuvent utiliser des formes multiples (textes, bases de connaissance, vidéos, infographies…) mais demeurent générales avec un degré de profondeur, de complétude, de fiabilité ou de précision faible.
- fortes voire exemplaires : les explications sont personnalisées, voire interactives, précisent les critères, les intentions, les modalités, documentent les résultats avec un degré de précision, de loyauté et de clarté élevé et permettent aux utilisateurs de contester voire leurs offrent des garanties et des modalités de recours en cas d'explication fausse ou insuffisante.

L'explication, une obligation légale et morale

Expliquer – c'est-à-dire motiver les décisions – est essentiel. Comme le rappelait très justement Simon Chignard (@schignard), conseiller stratégie à Etalab : l'explicabilité est aussi et avant tout une obligation légale. Apporter une explication raisonnable, même imparfaite, est nécessaire pour établir une relation de confiance entre le calcul et les calculés, entre un système et ses administrés. Le RGPD a d'ailleurs introduit un droit d'explication des décisions automatisées. Reste que ces explications doivent être également adaptées au public et justes. Or, le risque bien souvent est de fournir des explications inadaptées aux publics auxquels elles s'adressent, partielles voire partiales, incomplètes ou trop simplistes, ou pire, fausses et mensongères, comme lors d'un refus de crédit qui peut vous donner une raison de refus, alors que la raison principale est tout autre (le fait que vous viviez seul par exemple, que vous soyez une femme ou une personne de couleur… à l'image des très récents soupçons à l'égard de la carte de crédit d'Apple). Un enjeu qui pose une question de fond sur la loyauté des explications elles-mêmes et donc leur contrôle. En ce sens, l'explicabilité est donc à la fois une obligation légale – car le résultat et la modalité d'un calcul doivent pouvoir être rendus au calculé – ce qui suppose des modalités de contrôle. Et une obligation morale : car un système qui délivre un résultat à quelqu'un doit pouvoir le justifier, précisément, fidèlement et factuellement.

L'explicabilité n'est donc pas négociable. Ce qui n'est pas explicable ne peut entrer en discussion avec la société. *« L'absence d'interprétabilité n'est pas une option »*, disions-nous. La philosophe Antoinette Rouvroy nous rappelait qu'il n'y a pas de société sans motivation des décisions, que les

finalités d'un traitement doivent nous être transparentes. Que les motivations d'une décision doivent être communiquées et communicables sous une forme intelligible. Pourtant, trop souvent encore, l'explication résiste, notamment parce que les ingénieurs soulignent que fournir des explications fait peser un risque sur la qualité ou la robustesse d'un système. Pour beaucoup de responsables des traitements, les expliquer ou les ouvrir fait peser un risque de subversion, de détournement, comme si la transparence de l'explication faisait peser un risque sur la qualité du traitement lui-même. Chez les *data scientists* qui conçoivent les algorithmes et les systèmes automatisés, il y a partout une réticence à l'explication par peur d'un « effet rebond » à l'image du Google Bombing qui a longtemps permis aux internautes de se jouer des résultats de Google en y introduisant des formes de détournement. Or, l'opacité des calculs permet surtout de masquer leurs lacunes, leurs faiblesses, leurs failles, c'est-à-dire leurs défauts, les injustices, les biais et les inégalités que les chiffres recouvrent, expliquait brillamment la mathématicienne Cathy O'Neil. Idéalement, un calcul qui mesure parfaitement ce qu'il prétend ne devrait pouvoir être subverti par sa propre transparence. Comme le pointait très récemment la commission éthique des données du gouvernement allemand, plus un système a un impact social, c'est-à-dire un impact sur la vie et l'existence d'individus, plus ceux-ci doivent être contraints à la transparence et contrôlés afin qu'ils mesurent bel et bien ce qu'ils annoncent mesurer.

Où sont les explications ? Vers une échelle des niveaux d'explications ?

L'explication est le parent pauvre des systèmes techniques. Bien souvent, les explications sont tout simplement absentes, comme si elles n'étaient pas nécessaires au fonctionnement des systèmes. Dans le monde de l'innovation, on nous propose souvent une solution avant même de nous expliquer comment elle fonctionne. L'explication vient parfois plus tard. Et nombre de systèmes fonctionnent sans fournir beaucoup d'explications sur ce qu'ils font, comment ils le font, dans quel but et sans proposer de garanties ne serait-ce que sur la fiabilité de leurs explications donc…

Le refus d'explication
Outre leur absence, bien souvent, les explications nous sont refusées. C'est le cas par exemple des « recettes » utilisées par les commissions d'examen des voeux de chaque filière pour trier les candidatures sur Parcoursup. Sous prétexte de protéger les délibérations pédagogiques des jurys d'admission, les algorithmes locaux permettant aux filières de trier les milliers de candidatures qu'elles reçoivent ont été exclus de la règle de communication au public (au prétexte que la loi n'oblige les universités à n'informer que les candidats qui en font la demande). Or, nombre de filières (on estime que c'était le cas de 25 % d'entre elles en 2019) ont mis en place des outils de classement des candidats à l'image des feuilles Excel que montrait Julien Gossa en 2018, permettant par exemple de trier les élèves selon leur lycée de provenance (une pratique discrète qui entérine la sélection sociale). Ce dernier pointait d'ailleurs très bien combien l'évaluation des candidats était compliquée, pour ne pas dire impossible. Il n'empêche que mieux saisir les attendus des filières, leurs critères de classements et de tris, demeure un moyen de comprendre les motivations des sélections… mais dans parcoursup, cette compréhension qui est refusée au candidat est bien souvent démultipliée par le fait que chaque filière va utiliser une méthode qui lui est propre, comme autant de recettes auxquelles il est difficile de se conformer.

Ce n'est bien sûr pas le seul endroit où les explications nous sont refusées. Nombres de demandes que nous pouvons accomplir auprès de systèmes qui s'appuient sur des traitements et calculs prennent en compte des critères qui ne sont pas tous explicites. Les demandes d'affectation pour un poste, les demandes de crédits ou les systèmes de tris automatisés de candidatures à des emplois font parti des systèmes automatisés qui délivrent peu ou pas d'information sur leurs modalités, sur les critères utilisés ou d'explications sur leurs décisions. Nous sommes là confronté à un déni

d'information et un déni démocratique qu'il est nécessaire d'armer, en obligeant à la démultiplication des explications des procédures de calcul.

Le code et la transparence : l'explication technique

Au mieux, pour les plus ouverts des systèmes de traitement, on trouve une explication technique : le code, voire la formule du calcul est directement accessible. Parfois elles sont accompagnées d'indications sur les données utilisées et leur pondération dans le calcul… C'est là une explication technique pour technicien. Elle est souvent précieuse, primordiale, mais ne parle qu'aux techniciens, pas à la société à laquelle des explications sont dues. Elle nécessite d'être décodée, dépliée, interprétée, argumentée… Mais, depuis le code, des tiers (associations, citoyens, médias…) peuvent s'en emparer pour interroger la construction qui a eu lieu. C'est le cas notamment de nombre de projets publics qui publient leur code source en accès libre (sur Github) comme LexImpact, un outil qui permet d'estimer l'impact de réformes législatives sur des foyers fiscaux types. Ou encore La Bonne Boîte, un outil qui indique les entreprises qui embauchent pour permettre aux demandeurs d'emploi de mieux cibler leurs candidatures spontanées.

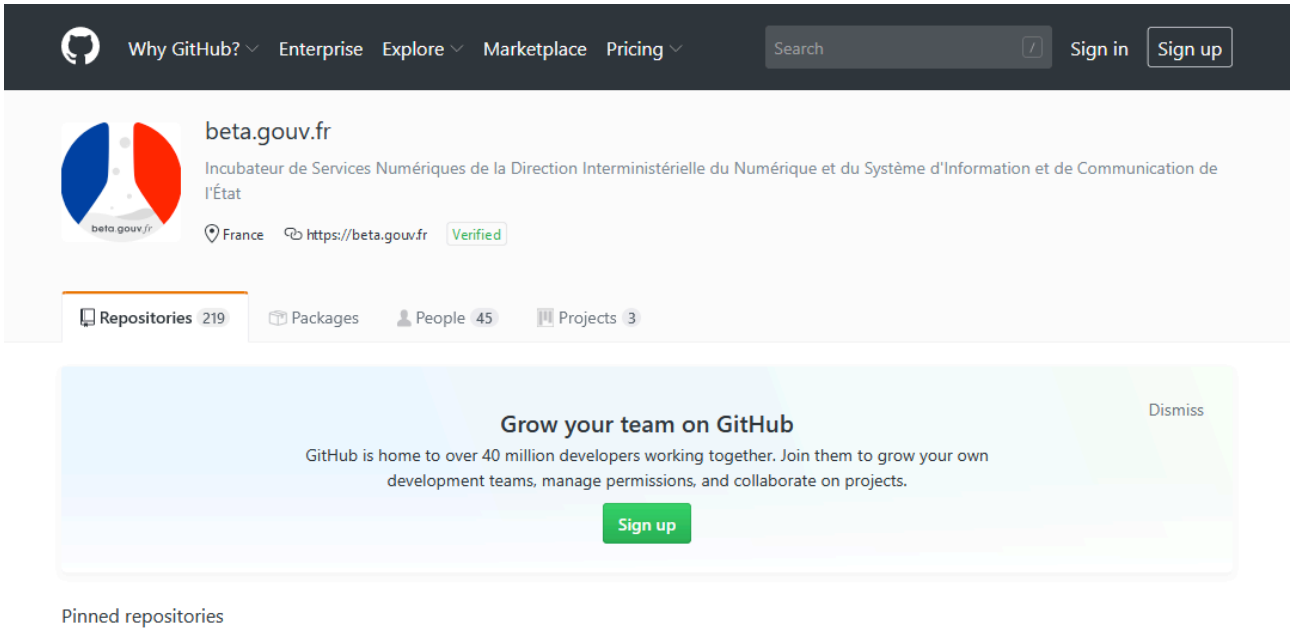


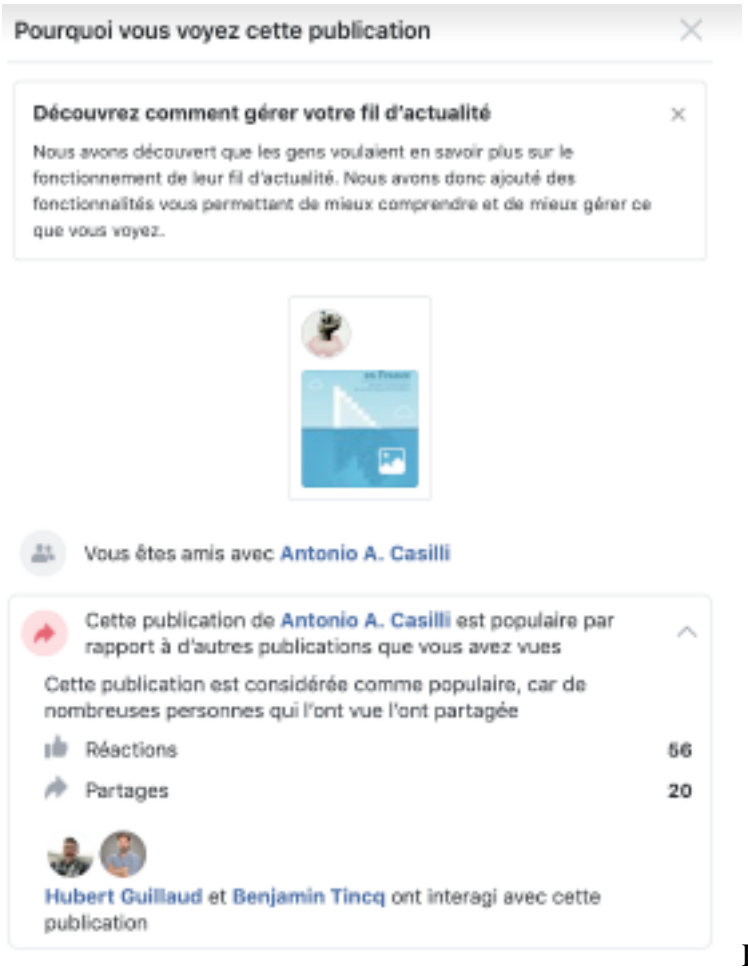
Image : Le répertoire GitHub des startups et projets d'Etat.

La vraie limite de cette explicabilité technique est qu'elle ne s'impose qu'aux projets qui ont une obligation légale de transparence, et particulièrement aux projets portés par des acteurs publics ou militants, et pas du tout à la plupart des projets privés, même si ceux-ci ont un impact fort sur la vie des gens (comme des systèmes d'attribution de crédit, de police ou de justice, ou de sélection à l'emploi).

La multimodalité des niveaux d'explications

Nombre de systèmes font néanmoins des efforts pour expliquer ce qu'ils font. Certains livrent des explications très simples et basiques, accessibles au plus grand nombre, mais souvent lacunaires voire parcellaires (et là on peut se poser des questions sur l'intention de ces explications et surtout sur leurs limites, visant parfois à masquer ce qui n'est pas expliqué). Beaucoup proposent des explications procédurales, c'est-à-dire qui expliquent les procédures à suivre, plus que les finalités ou les objectifs du système. Certains proposent des explications assez complètes, à l'image du Score Coeur (.pdf), l'algorithme d'attribution des greffons cardiaques réalisé par l'Agence de biomédecine ou encore cette notice détaillant le calcul de la taxe d'habitation sous forme d'un mode d'emploi synthétique… Deux exemples qui soulignent combien l'explication, quand elle se veut

loyale et exhaustive, scientifique, demeure néanmoins un processus touffu, dense, compliqué et qui est donc, par nature, difficile d'accès au plus grand nombre.



Enfin, bien plus rares sont les explications qui permettent de personnaliser les éléments du calcul qui ont été opérés, de délivrer des explications personnalisées donc. Ainsi, Facebook par exemple propose de multiples formes d'explications et d'aide sur son fonctionnement (sur l'utilisation des données, sur le fonctionnement du fil d'actualité et sur l'ordre des publications dans le fil d'actualité comme sur les publicités). Cette large gamme d'explications plutôt claires (on y trouve à la fois des textes, des vidéos, des datavisualisations, des graphes… et même des modules interactifs comme le fait d'expliquer par un exemple d'un article de votre flux provenant de votre compte ce qui justifie que vous le voyez apparaître sur votre mur), est-elle pour autant accessible à tous les utilisateurs de FB ? Est-elle complète ? Est-elle loyale ? Ainsi, FB explique par exemple que l'ordre des publications dans notre fil d'actualité est lié à la fréquence d'interaction, au type de publication, aux réactions qu'ont déjà reçu ces publications et à leur caractère récent. Mais des milliers d'autres facteurs entrent en ligne de compte explique la page. Aucun mot n'est dit par exemple sur le fait que FB nous partage plus volontiers des contenus de nos amis qui rejoignent des thématiques qu'il a identifiées comme importantes pour nous, alors que c'est là un critère fort de l'affichage. Par contre, s'il nous propose des modalités pour réguler cet affichage, notamment en sélectionnant mieux les personnes, les pages ou les groupes que l'on suit, mais pas les thématiques qu'il nous attribue. Les explications sur le fonctionnement de FB sont pourtant parmi les plus riches qu'on puisse trouver, mais on voit rapidement leurs limites. Qu'est-ce qui nous est tue ? Qu'est-ce que l'effort d'intelligibilité masque ? Pourquoi le lien entre ce qui nous est affiché et le modèle économique de Facebook n'est-il pas plus clair ?

Simulation et jouabilité : l'explication actionnable

Les simulations et les possibilités de jouabilité sont plus rares encore, mais elles sont

particulièrement stimulantes notamment parce qu'elles offrent souvent l'avantage de permettre une compréhension plus fine et donnent à celui qui reçoit les explications des moyens pour contextualiser et adapter sa réponse aux explications. Les systèmes jouables permettent souvent de montrer un peu mieux la complexité qui opère. La jouabilité a pourtant des limites : rares sont les explications qui expliquent et motivent les choix qui ont été faits, les décisions qui ont été prises, les critères qui ont été retenus et écartés. Rares donc sont les spécifications qui proposent de rendre les choix qui ont été faits plus transparents. Peu proposent des descriptions du calcul, des données utilisées. Encore moins proposent de montrer ou d'évaluer une autre méthode de calcul qu'il aurait été possible de proposer.

Dans cette gamme d'explications jouables, on trouve par exemple, les explications à explorer (voire notre dossier : Vers un design de la médiation, jouer avec les interfaces). Des explications interactives permettant une compréhension active via des procédés interactifs, afin de développer *« une intuition sur le fonctionnement d'un système »*, comme le disait leur concepteur, Victor Bret. Parmi les nombreux outils de ce type, on peut également signaler celui réalisé par le studio Dataveyes pour Outbrain, l'agence de recommandation de contenus sponsorisés qui avait besoin d'expliquer à ses clients comment leur algorithme allait placer leurs contenus publicitaires (voir nos explications détaillées : « Vers des algorithmes exemplaires »). Récemment, la *Technology Review* a proposé une explication explorable pour aider à comprendre les limites de Compas, le très controversé algorithme de calcul du risque de récidive utilisé par la justice américaine. Dans le cadre de son effort vers une *« intelligence artificielle de confiance »*, IBM Research a publié une boîte à outils open source permettant de comprendre plusieurs modèles d'apprentissage automatique, notamment un petit outil pour comprendre les résultats de l'évaluation Fico, qui attribue un score à chaque Américain souhaitant emprunter (selon divers critères, notamment l'historique de paiement et les emprunts que vous avez déjà à charge). Le petit démonstrateur montre d'ailleurs que les explications du calcul sont différentes selon qui l'utilise : il fournit des réponses différentes selon qu'on est un spécialiste des données, un agent chargé de calculer le crédit ou un client.

L'explication forte

On pourrait enfin proposer un ultime niveau sur l'échelle de l'explication : une forme d'explication forte, qui reprendrait toutes les qualités des précédents niveaux et y ajouterait une forme de garantie, de responsabilité à l'égard de ceux à qui on explique. C'est-à-dire qui garantirait un droit de réponse, de contestation, d'amélioration… Voire même des contreparties ou des procédures facilitant la contestation ou les compensations, en cas de défaut d'explication. Bref, des formes de *« symétrie »* des traitements où les explications délivrées aux calculateurs seraient équivalentes aux explications livrées aux calculés, où les stratégies des calculateurs seraient transparentes aux calculés et garanties.

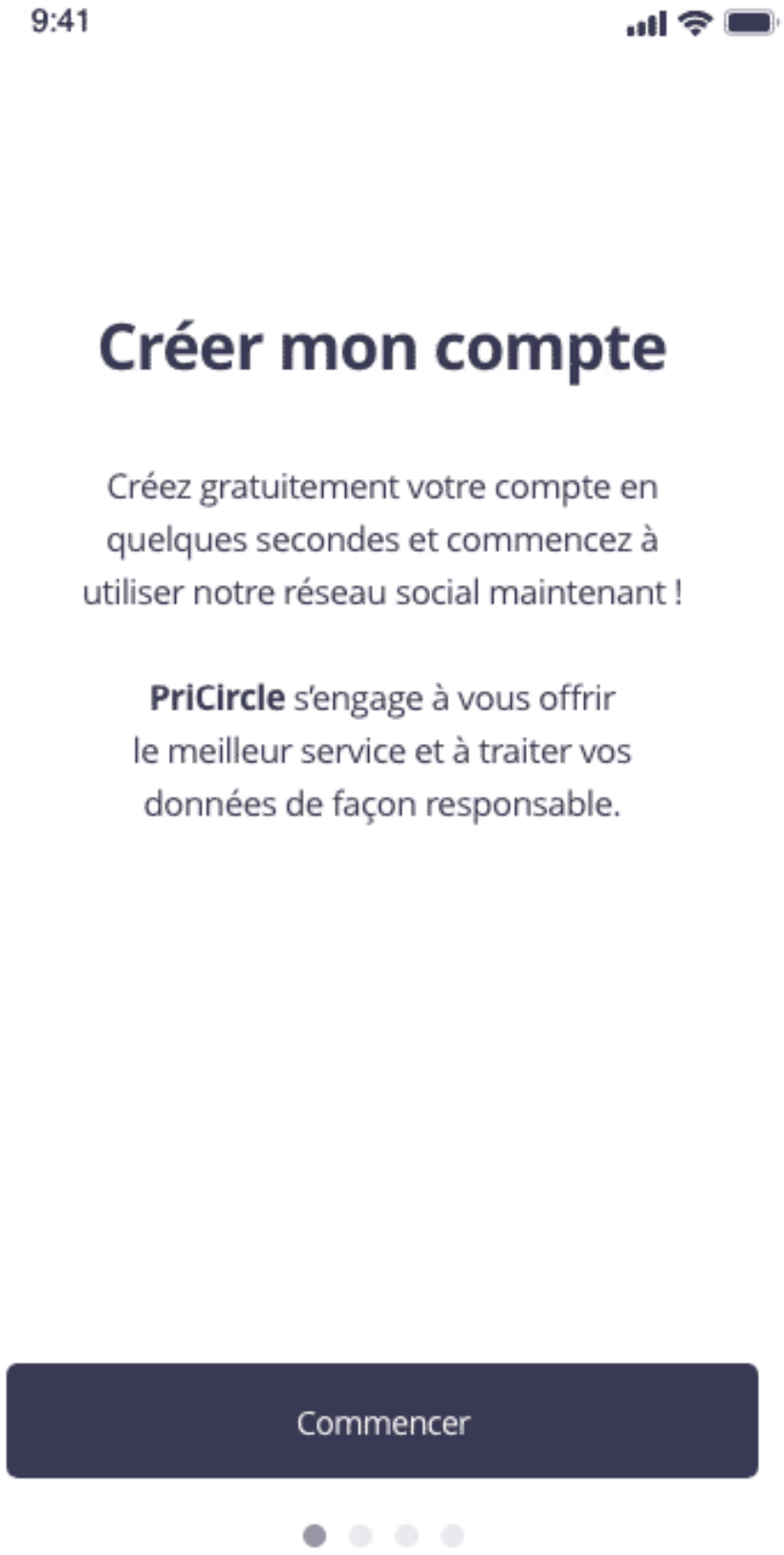
Pour l'instant, nous n'avons pas trouvé d'exemple pour illustrer cet idéal.

Les explications pour qui ? Les explications comment ?

Cette échelle des niveaux d'explication est pourtant incomplète. Il lui manque notamment deux facteurs : les publics et les interfaces.

La question de savoir à qui se destinent les explications est importante. Bien souvent, on ne trouve des explications que pour un seul niveau de public, alors que celui-ci est bien souvent multiple. Dans les systèmes techniques, l'explication demeure majoritairement technique : elle est avant tout destinée aux ingénieurs, aux développeurs voir aux agents qui vont être amené à interagir avec le système, qui vont devoir développer des outils qui vont s'y brancher, qui vont discuter avec. Bien souvent, elle vise à rendre les ambiguïtés interprétables plus qu'explicables, c'est-à-dire à expliciter les choix de paramètres, leur pondération… Mais malgré leur technicité, ces explications ne sont

pas complètes : elles cachent souvent le coeur du système qui relève du secret commercial, privilégiant seulement les options de branchements, les modalités de choix, les procédures accessibles (pas celles qui ne le sont pas). L'expérience montre que bien souvent un système automatisé doit rendre des comptes et des explications à plusieurs types de publics : des autorités de contrôle, des agents chargés de procéder ou d'accompagner le calcul et le public (qui n'est pas non plus unique : les explications ne sont pas les mêmes pour tous publics, pour des enfants ou des adultes par exemple, pour des patients ou pour leur famille, pour des personnes en interne et des personnes extérieures…). Les publics des calculs sont donc multiples, pluriels… Et les explications qu'il faut délivrer aux uns ne sont pas toujours de mêmes natures de celles qu'il faut délivrer à d'autres. Nous avons pour notre part constaté que nombre de projets ont tendance à réduire leurs explications à un utilisateur idéal, oubliant la pluralité des publics qui sont les leurs, et ce alors que la mise en place de systèmes automatisés perturbe toute la chaîne de compréhension du fonctionnement d'une organisation ou d'un calcul.



Enfin, la question de savoir

où sont disponibles ces explications est importante. Bien souvent, les explications sont décorrélées de l'outil lui-même, à la manière d'un mode d'emploi papier pour une machine électronique. Les explications sont trop souvent encore à aller les chercher à côté des systèmes. Il faut aller les trouver dans une nasse d'information complexe ou dans des documents extérieurs aux outils eux-mêmes. Elles sont rarement là où on en a besoin, au coeur des interfaces qu'on explore et qu'on ne comprend pas. La plateforme Données & Design de la Cnil donne un exemple parlant d'information contextualisée intégrée au parcours utilisateur : chaque demande d'information relative aux données personnelles lors du parcours d'inscription est explicite. À chaque étape de l'inscription, d'une interaction, des explications sont délivrées, exactement là où elles sont nécessaires.

Polysémiques, multifacettes, continues… : à quoi servent les explications ?

Expliquer semble simple. Mais comme on vient de la voir, à l'heure des systèmes techniques, l'explicabilité n'est pas seulement polysémique, elle est également multifacette.

Comme le soulignent Robert Hoffman, Shane Mueller et Gary Klein, dans leur article *« Explaining Explanation for « Explainable AI » »*, il est nécessaire d'interroger à quoi servent les explications, de comprendre ce qu'elles accomplissent, leur but. Ainsi, rappellent les chercheurs, l'explication est un processus continu. Fournir une explication n'est pas une fin en soi pas plus qu'il n'est une chose que l'on fait une fois pour toutes : cela ne consiste pas à fournir un matériel didactique plus ou moins satisfaisant. Une explication sert à créer de la confiance à toutes les étapes des interactions, ce qui signifie que l'utilisateur doit être capable d'explorer activement les choix dont il dispose et notamment les erreurs possibles.

L'explication est également un processus de co-adaptation : ce n'est pas seulement des instructions ou des informations qu'on délivre à une personne, c'est une collaboration entre celui qui explique, celui qui reçoit l'explication et le système. L'explication est par nature participante et par exemple, doit savoir s'adapter à ce que le destinataire comprend et même, idéalement, s'adapter en retour.

L'explication doit être « déclenchable » insistent-ils : tout n'a pas besoin d'être expliqué. Par contre, il faut saisir les éléments qui nécessitent de déclencher des explications appropriées.

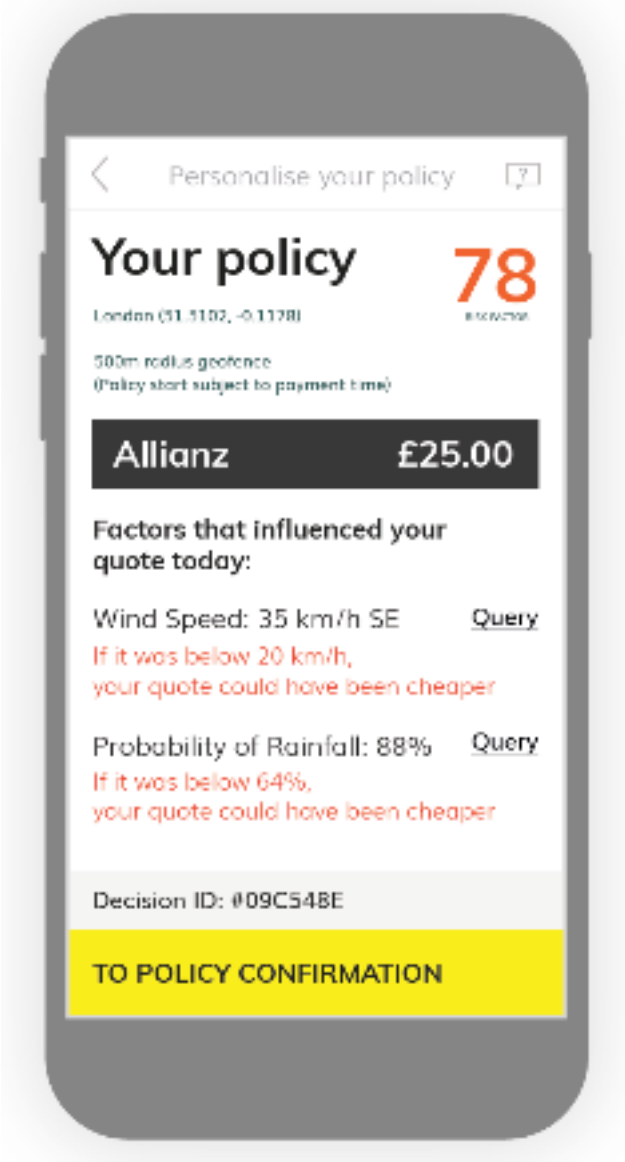
L'explication est également une exploration : elle doit aider l'utilisateur à comprendre les limites du système. Elle doit aider l'utilisateur à comprendre de lui-même : montrer ce qui n'est pas fait pour mieux montrer ce qui est accompli.

Pour le dire plus simplement, une explication sert à comprendre, à discuter et à contester. Elle n'est pas un outil de communication qui doit faire disparaître ses biais, ses erreurs ou ses choix. Elle est un outil loyal qui doit montrer ses lacunes et ses failles, qui doit aider à améliorer le système et le calcul.

Qu'est-ce qu'une bonne explication ? Les caractéristiques d'une bonne explication

Vers des explications contrastables, actionnables
L'iconoclaste Tim Miller (@tmiller_unimelb) a étudié plus de 250 articles scientifiques dans le domaine de la philosophie, de la psychologie et des sciences cognitives pour mettre en avant que la principale caractéristique d'une bonne explication est d'être *contrastive* (qu'on pourrait traduire par « contrastable » ou « comparable »). Les gens ne demandent pas « pourquoi il y a quelque

chose ? », mais plutôt « pourquoi il y a quelque chose plutôt que rien », rappelle-t-il. C'est-à-dire qu'une explication ne se déroule pas pour elle-même, mais en contexte, en regard d'autres éléments. Le caractère « contrasté » permet aux gens de mieux mesurer la différence entre ce à quoi ils s'attendaient et le résultat (et d'interroger ce qu'ils ne savent pas). Donner une explication contrastée est souvent d'ailleurs plus simple pour celui qui doit fournir l'explication. L'enjeu n'est pas de donner toutes les explications ou toutes les causes, mais de permettre de les balancer entre elles, de rendre les explications actionnables, jouables !



Pour mieux comprendre ce qu'est une explication contrastable ou actionnable, donnons un exemple. L'un des meilleurs est celui qu'a conçu l'agence de design britannique IF pour Flock, une assurance pour drone qui peut être souscrit à la volée. Ce travail a consisté à montrer, dans le design même de l'application, les enjeux et impacts des décisions. Ainsi, l'interface calcule un prix pour assurer un drone à la volée, au moment où l'on souhaite l'utiliser, selon le lieu et les conditions météo du moment (voir ces explications). Mais, une fois qu'elle a calculé le prix de l'assurance, elle vous indique également que si le vent était moins fort, ou que si la probabilité de pluie était moins forte, vous payeriez moins cher. D'autres paramètres sont pris en compte pour vous expliquer la tarification qui s'applique à vous : la proximité d'une école ou d'un hôpital peut faire augmenter le prix et l'application peut vous indiquer que si vous vous en éloignez, cela pourrait en faire diminuer le prix. Dans le cadre d'un travail de prototypage pour une assurance auto basée sur le comportement de l'automobiliste, l'agence IF a également esquissé des interfaces qui permettent au conducteur de voir la différence

entre son comportement et la norme. Ces exemples, stimulants, permettent de montrer quels facteurs contribuent aux décisions et comment les atténuer ou agir sur ces éléments.

Vers des explications sélectives – mais loyales
Contrairement à ce à quoi on pourrait s'attendre, une bonne explication pour Miller n'est pas une explication complète ou exhaustive. Une bonne explication est « choisie », sélective (et donc biaisée). Les gens ne s'attendent pas à voir une cause unique et complète ni une liste causale exhaustive. Son caractère « sélectionné » est bien sûr un compromis. Son caractère partiel ou incomplet doit pourtant rester fidèle et loyal à ce qu'il se passe réellement.

Les probabilités n'ont pas tant d'importance que l'on croit, explique encore Tim Miller. Si la vérité est importante, se référer aux probabilités ou aux relations statistiques d'une explication n'est pas aussi efficace que se référer aux causes.

Vers des explications sociales
Enfin, les explications sont profondément sociales… Elles relèvent de l'échange, d'un transfert de connaissance, d'une interaction. En ce sens, elles demeurent profondément contextuelles. A toute question posée, plusieurs réponses peuvent être produites, rappelle Miller. En ce sens, la médiation et l'explication sont profondément dépendantes l'une de l'autre. Un système de décision automatisé ne peut donc faire l'économie de médiations adaptées, d'équipes et de solutions dédiées aux dialogues avec les usagers, et se doit d'organiser ce dialogue entre son système et la société.

Tant mieux. Cela nous rappelle que l'explication n'est en rien réductible à une simple technique, à une question objectivable, mais a tout à voir avec une relation et sa complexité.

Hubert Guillaud
(*Avec la complicité de Simon Chignard et Thierry Marcou*).

PS : Amis lecteurs, si vous avez vu passer des services qui proposent des modalités d'explications remarquables, n'hésitez pas à nous les signaler.

- algorithme
- nossystemes

À lire aussi sur internetactu.net

- **Attention à l'attention**
- **Technosciences : de *la démocratie des chimères*… aux chimères de la démocratie**
- **Vers une science de la causalité**
- **La géo-ingénierie à petite échelle**
- **Vers des interactions automatisées et empathiques à la fois**
- **La vie, telle qu'elle pourrait être (1/3) : de nouvelles lettres pour l'alphabet du vivant**

- **Défaire la tyrannie du numérique ?**

- **Ce que les catastrophes technologiques disent de la technologie**

- **Interdire la reconnaissance faciale (1/3) : la reconnaissance faciale n'est pas une technologie, c'est une idéologie !**