Version pénultième – à paraître dans *Éthique publique*

Agencéité et responsabilité des agents artificiels dans un monde de connaissances

Louis Chartrand

Résumé

Les agents artificiels et les nouvelles technologies de l'information, de par leur capacité à établir de nouvelles dynamiques de transfert d'information, ont des effets perturbateurs sur les écosystèmes épistémiques. Se représenter la responsabilité pour ces chambardements représente un défi considérable : comment ce concept peut-il rendre compte de son objet dans des systèmes complexes où il est difficile de rattacher l'action à un·e agent·e ? Cet article présente un aperçu du concept d'écosystème épistémique et de la force de changement que représentent les nouvelles technologies pour celui-ci, puis illustre les difficultés rencontrées avec la notion classique de responsabilité dans les écosystèmes épistémiques.

Abstract

Artificial agents and new information technologies, through their capacity to foster new information dynamics, have disruptive effects on epistemic ecosystems. Making sense of responsibility for these changes represent a challenge: the notion of responsibility struggles in complex systems, where actions are often hard to connect to an agent. This paper presents an overview of the concept of epistemic ecosystem and of the potential for disruption in new technologies, and then illustrates the difficulties with the classical notion of responsibility in such cases.

Introduction

La connaissance, quand on la considère comme un phénomène social, est quelque chose de vertigineux. Pour s'en convaincre, il suffit de partir du *leitmotiv* de l'épistémologie sociale : « le plus gros de notre connaissance nous provient de son partage par autrui », et de considérer ce que l'on sait sur des choses qui nous sont extrêmement distantes. Par exemple, je sais, avec certitude, que Jules

César a battu Vercingétorix à Alésia. Pourtant, 2 000 ans d'histoire me séparent de cet événement qui concerne des peuples qui n'existent plus et qui s'est produit dans un lieu lointain que je n'ai jamais visité. Pour que cette information me soit transmise, il a fallu le travail de nombreux témoins, scribes, chercheur·ses et enseignant·es qui se la sont passée pour me la transmettre.

Or, on sait que notre mémoire humaine nous fait souvent défaut, qu'en contant, on embellit volontiers, qu'en écoutant, on interprète avec les concepts auxquels on a accès, et qu'en conséquence on comprend souvent mal. Comment se fait-il qu'on soit capable de savoir, alors que les chaînes de transmission qui nous séparent de certains événements sont si longues, et les chances de défaillances si grandes ? Pourquoi ne tombe-t-on pas, pour reprendre l'expression de Lorraine Code (Code, 2006), dans un *chaos épistémique* ?

La réponse de Code est qu'individuellement, nous sommes tenu·es à des responsabilités épistémiques, c'est-à-dire que nous avons des obligations spécifiques envers nous-mêmes, envers les gens avec qui on transmet des informations, et envers la société en général, et que ces obligations relèvent de la fonction d'une société à produire et à cultiver la vérité.

Ces obligations peuvent prendre des formes différentes compte tenu des moyens et des contraintes qui sont données à la communauté qui veut produire de la connaissance.

Dans la suite, nous abordons, dans une perspective d’élargir le débat au-delà des frontières disciplinaires et institutionnelles, la capacité qu’ont les nouvelles technologies à agir sur et à l’intérieur de ces réseaux d’obligations et de contraintes, et les difficultés à appliquer une notion traditionnelle de responsabilité pour régir ces capacités. Dans un premier temps, nous présentons un aperçu de l’écosystème épistémique qui soutient nos pratiques de production et de maintien des connaissances. Puis, nous donnons un aperçu du potentiel perturbateur des nouvelles technologies et des agents artificiels pour cet écosystème. Nous faisons ensuite un survol du concept traditionnel de responsabilité et de ses limites lorsqu’on l’emploie pour un agent artificiel. Enfin, nous présentons quelques unes des pistes de solutions radicales qui s’offrent à nous.

Écosystème épistémique

L'obligation la plus simple consiste en celle qui se rapporte au rapport de l'individu à l'objet qu'il connaît: la communauté ne peut développer de bonnes connaissances si les gens qui y participent se représentent mal les objets qu'ils côtoient. Donc, si la réfraction de l'eau me donne l'impression que ma cuillère est tordue, j'ai une obligation de regarder d'un autre angle ou de sortir la cuillère pour expliquer l'anomalie et, d'une certaine façon, calibrer mes sens.

Cependant, comme la plupart de nos connaissances nous sont transmises par d'autres, il faut également que je puisse avoir certaines garanties de ce que je sais d'autrui. Par exemple, je sais de mon père que, pendant l'été, mon grand-père employait des joueurs des Canadiens de Montréal. Venant d'une autre personne, je pourrais facilement croire qu'il s'agit d'un mensonge ou d'un embellissement, mais j'ai une relation avec mon père qui me permet d'avoir confiance en sa parole. D'une part, je sais quand mon père ment, quand il embellit, et je sais les types de connaissances pour lesquelles il est moins fiable. Mais surtout, notre relation est le fruit d'une coopération : si mon père devait faillir à sa responsabilité épistémique envers moi, ma confiance serait ébranlée, et notre relation en serait affectée. En ce sens, notre relation permet à mon père de prendre la responsabilité de la véracité de ce qu'il me dit. Dans une communauté isolée et de petite taille, ce genre de mécanisme peut permettre de transmettre des connaissances sur plusieurs générations. Mais qu'en est-il de connaissances provenant de parfaits inconnus qui ne rencontrent jamais l'essentiel de leur audience ? Une partie de nos connaissances nous provient non pas de conversations avec des gens que l'on connaît, mais de nos lectures ou de paroles entendues à la télévision. Le papier et les autres technologies de diffusion jouent alors un rôle de médiation. Dans une telle situation, le rapport entre agent·es épistémiques est réduit, mais le coût de communiquer de l'information baisse lui aussi, étant donné qu'un seul acte de communication peut rejoindre des milliers de personnes. Le coût de l'information étant moindre, elle devient plus accessible, ce qui entraîne également des opportunités pour développer de nouveaux moyens de valider les connaissances qui nous sont transmises.

Une première opportunité provient de la possibilité qui est ainsi ouverte de raffiner l'évaluation de la crédibilité des sources. Bien que l'on ne dispose plus de notre propre expérience d’une relation profonde avec la personne qui nous transmet de l'information pour en évaluer sa crédibilité, on peut néanmoins connaître

sa réputation auprès de son audience. Celle-ci se manifeste souvent par le biais de reconnaissances institutionnelles, qui jouissent elles-mêmes d'une réputation publique. Par exemple, qu'une personne s'exprime comme journaliste du Devoir ou du Monde lui donne une certaine crédibilité en vertu de la réputation du journal, qui est attestée entre autres par l'opinion de la communauté des journalistes. À ce niveau, ce n'est donc plus la relation personnelle entre deux personnes qui est garante de la qualité de l'information, mais la relation entre une agente, un agent ou une institution d'une part et un certain public d'autre part. La réputation donne donc, comme les relations entre gens qui se connaissent bien, des raisons de croire qui ne relèvent pas d'éléments probants. Sauf qu'ici, plutôt que de faire porter le poids de l'assurance sur la relation entre deux personnes, on la fait porter sur la relation entre la source et l'ensemble de la communauté.

Par ailleurs, il devient plus facile de trouver de l'information pour corroborer celle qui nous a déjà été transmise. Ainsi on peut, par exemple, aller chercher des points de vue différents sur un même événement pour voir s'ils concordent sur les points importants. On peut également remonter la piste de transmission de connaissances qui a permis qu'on acquière ce que l'on sait de cet événement, et déterminer les raisons qui ont justifié la crédibilité et la validité de ce savoir en amont. Par exemple, je pourrais croiser l'histoire de mon père avec d'autres témoignages, venant de ma tante ou de ma grand-mère, qui ont peut-être elles aussi croisé des joueurs du Canadien au commerce de mon grand-père.

Ce processus de vérification peut lui-même être sous-traité vers la communauté, comme on le voit dans des entreprises de connaissance qui demandent des ressources considérables (par exemple, Wikipedia, ou le *Large Hadron Collider*). Les tâches contribuant à la vérification et à l'amélioration des connaissances sont alors accomplies par des personnes différentes et qui, très souvent, ne se sont jamais rencontrées. Par exemple, au sein de la communauté de Wikipedia, on retrouve des gens qui se spécialisent à créer des pages sur un sujet qui les intéresse, d'autres qui soulignent les inconsistances, d'autres qui recherchent des sources, etc. (Goldenberg, 2011) La confiance que l'on peut avoir ne repose donc plus sur la réputation d'une personne ou d'une institution, mais sur la confiance que l'on a envers le processus de développement des connaissances.

Bref, dans une société de connaissance, les obligations et responsabilités épistémiques des individus se conjuguent dans des relations qu'ils entretiennent soit avec un autre individu, soit avec une communauté qui possède les moyens de faire un suivi de sa crédibilité. Ces relations peuvent être médiatisées par des

technologies de diffusion, comme le papier ou le web, ce qui contraint les stratégies de vérification, mais crée la possibilité de nouvelles stratégies. Et enfin, ces obligations peuvent être sous-traitées et redistribuées de différentes façons, notamment en effectuant un partage des tâches épistémiques.

Les obligations et responsabilités se rapportent ainsi à un écosystème épistémique (cf. Code, 2006). À travers ces différents rapports d'obligations et de responsabilités épistémiques, les humains participent à des dynamiques qui ont pour objectif de cultiver et de développer les connaissances de leur communauté ou de leur société. Ces dynamiques, prises ensemble, favorisent l'essor de certains types de connaissances, guident les individus vers certains rôles, favorisent certaines vertus et capacités épistémiques, etc. En ce sens, une responsabilité épistémique joue un rôle dans le projet plus général de culture de la connaissance dans une société. Aussi, pour rendre compte de la responsabilité épistémique, il faut comprendre la fonction propre (au sens de Millikan, 1984) de cette responsabilité à l'intérieur de son écosystème épistémique, c'est-à-dire la fonction ou l'ensemble des fonctions pour laquelle ou lesquelles cette responsabilité s'est développée et a pris sa place dans l'écosystème épistémique.

Perturbations technologiques

Les nouvelles technologies interviennent dans ces écosystèmes, et risquent souvent, pour le meilleur ou pour le pire, de perturber leur équilibre et d'engendrer des changements importants dans leurs dynamiques. Pour l'essentiel, ce potentiel de perturbation provient du rôle de médiation que prennent les technologies. Elles s'interposent entre le sujet et l'objet, ou entre source et audience d'une façon qui change les rapports et les pratiques épistémiques.

Perturbations de premier ordre : médiation entre humain et objet

Le rôle de médiation des nouvelles technologies se présente d'abord au niveau de la perception: elles nous permettent de magnifier des détails que l'on ne saurait voir autrement, de détecter des données portant sur des objets auxquels nos sens sont aveugles, ou de reconfigurer les données de façon à faire apparaître des associations ou des formes que l'on n'aurait pas su voir autrement. Par exemple, on peut employer un télescope pour magnifier l'image d'une étoile ou pour détecter des radiations en dehors du spectre lumineux visible. On peut ensuite prendre de telles images d'une même étoile sur plusieurs jours et les traiter

à l'ordinateur pour détecter la réflexion d'une exoplanète dans le halo lumineux de l'étoile. Ces technologies changent notre rapport avec l'objet, qui n'est plus soumis aux modalités habituelles de l'observation avec les yeux. Pour avoir confiance en la représentation qu'ils nous donnent, il faut avoir ajusté les instruments de façon à contrôler les facteurs externes qui pourraient les affecter, et avoir validé ces ajustements à l'aide d'autres observations (Humphreys, 2004).

Ce genre de médiation devient perturbatrice notamment lorsque les nouvelles technologies, devenant incontournables, valorisent une expertise technique au détriment d'autres connaissances. Ce transfert dans la valorisation des savoirs et savoir-faire s'accompagne souvent d'un transfert d'autorité, puisque crédibilité et autorité viennent souvent de pair : les personnes qui détiennent le savoir technique en viennent ainsi à prendre des décisions là où d'autres personnes détenant d'autres savoirs auraient pu aspirer à avoir ce pouvoir décisionnel. Un exemple fréquemment discuté dans la littérature féministe provient d'accompagnement médical de l'accouchement : la technicisation de celui-ci a donné un pouvoir important aux médecins et au personnel médical. Ce pouvoir donne parfois lieu à des abus, lesquels peuvent aller à l'ignorance des choix de la patiente à des interventions abusives, voire dangereuses, comme le « point du mari », ou humiliantes, comme la stérilisation forcée. Ici, le transfert d'expertise contribue à décrédibiliser les patientes à un tel point que leurs connaissances de leur propre corps et de leur vie sont évacuées du processus de décision.

Perturbations de second ordre : médiation entre humains

Le potentiel perturbateur est d'autant plus grand si la technologie concerne directement le rapport entre deux agent·es épistémiques. On a vu comment l'introduction de médias qui permettent de rejoindre de plus larges audiences a pu dégrader la relation entre source et audience parce qu'il ne permet plus à l'audience de développer une relation personnelle avec la source qui pourrait garantir l'authenticité des propos. Inversement, la réintroduction sur le web d'un canal de communication de l'audience vers la source, permettant de commenter un article ou de s'adresser directement à son auteur·e par les réseaux sociaux a perturbé à nouveau la relation entre auteur·e et audience : par exemple, des journalistes, chroniqueur·ses, blogueur·ses et autres voix dans le débat public sont souvent la cible de flots de menaces, commentaires haineux, divulgation de données confidentielles, etc. Ces messages, qui font parfois l'objet de stratégies coordonnées, sont souvent envoyés avec l'intention assumée de décrédibiliser ou de forcer la

personne en question à se retirer du débat public, et donc à empêcher que certains témoignages et certaines connaissances y soient traitées. En l'absence de stratégies efficaces pour répondre à ces attaques, les victimes sont souvent forcées d'abandonner leur tribune, et le débat public s'en trouve appauvri.

Cependant, on aurait tord de penser que les dangers se limitent aux emplois de la technologie qui vont au-delà des usages attendus par les gens qui l'ont conçue. Les usages attendus peuvent aussi avoir des conséquences indésirables. Par exemple, dans un contexte de polarisation au niveau politique, on critique souvent les réseaux sociaux parce qu'ils favorisent des phénomènes de « caisses de résonance ». Étant confrontés aux opinions et aux pensées de gens qui pensent comme eux, les gens qui emploient régulièrement les médias sociaux ont alors plus de mal à contempler des positions alternatives et à évaluer de façon critique leurs propres croyances. Il arrive également que l'usage attendu d'une technologie – qui est souvent, après tout, financée par des gens en position de pouvoir – vienne exacerber une hiérarchie déjà présente et effacer les contributions épistémiques des gens qui sont dans le bas de l'échelle (Noorman, 2016; Bovens et Zouridis, 2002). C'est pourquoi les logiciels conçus pour le fonctionnement interne des entreprises suivent souvent une tendance plus générale de centralisation des processus décisionnels. Ainsi, il arrive souvent que les employé·es au service à la clientèle se trouvent avoir très peu de marge de manœuvre, même s'illes sont les mieux positionnés pour connaître les besoins de la clientèle, et ce parce que les logiciels qui enregistrent et effectuent les transactions ne proposent sciemment qu'un nombre très limité d'options.

Par ailleurs, on fait souvent grand cas de la façon dont certains sites (dont Wikipédia) ont organisé le travail épistémique des humains de façon à permettre l'émergence de connaissances fiables, et ce sans exploiter des indices de réputation traditionnels, comme la formation ou l'affiliation à une institution. Dans ces cas de figure, l'organisation des processus de production et de vérification de connaissances repose énormément sur la conception du site, qui dirige les contributrices et contributeurs vers des actions désirées : créer une page manquante, ajouter une source, discuter d'un problème, etc. Cependant, l'abandon des indices de réputation traditionnels signifie qu'on se prive aussi de leurs acquis. Ainsi, même dans des domaines où il y a une expertise considérable provenant de femmes, elles sont peu nombreuses à contribuer à l'édition du contenu de Wikipédia (Hill, 2013). Or ces différences sont souvent attribuées aux rapports antagonistes que les gens qui y écrivent entretiennent entre eux pour imposer

leur vision à un article ou pour acquérir de la réputation sur le site. En conséquence, on observe une sous-représentation des contenus portant sur des femmes notoires comparativement, par exemple, à l'*Encyclopedia Brittanica* (Reagle et Rhue, 2011).

Perturbations des agents épistémiques artificiels

Le rôle médiateur des nouvelles technologies, incluant celles qui relèvent de l'intelligence artificielle, est donc un facteur important à prendre en compte lorsque l'on réfléchit sur leurs apports à nos écosystèmes épistémiques. Cependant, les développements de l'intelligence artificielle ouvrent la voie à de nouveaux types d'implication dans les écosystèmes épistémiques.

Ainsi, les programmes informatiques peuvent jouer un rôle plus actif du moment où on leur donne des tâches à accomplir. À partir de scripts très précis, on peut amener un programme à effectuer des tâches répétitives ou à prendre des décisions à partir des données qui lui sont disponibles. Par exemple, Wikipédia possède de nombreux « robots » qui sont souvent cantonnés aux tâches de maintenance (classification des articles, gestion du vandalisme, etc.), mais qui jouent parfois des rôles plus élaborés, comme la création de nouvelles pages à partir de contenu externe. Ces programmes possèdent une certaine autonomie, dans la mesure où ils prennent des décisions éditoriales.

Comme ces « robots » ne font que suivre leur programmation, on peut douter qu’ils disposent de suffisamment d’autonomie pour s’affranchir des intentions des gens qui les conçoivent. Ceci dit, d'une part, même un robot possédant un script clairement défini peut former des interactions complexes et imprévisibles. Les réalités du terrain échappent souvent aux attentes des gens qui font la conception. Par exemple, dans la « botosphère » de Wikipédia, il arrive souvent que les robots se contredisent et se livrent des guerres d'éditions (Tsvetkova et collab., 2017). D'autre part, les progrès de l'apprentissage machine ouvrent la voie à des formes d'autoprogrammation. L'idée est alors de développer des comportements, des stratégies ou représentations de façon à optimiser certains indicateurs, lesquels peuvent indiquer une ressemblance avec des comportements ou certains succès communicationnels. Ce genre de robot est, de par sa conception, imprévisible, puisque le travail de déterminer son mode d'action lui revient. Cependant, sa capacité d'apprentissage lui permet de résoudre davantage de problèmes et de s'adapter à différents environnements, et donc à jouer davantage de rôles dans l'écosystème épistémique.

D'une façon ou d'une autre, ces robots en viennent à des actions qui comptent pragmatiquement comme des actes de langage contribuant aux dynamiques épistémiques. Dans Wikipedia, une édition faite par un robot compte au même titre qu'une contribution humaine. Mais même dans les conversations où ils sont identifiés comme tels, les humains ont tendance à interpréter les actes de langage des robots comme des actes de langage humains. Comme le note Zoe Quinn, une victime de Gamergate, le harcèlement psychologique qui est produit par des robots est quand même ressenti comme du harcèlement psychologique. En conséquence, il cause des effets épistémiques similaires.

Le danger de l'évolution des agents artificiels, dans ce contexte, est que chaque nouvelle trouvaille, qu'il s'agisse d'une technique qui étend leurs capacités d'interaction ou simplement d'une application ingénieuse dans les réseaux sociaux, a un potentiel perturbateur. Lorsqu'un tel développement survient, la dynamique de la communauté épistémique est changée, et les processus par lesquels elle en vient à accepter, diffuser et enrichir certaines connaissances sont affectés; certaines personnes gagnent en crédibilité et en impact, d'autres y perdent. En ce sens, le développement d'agents artificiels peut être un outil d'émancipation s'il permet de déconstruire la hiérarchie des débats publics, de donner la parole aux gens dont les connaissances et les expériences sont ignorées, et de faciliter le dialogue et la compréhension entre gens qui n'ont pas l'habitude de se côtoyer. En revanche, il peut aussi, au contraire, être employé pour contraindre au silence des voix légitimes, faciliter la diffusion de propagande ou crédibiliser des informations fausses ou trompeuses.

Aussi, l'important à noter ici n'est pas que les nouvelles technologies et l'intelligence artificielle ont nécessairement un impact négatif sur le développement et la culture de nos connaissances. Au contraire, des communautés comme Wikipédia les emploient avec succès pour promouvoir ces fins. C'est plutôt qu'il faut tenir compte des effets perturbateurs, parce qu'ils créent des vulnérabilités qui, une fois exploitées, peuvent avoir des effets dévastateurs.

Responsabilité

Ayant fait un portrait rapide des écosystèmes épistémiques et des risques que posent pour eux les applications de l'intelligence artificielle, je vais maintenant passer au concept de responsabilité, et au rôle qu'il peut jouer pour structurer le débat sur la gestion des risques liés aux perturbations des nouvelles technolo-

gies. Pour ce faire, il faut d'abord clarifier ce que l'on entend par « concept de responsabilité », puisque l'on emploie ce terme pour désigner différentes choses.

Distinctions

Premièrement, il faut distinguer entre une responsabilité conventionnelle, qui correspondrait à peu près à celle dont il est question dans le droit civil, d'une responsabilité plus « réelle » qui rend compte de l'apport d'un·e agent·e à une situation. La responsabilité conventionnelle (*attributability*) pour un événement est accordée à ou appropriée par une personne qui n'a pas nécessairement besoin d'avoir un rôle dans l'histoire causale de l'événement. La personne responsable au sens conventionnel est celle à qui incombe le fardeau des conséquences. Par exemple, dans l'arrêt Donoghue c. Stevenson (1932), un producteur de bière de gingembre est tenu responsable de la présence d'un escargot dans un de ses produits, et ce parce que la santé publique demande qu'une attention particulière soit mise sur ce qui peut se retrouver accidentellement dans les produits que l'on mange et que l'on boit.

À l'opposé, la responsabilité réelle (*accountability*) n'est attribuable qu'à une personne qui a un rôle important dans l'histoire causale de l'événement. Elle désigne une connexion réelle entre la personne et ce dont elle est responsable[1].

Par ailleurs, on peut distinguer une responsabilité générale, qui relève des devoirs que toute personne a dans une société ou dans une communauté, d'une responsabilité circonstancielle qui relève des particularités d'une situation. La responsabilité circonstancielle peut découler, par exemple, des obligations qui résultent d'un contrat, ou des attentes d'un·e client·e. La responsabilité générale, en revanche, découle d'obligations qui viennent avec l'appartenance à une socié-

1 Ceci n'équivaut pas à dire de la personne qu'elle n'est responsable que de conséquences de ses actions. On peut être responsable de ce que l'on n'a pas fait – comme d'avoir oublié un anniversaire, par exemple. Dans un tel cas, la faute repose dans nos attitudes et notre caractère, de sorte que l'on reste l'agent·e principal dans l'histoire causale de cet oubli (cf. Smith 2005).

té, comme les obligations morales ou celles qui proviennent de notre statut d'agent·es épistémiques.

Enfin, à l'intérieur de la responsabilité générale et réelle, on distingue la responsabilité morale et la responsabilité épistémique[2]. Cette distinction repose non pas sur la façon dont on relie l'action à son agent·e, mais sur la façon dont on évalue l'action. Notre responsabilité morale à une action fait en sorte qu'elle nous révèle ou nous constitue dans notre caractère moral : le fait d'être une personne moralement bonne repose sur la qualité morale des actions que l'on fait. Similairement, nos actions épistémiques, c'est-à-dire celles qui contribuent positivement ou négativement aux fonctions de l'écosystème épistémique de notre communauté, nous constituent comme agent·es épistémiques. Aussi, nous sommes épistémiquement responsables de quelque chose dans la mesure où cette contribution à la culture et au développement de la connaissance révèle notre caractère comme agente ou agent épistémique. Autrement dit, ce qui fait la distinction entre responsabilité morale et responsabilité épistémique est le type de régime normatif auquel on réfère.

Composantes de la responsabilité

S'il est facile de faire certaines distinctions quant au sens du mot « responsabilité », il est beaucoup plus difficile de déterminer ce qui fait la responsabilité, notamment parce qu'elle est associée à des critères variés, qui renvoient à des dimensions très différentes de l'agencéité. Aussi, non seulement les philosophes

2 On pourrait interpréter la responsabilité épistémique comme une forme particulière de responsabilité morale, ou comme une forme de responsabilité distincte, par exemple fondée sur l'utilité des ressources que l'écosystème épistémique prodigue aux individus. Dans la première interprétation, les obligations qui relèvent de la responsabilité épistémique relèvent ultimement de notre responsabilité morale. Autrement dit, agir en agent·e épistémique responsable, c'est également agir en tant qu'agent·e moralement responsable. Dans la seconde interprétation, les obligations épistémiques relèvent plutôt d'un impératif de coopération qui est une condition de possibilité de systèmes de connaissance efficaces. Autrement dit, pour maintenir notre mode de vie, il faut connaître, et pour connaître, il faut maintenir les systèmes qui soutiennent la connaissance au niveau des communautés.

ne s'entendent-illes pas sur l'analyse que l'on doit faire du concept, mais illes la font souvent reposer sur des critères différents.

Le premier de ces critères est la conscience : on dira, par exemple, qu'une entité ne pourra pas être susceptible d'être responsable d'une action si cette entité n'est pas consciente[3]. Pour qu'une entité puisse être consciente, elle doit avoir la capacité de *ressentir* des choses comme la douleur ou la joie, par opposition à simplement disposer d'information sur ce qui cause ces sensations.

Or il y a deux difficultés majeures à intégrer un critère comme le ressenti pour déterminer la capacité d'une entité à être responsable. D'une part, la conscience comme ressenti pose ce qu'on appelle le problème « dur » de la conscience (Chalmers, 1995) : la conscience, étant irréductiblement subjective, personnelle, et impossible à partager, ne peut être adéquatement expliquée en employant le vocabulaire public que l'on emploie pour décrire des choses que l'on peut partager comme, par exemple, le monde physique autour de nous et son fonctionnement. Ainsi, on peut la pointer, mais il n'est pas du tout clair qu'on puisse jamais en rendre compte. D'autre part, et en conséquence, on fait face au problème des autres esprits (Hyslop, 2016) : étant donné qu'on ne peut rendre compte de la conscience, il est très difficile de la reconnaître chez autrui avec certitude, en particulier si cet autrui est du type d'entité qui est *a priori* très différente des humains, comme un animal ou un agent artificiel. Il est donc difficile de justifier l'application de ce critère en se basant sur quelque chose de plus solide que l'intuition.

Vient ensuite le critère de causalité: une personne est susceptible d'être responsable d'une action si elle a causé cette action. Ou, autrement dit : on ne peut dire

3 On pourrait ici se demander pourquoi je me concentre sur la conscience phénoménale (c'est-à-dire celle qui correspond au *ressenti* de l'agent·e) alors que la littérature fait souvent état du critère de conscience d'accès (c'est-à-dire la conscience que l'agent·e a des états et processus épistémiques internes qui déterminent son comportement) comme critère de responsabilité (par exemple, Holroyd 2012, Machery et collab. 2010). Alors que la capacité de conscience phénoménale correspond à une capacité individuelle, qui peut, si le critère est juste, contribuer à faire de l'entité individuelle en question une entité capable de responsabilité, la conscience d'accès est plutôt employée pour déterminer si un acte ou un événement peut être la responsabilité d'une entité dont on a déjà déterminé qu'elle était capable de responsabilité. En ce sens, la conscience phénoménale correspond davantage au questionnement que l'on a concernant le statut des agents artificiels. Ceci dit, il serait possible de produire à partir de la conscience d'accès un critère pour déterminer si une entité est capable de responsabilité – mais ce travail excède les ambitions de cet article. Je souhaite remercier un·e réviseur·se anonyme pour cette suggestion, et Sarah Arnaud pour ses contributions à cette section.

qu'une personne est responsable de quelque chose si elle n'a pas pu y contribuer causalement. La principale difficulté avec ce critère provient de la complexité de l'histoire causale des événements. Si je demande à un logiciel comme Siri ou Watson de me donner la date de la bataille des plaines d'Abraham et qu'il me répond « 1759 », la cause de cette réponse est peut-être que Siri a été la chercher sur Wikipédia. Mais on pourrait tout aussi bien dire que le travail des concepteurs et conceptrices, la machine qui fait tourner le logiciel ou même la batterie qui alimente mon cellulaire sont des causes de cette réponse. Chaque événement dispose d'une infinité de causes secondaires qui apparaissent en arrière-plan de l'histoire causale que l'on raconte. Celles-ci, à leur tour, ont leurs propres causes, et tous ces facteurs peuvent être liés causalement.

L'histoire causale est donc toujours racontée sur un fond de conditions d'arrière-plan. Dans ce contexte, déterminer si la contribution d'un facteur mérite ou non de figurer parmi les causes plutôt que parmi les conditions d'arrière-plan est important pour déterminer sa responsabilité dans l'action: si Siri me répondait « 1755 », on ne dirait pas que la source d'électricité est responsable de son erreur. Cependant, il est souvent difficile de déterminer ce qui appartient aux conditions d'arrière-plan, et ce qui appartient aux causes pertinentes. Ce problème est particulièrement ressenti dans le contexte des nouvelles technologies où les infrastructures complexes font en sorte qu'il devient difficile retracer les chaînes causales.

Un autre critère souvent employé est celui du contrôle : un·e agent·e ne pourrait être responsable d'un événement que dans la mesure où ille a du contrôle sur l'événement. Ce critère est plus fort que la causalité, puisqu'on peut souvent causer une situation sans en avoir le contrôle. Ainsi, la contribution de la personne qui fait l'entretien des machines sur lesquelles roule Siri ne serait pas responsable des actions de Siri, puisqu'elle a peu d'influence sur son comportement, et qu'elle est contrainte à faire des tâches précises, préalablement décidées. La responsabilité écherrait plutôt sur les gens qui ont un pouvoir décisionnel important.

Cependant, la séparation du travail dans les organisations implique également une séparation des pouvoirs. Non seulement la responsabilité d'une personne participant au développement de Siri s'en trouve-t-elle diluée, mais on se retrouve avec le « problème des mains multiples » : des défaillances mineures et relativement normales à plusieurs niveaux de la conception peuvent s'accumuler et interagir entre elles jusqu'à mener à des défaillances systémiques majeures,

pour lesquelles personne n'est tout à fait responsable. De plus, étant donné la tendance de centralisation des décisions, des normes et des logiciels en viennent de plus en plus à prendre les décisions à la place des humains qui sont sur le terrain (Bovens et Zouridis, 2002). Dans ce contexte, il est de plus en plus difficile d'identifier des individus qui ont un contrôle sur la situation.

Un dernier critère tourne autour de la notion de « raison », au sens des raisons que l'on donne pour justifier ou motiver une action ou un jugement. Par exemple, on dira que quelqu'un peut être responsable de son action s'ille peut évoquer les raisons qui l'ont amené à la poser. Ou alors, on dira que quelqu'un est responsable d'un comportement s'ille peut le modifier en réponse à des raisons.

Lorsqu'une nouvelle technologie a un impact à travers son rôle de médiation, ce critère nous renvoie aux raisons qui ont motivé des choix de conception. Même si les effets de la technologie sont souvent difficiles à prédire, les gens qui les développent ont une responsabilité de ne pas introduire des produits qui, de par leur conception, promeuvent la prolifération d'informations fausses ou portant à confusion.

Cependant, lorsqu'une intelligence artificielle pose un acte épistémique, elle le fait souvent sur la base de raisons qui lui appartiennent en propre. C'est particulièrement vrai dans le cas de systèmes reposant sur des algorithmes d'apprentissage machine. La connaissance sur laquelle le système se base pour prendre des décisions ne vient alors typiquement pas de la conceptrice ou du concepteur, mais a été apprise par le système lui-même. Par ailleurs, les humains qui l'ont conçu n'ont le plus souvent même pas accès aux raisons qui déterminent le comportement de leur création. Il semble donc que seul l'agent artificiel peut être crédité de ces raisons. C'est pourquoi, par exemple, lorsqu'un robot en vient à dire des choses racistes sur Twitter, on ne tiendra pas les gens qui l'ont conçu pour des personnes racistes.

Agents artificiels responsables

On constate donc que l'application des critères de responsabilité nous mène à des difficultés lorsqu'on tente de les employer dans le contexte d'un écosystème épistémique où les nouvelles technologies jouent un rôle important. Le problème ne vient cependant pas des critères eux-mêmes, mais plutôt du fait que le concept de responsabilité est calqué sur un prototype de l'action individuelle. Selon celui-ci, la responsabilité se réduit à une forme de relation entre un événe-

ment et une agente ou un agent, lesquels sont tous deux discrets et facilement identifiables.

Cette conception de l'action est problématique pour deux raisons.

Premièrement, elle suppose une forme d'internalisme de l'esprit : l'individu produisant tous ses raisonnements par lui-même, il posséderait tout ce qui lui permettrait d'être réellement responsable de ses actions. Or, on sait que l'esprit tend à « sous-traiter » son travail cognitif, notamment en faisant confiance à certaines sources d'information, en adoptant des heuristiques qui exploitent les régularités de son environnement, ou en s'appropriant des outils qui facilitent la réflexion ou autres activités cognitives (Clark, 2008). Si la personne a une certaine familiarité avec l'objet qui effectue cette sous-traitance, on peut continuer à dire qu'elle possède l'activité cognitive en question, et donc que l'emploi de cette sous-traitance ne met pas en danger sa capacité d'être responsable des décisions qu'elle prend par ce moyen. Par exemple, si je prends des notes lorsque je rencontre ma conseillère en placement et que j'utilise ces notes pour prendre des décisions financières, l'emploi de mon carnet de notes ne m'aliène pas les processus cognitifs qui constituent ma délibération en vue de ces décisions. Étant donné que j'ai le contrôle sur ce qui apparaît dans mon carnet, et que je peux évaluer la fiabilité des informations que j'y ai inscrites, je ne perds pas en responsabilité en prenant des notes. Cependant, les nouvelles technologies ne permettent souvent pas un tel degré de contrôle ; même lorsqu'on les connaît bien, les résultats sont souvent imprévisibles.

Deuxièmement, la conception de l'action individuelle que l'on retrouve, dans le concept courant de responsabilité, se figure l'agent·e comme étant une personne individuelle, et emploie des modes de raisonnement et de représentation qui supposent qu'il n'y a qu'une personne qui agit pour un événement. Il s'agit souvent d'une fiction utile, dans la mesure où, là où la situation tend à s'éloigner du modèle, nos institutions tendent à se structurer de façon à rétablir des formes d'imputabilité correspondant au modèle de l'individu humain responsable. Cependant, le mode de fonctionnement des nouvelles technologies tend à faire en sorte que les institutions peinent à combler le fossé entre l'idéal de la responsabilité individuelle et la réalité complexe des actions et perturbations que l'on constate dans les écosystèmes épistémiques pénétrés par ces technologies. D'une part, en facilitant une influence distante, elles introduisent une intentionnalité extérieure dans l'action des individus. Comme on l'a mentionné plus tôt, lorsque quelqu'un emploie un logiciel, ille suit en grande partie les intentions de la personne

qui a conçu ce logiciel. Cette personne ne peut donc pas être tenue totalement responsable si, en employant le logiciel tel que le concepteur l'a voulu, elle produit des conséquences néfastes. D'autre part, si l'on veut savoir qui est responsable des effets d'un logiciel ou d'une technologie, on est souvent confronté au fait que, la conception et la maintenance de ces systèmes nécessitant souvent de grosses équipes, il est rarement possible d'accorder la responsabilité pour une défaillance à une personne.

Une façon d'adresser ce problème pourrait être de continuer d'employer le modèle individualiste et internaliste de la responsabilité et de lui donner les moyens de s'adapter rapidement aux changements. Par exemple, on pourrait créer un nouveau concept de responsabilité conventionnelle qui accorde aux gens qui développent ou maintiennent un agent artificiel une part de responsabilité pour les conséquences de ses actions. Cependant, à refuser de vouloir développer un appareil conceptuel plus à même de représenter la responsabilité réelle, on se ferme la porte aux possibilités qu'il pourrait donner. Aussi, il y a une vertu à vouloir reconsidérer notre concept de responsabilité, et à travers lui, notre concept d'agencéité.

Une façon de ce faire consiste à donner aux agents artificiels un statut d'agent qui lui permette une certaine forme de responsabilité. Ce faisant, on peut adresser en partie le problème de l'internalisme : si on admet une forme d'agencéité épistémique aux algorithmes qui font une partie de notre travail cognitif, on peut rendre compte de ce qui dilue notre responsabilité en imputant la source de cette dilution à une entité précise.

Cependant, ce genre de projet se heurte à plusieurs difficultés. D'une part, il y a le risque de la déresponsabilisation : il y a un réel danger, par exemple, que les gens qui conçoivent des logiciels se servent de la responsabilité de leurs agents artificiels pour se défaire de la responsabilité liée à leur utilisation. D'autre part, les agents artificiels n'ont pas les mêmes capacités que les personnes humaines. C'est pourquoi même les gens qui défendent ce projet, comme Luciano Floridi ou Judith Simon, ne vont souvent pas jusqu'à leur accorder la capacité d'avoir de la responsabilité au même sens que pour les humains, et se suffisent soit d'une imputabilité (Floridi et Sanders, 2004), soit d'une forme de responsabilité correspondant aux capacités de la machine (Simon, 2015).

Un autre projet de réforme consiste à donner le statut d'agent épistémique capable de responsabilité réelle à des groupes d'humains. Selon les défenseurs de

ce projet, la responsabilité d'un groupe ne serait pas réductible à la responsabilité de ses membres. Cela permettrait, par exemple, de rendre compte de la responsabilité dans le problème des mains multiples, où les comportements individuels ne peuvent pas rendre compte de la faute collective.

Encore une fois, attribuer des responsabilités aux groupes vient avec les risques potentiels liés à la déresponsabilisation des individus. Et encore une fois, les groupes ont souvent des capacités différentes des personnes individuelles (Theiner, 2013), ce qui implique que leur responsabilité doit s'exprimer et se gérer de façon différente.

Ces difficultés ne sont pas dévastatrices pour ces deux projets ; cependant, elles soulignent la difficulté de la tâche de réformer le concept de responsabilité. Le danger de dilution de responsabilité, par exemple, soulève que l'agencéité des agents artificiels ou des groupes doit être encadrée, et que cet encadrement ne peut échoir sur une entité qui ne saurait répondre adéquatement à une critique ou à une punition. Cependant, donner un statut analogue à la responsabilité à des agents artificiels ou à des groupes ouvre également des opportunités d'encadrement. Par exemple, on note que la responsabilité épistémique des individus humains est encadrée par des institutions (écoles, médias, lois, etc.) et des relations interpersonnelles qui promeuvent des comportements vertueux et imposent des sanctions. Or, si les agents artificiels sont des agents distincts, mais dépendent des équipes d'entretien et de conception, on pourrait croire que celles-ci devraient être responsables d'encadrer les agents qui dépendent d'elles. Ainsi, la responsabilité des développeur·ses humain·es qui sont en charge d'agents artificiels irait au-delà d'une simple responsabilité pour les dommages : elle les obligerait à un encadrement spécifique qui viserait une certaine vertu et un certain régime de sanctions.

En somme, le pari du projet de réforme du concept de responsabilité épistémique est qu'à mieux rendre compte de la réalité qu'il est censé représenter, on obtiendra un concept qui est plus à même de remplir les fonctions du concept dans l'écosystème épistémique. À mieux représenter les nuances de l'action, on sera mieux à même de coordonner les agents épistémiques pour construire des écosystèmes épistémiques robustes. On peut donc espérer que l'émergence de nouvelles façons de concevoir la responsabilité, qui se limite encore à des cas-limite « monstrueux » (pour employer les mots de Lakatos, 1976), permettra d'ouvrir la voie à une entreprise de réforme plus générale de ce concept et de ceux qui peuvent l'aider à remplir ses fonctions sociales, institutionnelles et épistémiques.

Bibliographie

BOVENS, M. et S. ZOURIDIS (2002). « From Street-Level to System-Level Bureaucracies: How Information and Communication Technology is Transforming Administrative Discretion and Constitutional Control », *Public Administration Review*, vol. 62, n° 2, p. 174-184.

CLARK, A. (2008). *Supersizing the mind: Embodiment, action, and cognitive extension*, Oxford University Press.

CODE, L. (2006). *Ecological thinking: The politics of epistemic location*, Oxford University Press.

Donoghue v Stevenson [1932] UKHL 100, (26 mai 1932).

FLORIDI, L. et J. W. SANDERS (2004). « On the morality of artificial agents », *Minds and machines*, vol. 14, n° 3, p. 349-379.

GOLDENBERG, A. (2011). *La négociation des contributions dans les wikis publics: légitimation et politisation de la cognition collective*, thèse de doctorat, Université du Québec à Montréal.

HILL, A., Benjamin Mako AND Shaw (juin 2013). « The Wikipedia Gender Gap Revisited: Characterizing Survey Response Bias with Propensity Score Estimation », *PLOS ONE*, vol. 8, n° 6, p. 1-5.

HOLROYD, J. (2012). « Responsibility for implicit bias », *Journal of Social Philosophy*, vol. 43, n° 3, p. 274-306.

HUMPHREYS, P. (2004). *Extending ourselves: Computational science, empiricism, and scientific method*, Oxford University Press.

HYSLOP, A. (2016) « Other Minds », dans E. N. Zalta (dir.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University.

LAKATOS, I. (1976). *Proofs and Refutations: The Logic of Mathematical Discovery*, Cambridge University Press.

MACHERY, E., L. FAUCHER et D. R. KELLY (2010). « On the alleged inadequacies of psychological explanations of racism » *The Monist*, vol. 93, n° 2, p. 228-254.

MILLIKAN, R. G. (1984). *Language, Thought and Other Biological Categories*, MIT Press.

NOORMAN, M. (2016). « Computing and Moral Responsibility », dans E. N. Zalta (dir.), *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University.

REAGLE, J. et L. RHUE (2011). « Gender bias in Wikipedia and Britannica », *International Journal of Communication*, vol. 5, p. 21.

SIMON, J. (2015). « Distributed epistemic responsibility in a hyperconnected era », dans (dir.), *The Onlife Manifesto*, Springer, p. 145-159.

Smith, A. (2005). « Responsibility for Attitudes: Activity and Passivity in Mental Life », *Ethics*, vol. 115, n° 2, p.236-271.

THEINER, G. (2013). « Onwards and upwards with the extended mind: From individual to collective epistemic action », *Developing scaffolds*, p. 191-208.

TSVETKOVA, M., R. GARCÍA-GAVILANES, L. FLORIDI, et T. YASSERI (février 2017). « Even good bots fight: The case of Wikipedia », *PLOS ONE*, vol. 12, n° 2, p. 1-13.