

1 Le comportement des machines

1.1 Résumé

Les machines alimentées par l'intelligence artificielle (ia) médiatisent de plus en plus nos interactions sociales, culturelles, économiques et politiques. Comprendre le comportement des systèmes d'intelligence artificielle est essentiel à notre capacité à contrôler leurs actions, à maximiser leurs bénéfices et à minimiser leurs nuisances. Nous soutenons ici que cela nécessite un vaste programme de recherche scientifique pour étudier le comportement des machines, qui intègre et élargit la discipline de l'informatique et inclut des perspectives issues de l'ensemble des sciences. Nous commençons par exposer un ensemble de questions fondamentales pour ce domaine émergent, puis nous explorons les contraintes techniques, juridiques et institutionnelles qui pèsent sur l'étude du comportement des machines.

Dans son ouvrage phare de 1969, *The Sciences of the Artificial*, le lauréat du prix Nobel Herbert Simon écrivait : “La science naturelle est la connaissance des objets et phénomènes naturels. Nous nous demandons s'il ne peut pas aussi exister une science de l'“artificiel” - une connaissance des objets et phénomènes artificiels.” [1] Dans l'esprit de la vision de Simon, nous décrivons l'émergence d'un champ d'étude scientifique interdisciplinaire. Ce champ porte sur l'étude scientifique des machines intelligentes, non pas en tant qu'artefacts d'ingénierie, mais en tant que classe d'acteurs dotés de schémas comportementaux et d'une écologie particuliers. Ce champ recoupe, tout en s'en distinguant, l'informatique et la robotique. Il traite le comportement des machines de manière empirique. Cela s'apparente à la manière dont l'éthologie et l'écologie comportementale étudient le comportement animal en intégrant la physiologie et la biochimie - propriétés intrinsèques - à l'étude de l'écologie et de l'évolution - propriétés façonnées par l'environnement. Les comportements animaux et humains ne peuvent être pleinement compris sans l'étude des contextes dans lesquels ils se produisent. Le comportement des machines ne peut, de même, être pleinement compris sans l'étude intégrée des algorithmes et des environnements sociaux dans lesquels ces algorithmes opèrent [2]. À l'heure actuelle, les scientifiques qui étudient les comportements de ces agents d'intelligence artificielle (ia) virtuels et incarnés sont majoritairement les mêmes scientifiques qui ont créé ces agents eux-mêmes (dans tout cet article, nous utilisons le terme “agents d'ia” de manière large pour désigner aussi bien les algorithmes complexes que simples utilisés pour prendre des décisions). Comme ces scientifiques créent des agents pour résoudre des tâches particulières, ils s'attachent souvent à garantir que les agents remplissent leur fonction prévue (bien que ces domaines respectifs soient bien plus vastes que les exemples spécifiques cités ici). Par exemple, les agents d'ia doivent atteindre un seuil de précision dans la classification de documents, la reconnaissance faciale ou la détection visuelle d'objets. Les voitures autonomes doivent naviguer avec succès dans diverses conditions météorologiques ; les agents joueurs de jeux doivent vaincre divers adversaires humains ou machines ; et les agents d'exploration de données doivent apprendre quels individus cibler dans les campagnes publicitaires sur les réseaux sociaux.

Denise Vella-Chemla et deepseek, octobre 2026.

1. Référence : <https://www.nature.com/articles/s41586-019-1138-y>.

Ces agents d'ia ont le potentiel d'améliorer le bien-être humain de nombreuses manières. C'est d'ailleurs généralement la vision de leurs créateurs. Mais une considération plus large du comportement des agents d'ia est désormais critique. Les agents d'ia s'intégreront de plus en plus à notre société et sont déjà impliqués dans diverses activités, telles que le scoring de crédit, le trading algorithmique, le maintien de l'ordre local, les décisions de libération conditionnelle, la conduite, les rencontres en ligne et la guerre par drones [5, 6, 7, 8, 9]. Des commentateurs et des chercheurs de divers domaines - incluant, sans s'y limiter, l'ingénierie des systèmes cognitifs, l'interaction homme-machine, les facteurs humains, les sciences, technologies et société, et l'ingénierie de la sécurité - tirent la sonnette d'alarme quant aux conséquences larges et non intentionnelles des agents d'ia, qui peuvent exhiber des comportements et produire des effets sociétaux en aval - positifs comme négatifs - imprévus par leurs créateurs [5, 6, 7, 8, 9].

Outre ce manque de prévisibilité entourant les conséquences de l'ia, il existe une crainte de la perte potentielle de supervision humaine sur les machines intelligentes et des préjudices potentiels associés à l'utilisation croissante de machines pour des tâches autrefois accomplies directement par des humains [9]. Dans le même temps, des chercheurs décrivent les bénéfices que les agents d'ia peuvent offrir à la société en soutenant et augmentant la prise de décision humaine [10, 11]. Bien que les discussions sur ces questions aient conduit à de nombreuses avancées importantes dans de nombreux champs de recherche académique distincts [12], certains mettant en lumière les défis de sécurité des systèmes autonomes [13] et d'autres étudiant les implications en matière d'équité, de responsabilité et de transparence (par exemple, la conférence ACM sur l'équité, la responsabilité et la transparence), de nombreuses questions demeurent.

Cette revue cadre et passe en revue le champ interdisciplinaire émergent du comportement des machines : l'étude scientifique du comportement exhibé par les machines intelligentes. Nous exposons ici les thèmes de recherche clés, les questions et les études de recherche marquantes qui illustrent ce champ. Nous commençons par fournir un contexte sur l'étude du comportement des machines et la nature nécessairement interdisciplinaire de cette science. Nous proposons ensuite un cadre pour la conceptualisation des études du comportement des machines. Nous concluons par un appel à l'étude scientifique des écologies machine et homme-machine et discutons de certains des obstacles techniques, juridiques et institutionnels auxquels sont confrontés les chercheurs dans ce domaine.

1.2 Motivation pour l'étude du comportement des machines

Trois motivations principales sous-tendent la discipline scientifique du comportement des machines. Premièrement, divers types d'algorithmes opèrent dans notre société, et les algorithmes jouent un rôle toujours croissant dans nos activités quotidiennes. Deuxièmement, en raison des propriétés complexes de ces algorithmes et des environnements dans lesquels ils opèrent, certains de leurs attributs et comportements peuvent être difficiles ou impossibles à formaliser analytiquement. Troisièmement, en raison de leur ubiquité et de leur complexité, prédire les effets des algorithmes intelligents sur l'humanité - qu'ils soient positifs ou négatifs - représente un défi considérable.

1.2.1 Ubiquité des algorithmes

La prévalence actuelle de divers algorithmes dans la société est sans précédent (Fig. 1). Les algorithmes de classement de l'information et les bots des réseaux sociaux influencent les informations vues par les citoyens [14, 15, 16, 17, 18]. Les algorithmes de scoring de crédit déterminent les décisions de prêt [19, 20, 21, 22]. Les algorithmes de tarification en ligne façonnent le coût des produits

de manière différenciée selon les consommateurs [23, 24, 25]. Les logiciels de trading algorithmique effectuent des transactions sur les marchés financiers à grande vitesse [26, 27, 28, 29]. Les algorithmes façonnent la répartition et les schémas spatiaux du maintien de l'ordre local [30] et les programmes de détermination algorithmique des peines affectent le temps purgé dans le système pénal. Les voitures autonomes parcourent nos villes [31], et les algorithmes de covoiturage modifient les schémas de déplacement des véhicules conventionnels [32]. Les machines cartographient nos foyers, répondent aux commandes verbales [33] et accomplissent des tâches ménagères régulières [34]. Les algorithmes façonnent les correspondances romantiques pour les services de rencontres en ligne [35, 36]. Les machines sont susceptibles de remplacer de plus en plus les humains dans l'éducation de nos jeunes [37] et les soins à nos anciens [38]. Les agents autonomes sont de plus en plus susceptibles d'affecter les comportements collectifs, de la coordination à l'échelle du groupe au partage [39]. En outre, bien que la perspective de développer des armes autonomes soit hautement controversée, beaucoup dans le domaine exprimant leur opposition [40], si de telles armes finissaient par être déployées, alors les machines pourraient déterminer qui vit et qui meurt dans les conflits armés [41, 42].

1.2.2 Complexité et opacité des algorithmes

L'extrême diversité de ces systèmes d'ia, couplée à leur ubiquité, suffirait à elle seule à garantir que l'étude du comportement de ces systèmes pose un défi formidable, même si les algorithmes individuels étaient relativement simples. La complexité des agents d'ia individuels est actuellement élevée et croît rapidement. Bien que le code spécifiant l'architecture et l'entraînement d'un modèle puisse être simple, les résultats peuvent être très complexes, aboutissant souvent à des boîtes noires [43]. On leur donne une entrée et ils produisent une sortie, mais les processus fonctionnels exacts qui génèrent ces sorties sont difficiles à interpréter, même pour les scientifiques qui génèrent les algorithmes eux-mêmes [44], bien que des progrès en interprétabilité soient réalisés [45, 46]. En outre, lorsque les systèmes apprennent à partir de données, leurs défaillances sont liées aux imperfections des données ou à la manière dont les données ont été collectées, ce qui a conduit certains à plaider pour des mécanismes de déclaration adaptés pour les jeux de données [47] et les modèles [48]. La dimensionnalité et la taille des données ajoutent une couche supplémentaire de complexité à la compréhension du comportement des machines [49].

Ce qui complique encore ce défi, c'est le fait qu'une grande partie du code source et de la structure des modèles des algorithmes les plus fréquemment utilisés dans la société est propriétaire, tout comme les données sur lesquelles ces systèmes sont entraînés. Le secret industriel et la protection juridique de la propriété intellectuelle entourent souvent le code source et la structure des modèles. Dans de nombreux contextes, les seuls facteurs publiquement observables concernant les systèmes d'ia industriels sont leurs entrées et leurs sorties.

Même lorsqu'ils sont disponibles, le code source ou la structure du modèle d'un agent d'ia peuvent fournir une puissance prédictive insuffisante sur sa sortie. Les agents d'ia peuvent également démontrer des comportements nouveaux par leur interaction avec le monde et d'autres agents, impossibles à prédire avec précision [50]. Même lorsque les solutions analytiques sont mathématiquement descriptibles, elles peuvent être si longues et complexes qu'elles en deviennent indéchiffrables [51, 52]. En outre, lorsque l'environnement change - peut-être en raison de l'algorithme lui-même - anticiper et analyser le comportement devient bien plus difficile.

1.2.3 Effets bénéfiques et néfastes des algorithmes sur l’humanité

L’ubiquité des algorithmes, couplée à leur complexité croissante, tend à amplifier la difficulté d’estimer les effets des algorithmes sur les individus et la société. Les agents d’ia peuvent façonner les comportements humains et les résultats sociétaux de manières intentionnelles et non intentionnelles. Par exemple, certains agents d’ia sont conçus pour aider les enfants dans leurs apprentissages [53] et d’autres pour assister les personnes âgées [54]. Ces systèmes d’ia peuvent bénéficier à leurs humains cibles en incitant ces derniers à de meilleurs comportements d’apprentissage ou à une mobilité plus sûre. Cependant, avec le pouvoir d’inciter les comportements humains de manières positives ou intentionnelles vient le risque que les comportements humains soient incités de manières coûteuses ou non intentionnelles - les enfants pourraient être influencés pour acheter certains produits de marque et les anciens pourraient être incités à regarder certaines émissions de télévision.

La manière dont ces influences algorithmiques sur les individus se transforment en effets à l’échelle de la société - positifs comme négatifs - est une préoccupation critique. Par exemple, l’exposition d’un petit nombre d’individus à de la désinformation politique peut avoir peu d’effet sur la société dans son ensemble. Cependant, l’effet de l’insertion et de la propagation de cette désinformation sur les réseaux sociaux peut avoir des conséquences sociétales plus substantielles [55, 56, 57]. En outre, les questions d’équité ou de biais algorithmiques [58, 59] ont déjà été documentées dans divers contextes, notamment la vision par ordinateur [60], les plongements de mots [61, 62], la publicité [63], le maintien de l’ordre [64], la justice pénale [65] et les services sociaux [66]. Pour traiter ces questions, les praticiens seront parfois contraints de faire des compromis de valeurs entre des notions concurrentes et incompatibles de biais [58, 59] ou entre les biais humains et ceux des machines. Des questions supplémentaires concernant l’effet des algorithmes demeurent, comme la manière dont les algorithmes de rencontres en ligne altèrent l’institution sociétale du mariage [35, 36] et s’il existe des effets systémiques de l’interaction croissante avec des algorithmes intelligents sur les étapes et la vitesse du développement humain [53]. Ces questions deviennent plus complexes dans les systèmes hybrides composés de nombreuses machines et humains interagissant et manifestant un comportement collectif [39, 67]. Pour que la société puisse avoir un droit de regard et une supervision sur les conséquences en aval de l’ia, les chercheurs en comportement des machines doivent fournir des éclairages sur le fonctionnement de ces systèmes et sur les bénéfices, coûts et compromis présentés par l’utilisation ubiquitaire de l’ia dans la société.

1.3 L’étude interdisciplinaire du comportement des machines

Pour étudier le comportement des machines - en particulier les comportements des algorithmes boîtes noires dans des contextes réels -, nous devons intégrer des connaissances issues d’une variété de disciplines scientifiques (Fig. 2). Cette intégration en est actuellement à ses balbutiements et s’est produite en grande partie de manière ad hoc en réponse au besoin croissant de comprendre le comportement des machines. Actuellement, les scientifiques qui étudient le plus couramment le comportement des machines sont les informaticiens, roboticiennes et ingénieurs qui ont créé les machines en premier lieu. Ces scientifiques peuvent être des mathématiciens et ingénieurs experts ; cependant, ils ne sont généralement pas formés comme comportementalistes. Ils reçoivent rarement une instruction formelle sur la méthodologie expérimentale, les statistiques basées sur la population et les paradigmes d’échantillonnage, ou l’inférence causale observationnelle, sans parler des neurosciences, du comportement collectif ou de la théorie sociale. Inversement, bien que les scientifiques du comportement soient plus susceptibles de posséder une formation dans ces méthodes scientifiques,

ils sont moins susceptibles de posséder l’expertise requise pour évaluer avec compétence la qualité sous-jacente et la pertinence des techniques d’ia pour un domaine problématique donné, ou pour décrire mathématiquement les propriétés d’algorithmes particuliers.

Intégrer les pratiques scientifiques de multiples domaines n’est pas chose facile. Jusqu’à présent, l’objectif principal de ceux qui créent des systèmes d’ia a été de concevoir, mettre en œuvre et optimiser des systèmes intelligents pour accomplir des tâches spécialisées. Des progrès excellents ont été réalisés sur des tâches de référence, notamment les jeux de société comme les échecs, les dames et le Go [68, 69, 70, 71], les jeux de cartes comme le poker [72], les jeux informatiques comme ceux de la plateforme Atari [73], les marchés artificiels [74] et la Robocup Soccer [75], ainsi que sur des données d’évaluation standardisées, telles que les données ImageNet pour la reconnaissance d’objets [76] et les données Microsoft Common Objects in Context pour les tâches de légendage d’images [77]. Des succès ont également été obtenus en reconnaissance vocale, traduction automatique [80] et locomotion autonome. Ces références sont couplées à des métriques pour quantifier la performance sur des tâches standardisées - et sont utilisées pour améliorer la performance, un proxy qui permet aux constructeurs d’ia de viser des algorithmes meilleurs, plus rapides et plus robustes.

Mais les méthodologies visant à maximiser la performance algorithmique ne sont pas optimales pour mener une observation scientifique des propriétés et comportements des agents d’ia. Plutôt que d’utiliser des métriques au service de l’optimisation contre des références, les chercheurs en comportement des machines s’intéressent à un ensemble plus large d’indicateurs, tout comme les scientifiques sociaux explorent un large éventail de comportements humains dans le domaine des interactions sociales, politiques ou économiques. À ce titre, les chercheurs en comportement des machines consacrent des efforts considérables à définir des mesures de résultats micro et macro pour répondre à des questions larges telles que : comment ces algorithmes se comportent-ils dans différents environnements et les interactions humaines avec les algorithmes altèrent-elles les résultats sociétaux ? Les expériences randomisées, l’inférence observationnelle et les méthodes de statistiques descriptives basées sur la population - souvent utilisées dans les sciences comportementales quantitatives - doivent être centrales dans l’étude du comportement des machines. L’incorporation de chercheurs extérieurs aux disciplines qui produisent traditionnellement les machines intelligentes peut apporter une connaissance d’outils méthodologiques importants, d’approches scientifiques, de cadres conceptuels alternatifs et de perspectives sur les phénomènes économiques, sociaux et politiques que les machines influenceront de plus en plus.

1.4 Type de question et objet d’étude

Nikolaas Tinbergen, qui a remporté le prix Nobel de physiologie ou médecine en 1973 aux côtés de Karl von Frisch et Konrad Lorenz pour avoir fondé le domaine de l’éthologie, a identifié quatre dimensions complémentaires d’analyse qui aident à expliquer le comportement animal [83]. Ces dimensions concernent les questions de fonction, de mécanisme, de développement et d’histoire évolutive d’un comportement et fournissent un cadre organisateur pour l’étude du comportement animal et humain. Par exemple, cette conceptualisation distingue l’étude de la manière dont un jeune animal ou humain développe un type de comportement de la trajectoire évolutive qui a sélectionné ce comportement dans la population. Le but de ces distinctions n’est pas la division mais plutôt l’intégration. Bien qu’il ne soit pas faux de dire que, par exemple, le chant d’un oiseau s’explique par l’apprentissage ou par son histoire évolutive spécifique, une compréhension complète du chant nécessitera les deux.

Malgré les différences fondamentales entre les machines et les animaux, l'étude comportementale des machines peut bénéficier d'une classification similaire. Les machines ont des mécanismes qui produisent le comportement, subissent un développement qui intègre l'information environnementale dans le comportement, produisent des conséquences fonctionnelles qui font que des machines spécifiques deviennent plus ou moins communes dans des environnements spécifiques, et incarnent des histoires évolutives à travers lesquelles les environnements passés et les décisions humaines continuent d'influencer le comportement des machines. Les chercheurs en informatique ont déjà réalisé des gains substantiels dans la compréhension des mécanismes et du développement des systèmes d'ia, bien que de nombreuses questions demeurent. Un accent relativement moindre a été mis sur la fonction et l'évolution des systèmes d'ia. Nous discutons de ces quatre sujets dans les sous-sections suivantes et fournissons la Fig. 3 comme résumé.

1.4.1 Mécanismes de génération du comportement

Les causes proximales du comportement d'une machine ont trait à la manière dont le comportement est déclenché observationnellement et généré dans des environnements spécifiques. Par exemple, les premiers programmes de trading algorithmique utilisaient des règles simples pour déclencher des comportements d'achat et de vente. Des agents plus sophistiqués peuvent calculer des stratégies basées sur des heuristiques adaptatives ou une maximisation explicite de l'utilité espérée [85, 86]. Le comportement d'un algorithme d'apprentissage par renforcement qui joue au poker pourrait être attribué à la manière particulière dont il représente l'espace d'états ou évalue l'arbre de jeu, etc. Un mécanisme dépend à la fois d'un algorithme et de son environnement. Un agent plus sophistiqué, comme une voiture sans conducteur, peut exhiber un comportement de conduite particulier - par exemple, changer de voie, dépasser ou signaler aux piétons. Ces comportements seraient générés selon les algorithmes qui construisent les politiques de conduite et sont aussi fondamentalement façonnés par les caractéristiques du système de perception et d'actionnement de la voiture, notamment la résolution et la précision de son système de détection et de classification d'objets, et la réactivité et la précision de sa direction, entre autres facteurs [87]. Parce que de nombreux systèmes d'ia actuels sont dérivés de méthodes d'apprentissage automatique appliquées à des données de plus en plus complexes, l'étude du mécanisme derrière le comportement d'une machine, comme ceux mentionnés ci-dessus, nécessitera un travail continu sur les méthodes d'interprétabilité de l'apprentissage automatique [88, 89].

1.4.2 Développement du comportement

Dans l'étude du comportement animal ou humain, le développement fait référence à la manière dont un individu acquiert un comportement particulier - par exemple, par imitation ou conditionnement environnemental. Cela se distingue des changements évolutifs à plus long terme.

Dans le contexte des machines, nous pouvons nous demander comment les machines acquièrent (développent) un comportement individuel ou collectif spécifique. Le développement comportemental pourrait être directement attribuable à des choix d'ingénierie ou de conception humains. Les choix de conception architecturale faits par le programmeur (par exemple, la valeur d'un paramètre de taux d'apprentissage, l'acquisition de la représentation des connaissances et de l'état, ou un câblage particulier d'un réseau de neurones convolutionnel) déterminent ou influencent les types de comportements que l'algorithme exhibe. Dans un système d'ia plus complexe, comme une voiture sans conducteur, le comportement de la voiture se développe au fil du temps, à partir du développement

logiciel et des composants matériels changeants que les ingénieurs intègrent dans son architecture globale. Les comportements peuvent également changer à la suite de mises à niveau algorithmiques poussées vers la machine par ses concepteurs après le déploiement.

Un ingénieur humain peut également façonner le comportement de la machine en l'exposant à des stimuli d'entraînement particuliers. Par exemple, de nombreux algorithmes de classification d'images et de textes sont entraînés à optimiser la précision sur un ensemble spécifique de jeux de données étiquetés manuellement par des humains. Le choix du jeu de données - et les caractéristiques qu'il représente - peut influencer substantiellement le comportement exhibé par l'algorithme.

Enfin, une machine peut acquérir des comportements par sa propre expérience. Par exemple, un agent d'apprentissage par renforcement entraîné à maximiser le profit à long terme peut apprendre des stratégies de trading à court terme particulières basées sur ses propres actions passées et les retours concomitants du marché [90]. De même, les algorithmes de recommandation de produits font des recommandations basées sur un flux incessant de choix effectués par les clients et mettent à jour leurs recommandations en conséquence.

1.4.3 Fonction

Dans l'étude du comportement animal, la valeur adaptative décrit comment un comportement contribue à la fitness reproductive au cours de la vie d'un animal. Par exemple, un comportement de chasse particulier peut être plus ou moins efficace qu'un autre pour prolonger la vie de l'animal et, corrélativement, le nombre d'opportunités d'accouplement, la progéniture née et le succès reproducteur probable de la progéniture. L'accent mis sur la fonction nous aide à comprendre pourquoi certains mécanismes comportementaux se propagent et persistent tandis que d'autres déclinent et disparaissent. La fonction dépend critically de l'adéquation du comportement à l'environnement.

Dans le cas des machines, nous pouvons parler de la manière dont le comportement remplit une fonction contemporaine pour des parties prenantes humaines particulières. L'environnement humain crée des forces sélectives qui peuvent rendre certaines machines plus communes. Les comportements qui réussissent (qui améliorent la fitness) sont copiés par les développeurs d'autres logiciels et matériels ou sont parfois conçus pour se propager parmi les machines elles-mêmes. Ces dynamiques sont en fin de compte pilotées par le succès des institutions - telles que les entreprises, les hôpitaux, les gouvernements municipaux et les universités - qui construisent ou utilisent l'ia. L'exemple le plus évident est fourni par le trading algorithmique, où les stratégies de trading automatisées qui réussissent peuvent être copiées lorsque leurs développeurs passent d'une entreprise à une autre, ou sont simplement observées et rétro-ingénierées par des rivaux.

Ces forces peuvent produire des effets imprévus. Par exemple, des objectifs tels que maximiser l'engagement sur un site de réseaux sociaux peuvent conduire à ce qu'on appelle des bulles de filtres, ce qui peut accroître la polarisation politique ou, sans modération prudente, faciliter la propagation de fausses nouvelles [91]. Cependant, les sites web qui n'optimisent pas pour l'engagement des utilisateurs peuvent ne pas être aussi prospères en comparaison de ceux qui le font, ou peuvent faire faillite. De même, en l'absence de régulation externe, les voitures autonomes qui ne priorisent pas la sécurité de leurs propres passagers peuvent être moins attractives pour les consommateurs, entraînant moins de ventes [31]. Parfois, la fonction du comportement d'une machine est de faire face au comportement d'autres machines. Les attaques adversariales - des entrées synthétiques qui trompent un système pour qu'il produise une sortie non désirée [44, 92, 93, 94] - sur les systèmes d'ia et les réponses subséquentes de ceux qui développent l'ia à ces attaques [95] peuvent produire des dynamiques proie-prédateur complexes qui ne sont pas facilement comprises en étudiant chaque

machine isolément.

Ces exemples mettent en évidence comment les incitations créées par les institutions externes et les forces économiques peuvent avoir des effets indirects mais substantiels sur les comportements exhibés par les machines [96]. Comprendre l’interaction entre ces incitations et l’ia est pertinent pour l’étude du comportement des machines. Ces dynamiques de marché interagiraient, à leur tour, avec d’autres processus pour produire une évolution parmi les machines et les algorithmes.

1.4.4 Évolution

Dans l’étude du comportement animal, la phylogénèse décrit comment un comportement a évolué. Outre sa fonction actuelle, le comportement est influencé par les pressions sélectives passées et les mécanismes précédemment évolués. Par exemple, la main humaine a évolué à partir de la nageoire d’un poisson osseux. Sa fonction actuelle n’est plus la nage, mais sa structure interne s’explique par son histoire évolutive. Des forces non sélectives, telles que la migration et la dérive, jouent également un rôle important pour expliquer les relations entre différentes formes de comportement. Dans le cas des machines, l’histoire évolutive peut également générer une dépendance au chemin, expliquant des comportements autrement déroutants. À chaque étape, des aspects des algorithmes sont réutilisés dans de nouveaux contextes, contraignant à la fois le comportement futur et rendant possibles des innovations supplémentaires. Par exemple, les premiers choix concernant la conception des microprocesseurs continuent d’influencer l’informatique moderne, et les traditions dans la conception d’algorithmes - telles que les réseaux de neurones et les modèles bayésiens à espace d’états - intègrent de nombreuses hypothèses et guident les innovations futures en rendant certains nouveaux algorithmes plus faciles d’accès que d’autres. En conséquence, certains algorithmes peuvent prêter attention à certaines caractéristiques et en ignorer d’autres parce que ces caractéristiques étaient importantes dans les premières applications réussies. Certains comportements de machine peuvent se répandre parce qu’ils sont “évolables” - faciles à modifier et robustes aux perturbations - semblables à la façon dont certains traits des animaux peuvent être communs parce qu’ils facilitent la diversité et la stabilité [97].

Le comportement des machines évolue différemment du comportement animal. La plupart de l’hérédité animale est simple - deux parents, un événement de transmission. Les algorithmes sont beaucoup plus flexibles et ils ont un concepteur avec un objectif en arrière-plan. L’environnement humain influence fortement la manière dont les algorithmes évoluent en changeant leur système d’hérédité. Le comportement de réplication de l’ia peut être facilité par une culture de partage open source de logiciels, les détails de l’architecture réseau ou les jeux de données d’entraînement sous-jacents. Par exemple, les entreprises qui développent des logiciels pour voitures sans conducteur peuvent partager des bibliothèques open source améliorées pour la détection d’objets ou la planification de trajectoire, ainsi que les données d’entraînement qui sous-tendent ces algorithmes, afin de permettre aux logiciels améliorant la sécurité de se répandre dans toute l’industrie. Il est possible qu’une seule “mutation” adaptative dans le comportement d’une voiture sans conducteur particulière se propage instantanément à des millions d’autres voitures via une mise à jour logicielle. Cependant, d’autres institutions appliquent également des limites. Par exemple, les brevets logiciels peuvent imposer des contraintes sur la copie de traits comportementaux particuliers. Et les contraintes réglementaires - telles que les lois sur la protection de la vie privée - peuvent empêcher les machines d’accéder, de conserver ou d’utiliser autrement certaines informations dans leur prise de décision. Ces particularités soulignent le fait que les machines peuvent exhiber des trajectoires évolutives très différentes, car elles ne sont pas liées aux mécanismes de l’évolution organique.

1.5 Échelle d’investigation

Avec le cadre exposé ci-dessus et dans la Fig. 3, nous cataloguons maintenant des exemples de comportement de machine aux trois échelles d’investigation : machines individuelles, collectifs de machines et groupes de machines intégrés dans un environnement social avec des groupes d’humains dans des systèmes hybrides ou hétérogènes [39] (Fig. 4). Le comportement de machine individuelle met l’accent sur l’étude de l’algorithme lui-même, le comportement de machine collective met l’accent sur l’étude des interactions entre machines, et le comportement hybride homme-machine met l’accent sur l’étude des interactions entre machines et humains. Nous pouvons ici faire une analogie avec l’étude d’une espèce particulière, l’étude des interactions entre membres d’une espèce et les interactions de l’espèce avec son environnement plus large. Les analyses à chacune de ces échelles peuvent aborder tout ou partie des questions décrites dans la Fig. 3.

1.5.1 Comportement de machine individuelle

L’étude du comportement des machines individuelles se concentre sur des machines intelligentes spécifiques par elles-mêmes. Souvent, ces études se concentrent sur des propriétés intrinsèques aux machines individuelles et qui sont pilotées par leur code source ou leur conception. Les domaines de l’apprentissage automatique et du génie logiciel mènent actuellement la majorité de ces études. Il existe deux approches générales pour l’étude du comportement de machine individuelle. La première se concentre sur le profilage de l’ensemble des comportements d’un agent machine spécifique en utilisant une approche intra-machine, comparant le comportement d’une machine particulière dans différentes conditions. La seconde, une approche inter-machines, examine comment une variété d’agents machines individuels se comportent dans la même condition.

Une approche intra-machine pour l’étude des comportements de machines individuelles investigate des questions telles que : existe-t-il des constantes qui caractérisent le comportement intra-machine d’une ia particulière à travers une variété de contextes, comment le comportement d’une ia particulière progresse-t-il au fil du temps dans le même environnement ou des environnements différents, et quels facteurs environnementaux conduisent à l’expression de comportements particuliers par les machines.

Par exemple, un algorithme peut n’exhiber certains comportements que s’il est entraîné sur des données sous-jacentes particulières [98, 99, 100] (Fig. 3). Ensuite, la question devient de savoir si un algorithme qui évalue la probabilité de récidive dans les décisions de libération conditionnelle se comporterait de manière inattendue lorsqu’on lui présente des données d’évaluation qui divergent substantiellement de ses données d’entraînement. D’autres études liées à la caractérisation du comportement intra-machine incluent l’étude des comportements de récupération de robots individuels [101, 102], les attributs cognitifs des algorithmes et l’utilité d’utiliser des techniques de la psychologie dans l’étude du comportement algorithmique [103], et l’examen de caractéristiques spécifiques aux bots telles que celles conçues pour influencer les utilisateurs humains [104].

La seconde approche pour l’étude du comportement de machine individuelle examine les mêmes comportements tels qu’ils varient entre les machines. Par exemple, ceux qui s’intéressent à l’examen des comportements publicitaires d’agents intelligents [63, 105, 106] peuvent étudier une variété de plateformes publicitaires (et leurs algorithmes sous-jacents) et examiner l’effet inter-machines de la réalisation d’expériences avec le même ensemble d’entrées publicitaires sur toutes les plateformes. La même approche pourrait être utilisée pour les investigations des algorithmes de tarification dynamique [23, 24, 32] sur les plateformes. D’autres études inter-machines pourraient examiner les

différents comportements utilisés par les véhicules autonomes dans leurs schémas de dépassement ou les comportements de recherche de nourriture variés exhibés par les drones de recherche et sauvetage [107].

1.5.2 Comportement de machine collective

Contrairement à l'étude du comportement des machines individuelles, l'étude du comportement de machine collective se concentre sur les comportements interactifs et à l'échelle du système de collections d'agents machines. Dans certains cas, les implications du comportement de machine individuelle peuvent n'avoir que peu de sens tant que le niveau collectif n'est pas considéré. Certaines investigations de ces systèmes ont été inspirées par des collectifs naturels, tels que les essaims d'insectes, ou des groupes mobiles, tels que les oiseaux en vol ou les poissons en banc. Par exemple, les groupes animaux sont connus pour exhiber à la fois une détection émergente de caractéristiques environnementales complexes [108] et une prise de décision consensuelle efficace [109]. Dans les deux scénarios, les groupes exhibent une conscience de l'environnement qui n'existe pas au niveau individuel. Des domaines tels que les systèmes multi-agents et la théorie des jeux computationnelle fournissent des exemples utiles de l'étude de ce domaine du comportement des machines.

Les robots qui utilisent des algorithmes simples pour les interactions locales entre bots peuvent néanmoins produire un comportement intéressant une fois agrégés en grands collectifs. Par exemple, des chercheurs ont examiné les propriétés de type essaim de microrobots qui se combinent en agrégations ressemblant à des essaims trouvés dans les systèmes d'agents biologiques [110, 111]. Des exemples supplémentaires incluent les comportements collectifs d'algorithmes à la fois en laboratoire (dans le Jeu de la Vie [112]) ainsi qu'à l'état sauvage (comme on le voit dans les bots d'édition de Wikipédia [113]). D'autres exemples incluent l'émergence de nouveaux langages algorithmiques [114] entre machines intelligentes communicantes ainsi que les propriétés dynamiques des systèmes de transport entièrement autonomes. En fin de compte, de nombreuses questions intéressantes dans ce domaine restent à examiner.

La grande majorité des travaux sur le comportement animal collectif et la robotique collective s'est concentrée sur la manière dont les interactions entre agents simples peuvent créer des structures et des propriétés d'ordre supérieur. Bien que important, cela néglige le fait que de nombreux organismes, et de plus en plus aussi les agents d'ia, sont des entités sophistiquées avec des comportements et des interactions qui peuvent ne pas être bien caractérisés par des représentations simplistes. Révéler quelles propriétés supplémentaires émergent lorsque les entités en interaction sont capables de cognition sophistiquée reste un défi clé dans les sciences biologiques et peut avoir des parallèles directs dans l'étude du comportement des machines. Par exemple, semblables aux animaux, les machines peuvent exhiber un apprentissage social. Un tel apprentissage social n'a pas besoin d'être limité aux machines apprenant des machines, mais nous pouvons nous attendre à ce que les machines apprennent des humains, et vice versa, que les humains apprennent du comportement des machines. Les processus de rétroaction introduits peuvent fondamentalement altérer l'accumulation de connaissances, y compris à travers les générations, affectant directement la culture humaine et machine.

En outre, les systèmes d'ia créés par l'homme ne font pas nécessairement face aux mêmes contraintes que les organismes, et les assemblages collectifs de machines fournissent de nouvelles capacités, telles que la communication mondiale instantanée, qui peuvent conduire à des schémas comportementaux collectifs entièrement nouveaux. Les études sur le comportement de machine collective examinent les propriétés des assemblages de machines ainsi que les propriétés inattendues qui peuvent émerger

de ces systèmes complexes d'interactions.

Par exemple, certains des comportements collectifs les plus intéressants des algorithmes ont été observés dans les environnements de trading financier. Ces environnements opèrent sur des échelles de temps infimes, de sorte que les traders algorithmiques peuvent répondre aux événements et les uns aux autres avant tout trader humain [115]. Dans certaines conditions, les capacités à haute fréquence peuvent produire des inefficacités dans les marchés financiers [26, 115]. Outre la vitesse de réponse sans précédent, l'utilisation extensive de l'apprentissage automatique, l'opération autonome et la capacité de déploiement à grande échelle sont toutes des raisons de croire que le comportement collectif du trading machine peut être qualitativement différent de celui des traders humains. En outre, ces algorithmes financiers et systèmes de trading sont nécessairement entraînés sur certains jeux de données historiques et réagissent à une variété limitée de scénarios prévus, ce qui soulève la question de savoir comment ils réagiront à des situations nouvelles et imprévues dans leur conception. Les flash crashes sont des exemples de conséquences clairement non intentionnelles d'algorithmes (interagissant) [50, 115], ce qui conduit à la question de savoir si les algorithmes pourraient interagir pour créer une crise de marché plus large.

1.5.3 Comportement hybride homme-machine

Les humains interagissent de plus en plus avec les machines [16]. Elles médiatisent nos interactions sociales [39], façonnent les nouvelles [14, 17, 55, 56] et les informations en ligne [15, 116] que nous voyons, et forment avec nous des relations qui peuvent altérer nos systèmes sociaux. En raison de leur complexité, ces systèmes hybrides homme-machine posent l'un des domaines d'étude les plus difficiles techniquement mais aussi les plus importants pour le comportement des machines.

Les machines façonnent le comportement humain. L'un des domaines les plus évidents - mais néanmoins vitaux - de l'étude du comportement des machines concerne les façons dont l'introduction de machines intelligentes dans les systèmes sociaux peut altérer les comportements et les aspirations humains. Comme dans l'introduction de l'automatisation dans les processus industriels [117], les machines intelligentes peuvent créer des problèmes sociaux dans le processus d'amélioration des existants. De nombreux problèmes et questions se posent au cours de ce processus, comme la question de savoir si les algorithmes de correspondance utilisés pour les rencontres en ligne altèrent les résultats distributionnels du dating ou si les algorithmes de filtrage des nouvelles altèrent la distribution de l'opinion publique. Il est important d'étudier si des erreurs plus petites dans les algorithmes ou les données qu'ils utilisent pourraient se composer pour produire des effets à l'échelle de la société et comment les robots intelligents dans nos écoles, hôpitaux et centres de soins pourraient altérer le développement humain [118] et la qualité de vie [54] et potentiellement affecter les résultats pour les personnes handicapées [119].

D'autres questions dans ce domaine concernent le potentiel des machines à altérer le tissu social de manière plus fondamentale. Par exemple, les questions incluent dans quelle mesure et de quelles manières les gouvernements utilisent l'intelligence artificielle pour altérer la nature de la démocratie, la responsabilité politique et la transparence, ou la participation civique. D'autres questions incluent dans quelle mesure les machines intelligentes influencent la police, la surveillance et la guerre, ainsi que l'ampleur des effets de bots ont eu sur les résultats des élections [15] et si les systèmes d'ia qui aident à la formation de relations sociales humaines peuvent permettre une action collective.

Notamment, les études dans ce domaine examinent également comment les humains perçoivent l'utilisation des machines comme aides à la décision [120, 121], les préférences humaines pour et contre l'utilisation des algorithmes [121], et le degré auquel les productions humaines ou machines

réduisent ou créent une discorde chez les humains [122, 121]. Une question importante dans ce domaine inclut comment les humains répondent à la production croissante de biens et services économiques et au commerce avec des machines intelligentes [123]. En fin de compte, comprendre comment les systèmes humains peuvent être altérés par l'introduction de machines intelligentes dans nos vies est une composante vitale de l'étude du comportement des machines.

Les humains façonnent le comportement des machines. Les machines intelligentes peuvent altérer le comportement humain, et les humains créent, informent et modèlent les comportements des machines intelligentes. Nous façonnons le comportement des machines par la conception directe de systèmes d'ia et par la formation de ces systèmes sur des données d'entrée humaines actives et d'observations passives de comportements humains à travers les données que nous créons quotidiennement. Le choix des algorithmes à utiliser, des retours à fournir à ces algorithmes [124] et des données sur lesquelles les entraîner sont, à l'heure actuelle, des décisions humaines et peuvent directement altérer le comportement des machines. Une composante importante de l'étude du comportement des machines est de comprendre comment ces processus d'ingénierie altèrent les comportements résultants de l'ia, si les données d'entraînement sont responsables d'un comportement particulier de la machine, si c'est l'algorithme lui-même ou s'il s'agit d'une combinaison des deux. Le cadre exposé dans la Fig. 3 suggère qu'il y aura des questions complémentaires à chacune de ces questions. Examiner comment l'altération des paramètres du processus d'ingénierie peut altérer les comportements subséquents des machines intelligentes telles qu'elles interagissent avec d'autres machines et avec des humains dans des environnements naturels est central pour une compréhension holistique du comportement des machines.

Co-comportement homme-machine. Bien qu'il puisse être méthodologiquement commode de séparer les études sur les façons dont les humains façonnent les machines et vice versa, la plupart des systèmes d'ia fonctionnent dans des domaines où ils coexistent avec des humains dans des systèmes hybrides complexes [39, 67, 125, 134]. Les questions d'importance pour l'étude de ces systèmes incluent celles qui examinent les comportements qui caractérisent les interactions homme-machine, y compris la coopération, la compétition et la coordination - par exemple, comment les biais humains se combinent avec l'ia pour altérer les émotions ou croyances humaines [127, 126, 128], comment les tendances humaines s'associent aux algorithmes pour faciliter la propagation de l'information [55], comment les schémas de trafic peuvent être altérés dans les rues peuplées par un grand nombre de conducteurs sans conducteur et de conducteurs humains [32], et comment les schémas de trading peuvent être altérés par les interactions entre humains et agents de trading algorithmiques [28] ainsi que les facteurs qui peuvent faciliter la confiance et la coopération entre humains et machines [130, 128].

Un autre sujet dans ce domaine concerne l'automatisation robotique et logicielle du travail humain [129]. Ici, nous voyons deux types différents d'interactions homme-machine. L'une est que les machines peuvent améliorer l'efficacité humaine, comme dans la robotique et la chirurgie assistée par ordinateur. Une autre est que les machines peuvent remplacer les humains, comme dans le transport et la livraison sans conducteur. Cela conduit à des questions sur savoir si les machines finissent par faire plus de remplacement ou d'amélioration à long terme et comment les co-comportements homme-machine évolueront en conséquence.

Les exemples ci-dessus mettent en évidence que bon nombre des questions qui se rapportent au comportement hybride homme-machine doivent nécessairement examiner les boucles de rétroaction entre l'influence humaine sur le comportement des machines et l'influence des machines sur le comportement humain simultanément. Les chercheurs ont commencé à examiner les interactions

homme-machine dans des environnements de laboratoire formels, observant que les interactions avec des bots simples peuvent augmenter la coordination humaine [39] et que les bots peuvent coopérer directement avec les humains à des niveaux qui rivalisent avec la coopération humaine [130]. Cependant, il reste un besoin urgent de mieux comprendre les boucles de rétroaction dans des contextes naturels, dans lesquels les humains utilisent de plus en plus d’algorithmes pour prendre des décisions [131] et, par la suite, informent l’entraînement des mêmes algorithmes à travers ces décisions. En outre, à travers tous les types de questions dans le domaine de l’écologie comportementale des machines, il existe un besoin d’études qui examinent la dynamique à long terme de ces systèmes hybrides [132] avec un accent particulier sur les façons dont les interactions sociales humaines [133, 134] peuvent être modifiées par l’introduction de machines intelligentes [39].

1.6 Perspectives

Approfondir l’étude du comportement des machines est essentiel pour maximiser les bénéfices potentiels de l’ia pour la société. Les choix conséquents que nous faisons concernant l’intégration des agents d’ia dans la vie humaine doivent être faits avec une certaine compréhension des implications sociétales de ces choix. Pour fournir cette compréhension et cette anticipation, nous avons besoin d’une nouvelle discipline interdisciplinaire d’étude scientifique : le comportement des machines.

Pour que ce domaine réussisse, un certain nombre de considérations pertinentes existent. Premièrement, étudier le comportement des machines n’implique pas que les algorithmes d’ia aient nécessairement une agence indépendante ni n’implique que les algorithmes devraient porter la responsabilité morale de leurs actions. Si un chien mord quelqu’un, le propriétaire du chien est tenu responsable. Néanmoins, il est utile d’étudier les schémas comportementaux des animaux pour prédire un tel comportement aberrant. Les machines opèrent dans un cadre socio-technique plus large, et leurs parties prenantes humaines sont en fin de compte responsables de tout préjudice que leur déploiement pourrait causer.

Deuxièmement, certains commentateurs pourraient suggérer que traiter les systèmes d’ia comme des agents se concentre sur les données sous-jacentes sur lesquelles ces systèmes d’ia sont entraînés. En effet, aucun comportement n’est entièrement séparable des données environnementales sur lesquelles cet agent est entraîné ou développé ; le comportement des machines n’est pas d’exception. Cependant, il est tout aussi critique de comprendre comment les comportements des machines varient avec des entrées environnementales altérées qu’il est de comprendre comment les comportements des agents biologiques varient en fonction des environnements dans lesquels ils existent. À ce titre, les chercheurs en comportement des machines devraient se concentrer sur la caractérisation du comportement des agents à travers divers environnements, tout comme les scientifiques du comportement cherchent à caractériser les comportements politiques à travers différents contextes démographiques et institutionnels.

Troisièmement, les machines exhibent des comportements fondamentalement différents des animaux et des humains, nous devons donc éviter un anthropomorphisme et un zoomorphisme excessifs. Même si les méthodes scientifiques comportementales existantes peuvent s’avérer utiles pour l’étude des machines, les machines peuvent exhiber des formes d’intelligence et de comportement qualitativement différentes - même “extraterrestres” - de celles observées chez les agents biologiques. En outre, les scientifiques de l’ia peuvent disséquer et modifier les systèmes d’ia plus facilement et plus profondément que ce n’est le cas pour de nombreux systèmes vivants. Bien que des parallèles existent, l’étude des systèmes d’ia différera nécessairement de l’étude des systèmes vivants.

Quatrièmement, l'étude du comportement des machines nécessitera des efforts transdisciplinaires [82, 103] et entraînera tous les défis associés à ce type de recherche [135, 136]. Relever ces défis est vital [137]. Les universités et les agences de financement gouvernementales peuvent jouer un rôle important dans la conception d'études transdisciplinaires à grande échelle, neutres et fiables [138]. Cinquièmement, l'étude du comportement des machines nécessitera souvent une intervention expérimentale pour étudier les interactions homme-machine dans des contextes réels [139, 140]. Ces interventions pourraient altérer le comportement global du système, ayant éventuellement des effets indésirables sur les utilisateurs normaux [141]. Des considérations éthiques telles que celles-ci nécessitent une supervision prudente et des cadres standardisés.

Enfin, l'étude de systèmes algorithmiques ou robotiques intelligents peut entraîner des problèmes juridiques et éthiques pour les chercheurs qui étudient le comportement des machines. La rétro-ingénierie d'algorithmes peut nécessiter de violer les conditions d'utilisation de certaines plateformes ; par exemple, en créant de faux profils ou en masquant de véritables identités. Les créateurs ou mainteneurs des systèmes d'intérêt pourraient impliquer les chercheurs dans des poursuites judiciaires si la recherche nuit à la réputation de leurs plateformes. En outre, il n'est pas clair si la violation des conditions d'utilisation peut exposer les chercheurs à des sanctions civiles ou pénales (par exemple, par le biais du Computer Fraud and Abuse Act aux États-Unis), ce qui pourrait encore décourager ce type de recherche [142].

Comprendre les comportements et les propriétés des agents d'ia - et les effets qu'ils pourraient avoir sur les systèmes humains - est critique. La société peut bénéficier énormément des efficacités et de l'amélioration de la prise de décision qui peuvent provenir de ces agents. Dans le même temps, ces bénéfices pourraient faiblir sans minimiser les pièges potentiels de l'incorporation des agents d'ia dans la vie humaine quotidienne.

Reçu le 20 septembre 2018 ; accepté le 14 mars 2019 ; publié en ligne le 24 avril 2019.

Références

- [1] Simon, H. A. *The Sciences of the Artificial* (MIT Press, Cambridge, 1969). Simon demande s'il peut exister une science de l'"artificiel" qui produise une connaissance des objets et phénomènes artificiels.
- [2] Milner, R. A modal characterisation of observable machine-behaviour. In *Trees in Algebra and Programming, 6th Colloquium* 25-34 (Springer, 1981). Dans cette conférence invitée, Robin Milner expose l'idée d'étudier le comportement des machines à l'aide de la logique formelle.
- [3] Thomaz, A. L. & Breazeal, C. Teachable robots : understanding human teaching behavior to build more effective robot learners. *Artif. Intell.* **172**, 716-737 (2008).
- [4] Stone, P. et al. *Artificial Intelligence and Life in 2030. One Hundred Year Study on Artificial Intelligence : Report of the 2015-2016 Study Panel* <https://ai100.stanford.edu/2016-report> (Stanford University, 2016).
- [5] O'Neil, C. *Weapons of Math Destruction : How Big Data Increases Inequality and Threatens Democracy* (Broadway Books, 2016). Ce livre articule certains des risques posés par l'utilisation non critique des algorithmes dans la société et fournit une motivation pour l'étude du comportement des machines.
- [6] Future of Life Institute. Autonomous weapons : an open letter from AI & robotics researchers. <https://futureoflife.org/open-letter-autonomous-weapons/?cn-reloaded=1> (2015).
- [7] Dressel, J. & Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.* **4**, eaao5580 (2018).
- [8] Binns, R. et al. 'It's reducing a human being to a percentage' : perceptions of justice in algorithmic decisions. In *Proc. 2018 CHI Conference on Human Factors in Computing Systems* 377 (ACM, 2018).
- [9] Hudson, L., Owens, C. S. & Flannes, M. Drone warfare : blowback from the new American way of war. *Middle East Policy* **18**, 122-132 (2011).

- [10] Kahneman, D., Rosenfield, A. M., Gandhi, L. & Blaser, T. Noise : how to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review* <https://hbr.org/2016/10/noise> (2016).
- [11] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J. & Mullainathan, S. Human decisions and machine predictions. *Q. J. Econ.* **133**, 237-293 (2018).
- [12] Crawford, K. et al. *The AI Now report : The Social and Economic Implications of Artificial Intelligence Technologies in the Near-term*. https://ainowinstitute.org/AI_Now_2016_Report.pdf (2016).
- [13] Amodei, D. et al. Concrete problems in AI safety. Preprint at <https://arxiv.org/abs/1606.06565> (2016).
- [14] Bakshy, E., Messing, S. & Adamic, L. A. Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130-1132 (2015).
- [15] Bessi, A. & Ferrara, E. Social bots distort the 2016 U.S. Presidential election online discussion. *First Monday* **21**, 11 (2016).
- [16] Ferrara, E., Varol, O., Davis, C., Menczer, F. & Flammini, A. The rise of social bots. *Commun. ACM* **59**, 96-104 (2016).
- [17] Lazer, D. The rise of the social algorithm. *Science* **348**, 1090-1091 (2015).
- [18] Tufekci, Z. Engineering the public : big data, surveillance and computational politics. *First Monday* **19**, 7 (2014).
- [19] Lee, T.-S. & Chen, I.-F. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Syst. Appl.* **28**, 743-752 (2005).
- [20] Roszbach, K. Bank lending policy, credit scoring, and the survival of loans. *Rev. Econ. Stat.* **86**, 946-958 (2004).
- [21] Huang, C.-L., Chen, M.-C. & Wang, C.-J. Credit scoring with a data mining approach based on support vector machines. *Expert Syst. Appl.* **33**, 847-856 (2007).
- [22] Tsai, C.-F. & Wu, J.-W. Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Syst. Appl.* **34**, 2639-2649 (2008).
- [23] Chen, L. & Wilson, C. Observing algorithmic marketplaces in-the-wild. *SIGecom Exch.* **15**, 34-39 (2017).
- [24] Chen, L., Mislove, A. & Wilson, C. An empirical analysis of algorithmic pricing on Amazon marketplace. In *Proc. 25th International Conference on World Wide Web* 1339-1349 (International World Wide Web Conferences Steering Committee, 2016).
- [25] Hannak, A. et al. Bias in Online freelance marketplaces : evidence from TaskRabbit and Fiverr. In *Proc. ACM Conference on Computer Supported Cooperative Work and Social Computing* 1914-1933 (2017).
- [26] Cartlidge, J., Szostek, C., De Luca, M. & Cliff, D. Too fast too furious-faster financial-market trading agents can give less efficient markets. In *Proc. 4th International Conference on Agents and Artificial Intelligence* 126-135 (2012).
- [27] Kearns, M., Kulesza, A. & Nevmyvaka, Y. Empirical limitations on high-frequency trading profitability. *J. Trading* **5**, 50-62 (2010).
- [28] Wellman, M. P. & Rajan, U. Ethical issues for autonomous trading agents. *Minds Mach.* **27**, 609-624 (2017).
- [29] Farmer, J. D. & Skouras, S. An ecological perspective on the future of computer trading. *Quant. Finance* **13**, 325-346 (2013).
- [30] Perry, W. L., McInnis, B., Price, C. C., Smith, S. & Hollywood, J. S. *Predictive Policing : The Role of Crime Forecasting in Law Enforcement Operations* (RAND, 2013).
- [31] Bonnefon, J.-F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* **352**, 1573-1576 (2016).
- [32] Kooti, F. et al. Analyzing Uber's ride-sharing economy. In *Proc. 26th International Conference on World Wide Web* 574-582 (International World Wide Web Conferences Steering Committee, 2017).
- [33] Zeng, X., Fapojuwo, A. O. & Davies, R. J. Design and performance evaluation of voice activated wireless home devices. *IEEE Trans. Consum. Electron.* **52**, 983-989 (2006).
- [34] Hendriks, B., Meerbeek, B., Boess, S., Pauws, S. & Sonneveld, M. Robot vacuum cleaner personality and behavior. *Int. J. Soc. Robot.* **3**, 187-195 (2011).
- [35] Hitsch, G. J., Hortacsu, A. & Ariely, D. Matching and sorting in online dating. *Am. Econ. Rev.* **100**, 130-163 (2010).

- [36] Finkel, E. J., Eastwick, P. W., Karney, B. R., Reis, H. T. & Sprecher, S. Online dating : a critical analysis from the perspective of psychological science. *Psychol. Sci. Public Interest* **13**, 3-66 (2012).
- [37] Park, H. W., Rosenberg-Kima, R., Rosenberg, M., Gordon, G. & Breazeal, C. Growing growth mindset with a social robot peer. In *Proc. 2017 ACM/IEEE International Conference on Human-Robot Interaction* 137-145 (ACM, 2017).
- [38] Bemelmans, R., Gelderblom, G. J., Jonker, P. & de Witte, L. Socially assistive robots in elderly care : a systematic review into effects and effectiveness. *J. Am. Med. Dis. Assoc.* **13**, 114-120 (2012).
- [39] Shirado, H. & Christakis, N. A. Locally noisy autonomous agents improve global human coordination in network experiments. *Nature* **545**, 370-374 (2017). Dans cette étude hybride homme-machine, les auteurs montrent que des algorithmes simples injectés dans le jeu humain peuvent améliorer les résultats de coordination entre les humains.
- [40] Pichai, S. AI at Google : Our Principles. *Google Blog* <https://blog.google/topics/ai/ai-principles/> (2018).
- [41] Roff, H. M. The strategic robot problem : lethal autonomous weapons in war. *J. Mil. Ethics* **13**, 211-227 (2014).
- [42] Krishnan, A. *Killer Robots : Legality and Ethicality of Autonomous Weapons* (Routledge, 2016).
- [43] Voosen, P. The AI detectives. *Science* **357**, 22-27 (2017).
- [44] Szegedy, C. et al. Intriguing properties of neural networks. Preprint at <https://arxiv.org/abs/1312.6199> (2013).
- [45] Zhang, Q.-S. & Zhu, S.-C. Visual interpretability for deep learning : a survey. *Front. Inf. Technol. Electronic Eng.* **19**, 27-39 (2018).
- [46] Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. Preprint at <https://arxiv.org/abs/1702.08608> (2017).
- [47] Gebru, T. et al. Datasheets for datasets. Preprint at <https://arxiv.org/abs/1803.09010> (2018).
- [48] Mitchell, M. et al. Model cards for model reporting. Preprint at <https://arxiv.org/abs/1810.03993> (2018).
- [49] Lakkaraju, H., Kamar, E., Caruana, R. & Horvitz, E. Identifying unknown unknowns in the open world : representations and policies for guided exploration. In *Proc. 31st Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence* 2 (2017).
- [50] Johnson, N. et al. Abrupt rise of new machine ecology beyond human response time. *Sci. Rep.* **3**, 2627 (2013).
- [51] Appel, K., Haken, W. & Koch, J. Every planar map is four colorable. Part II : reducibility. *Illinois J. Math.* **21**, 491-567 (1977).
- [52] Appel, K. & Haken, W. Every planar map is four colorable. Part I : discharging. *Illinois J. Math.* **21**, 429-490 (1977).
- [53] Westlund, J. M. K., Park, H. W., Williams, R. & Breazeal, C. Measuring young children's long-term relationships with social robots. In *Proc. 17th ACM Conference on Interaction Design and Children* 207-218 (ACM, 2018).
- [54] Lorenz, T., Weiss, A. & Hirche, S. Synchrony and reciprocity : key mechanisms for social companion robots in therapy and care. *Int. J. Soc. Robot.* **8**, 125-143 (2016).
- [55] Vosoughi, S., Roy, D. & Aral, S. The spread of true and false news online. *Science* **359**, 1146-1151 (2018). Cette étude examine l'écologie hybride complexe des bots et des humains sur Twitter et constate que les humains propagent de fausses informations à des taux plus élevés que les bots.
- [56] Lazer, D. M. J. et al. The science of fake news. *Science* **359**, 1094-1096 (2018).
- [57] Roberts, M. E. *Censored : Distraction and Diversion Inside China's Great Firewall* (Princeton Univ. Press, 2018).
- [58] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. Algorithmic decision making and the cost of fairness. In *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 797-806 (ACM, 2017).
- [59] Kleinberg, J., Mullainathan, S. & Raghavan, M. Inherent trade-offs in the fair determination of risk scores. Preprint at <https://arxiv.org/abs/1609.05807> (2016).
- [60] Buolamwini, J. & Gebru, T. Gender shades : intersectional accuracy disparities in commercial gender classification. In *Proc. 1st Conference on Fairness, Accountability and Transparency* (eds Friedler, S. A. & Wilson, C.) **81**, 77-91 (PMLR, 2018).

- [61] Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. & Kalai, A. T. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proc. Advances in Neural Information Processing Systems* 4349-4357 (2016).
- [62] Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183-186 (2017).
- [63] Sweeney, L. Discrimination in online ad delivery. *Queueing Syst.* **11**, 10 (2013).
- [64] Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C. & Venkatasubramanian, S. Runaway feedback loops in predictive policing. Preprint at <https://arxiv.org/abs/1706.09847> (2017).
- [65] Angwin, J., Larson, J., Mattu, S. & Kirchner, L. Machine bias. *ProPublica* <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing> (2016).
- [66] Chouldechova, A., Benavides-Prado, D., Fialko, O. & Vaithianathan, R. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Proc. 1st Conference on Fairness, Accountability and Transparency* (eds Friedler, S. A. & Wilson, C.) **81**, 134-148 (PMLR, 2018).
- [67] Jennings, N. R. et al. Human-agent collectives. *Commun. ACM* **57**, 80-88 (2014).
- [68] Campbell, M., Hoane, A. J. & Hsu, F.-H. Deep blue. *Artif. Intell.* **134**, 57-83 (2002).
- [69] Schaeffer, J. et al. Checkers is solved. *Science* **317**, 1518-1522 (2007).
- [70] Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484-489 (2016).
- [71] Silver, D. et al. Mastering the game of Go without human knowledge. *Nature* **550**, 354-359 (2017).
- [72] Bowling, M., Burch, N., Johanson, M. & Tammelin, O. Heads-up limit hold'em poker is solved. *Science* **347**, 145-149 (2015).
- [73] Bellemare, M. G., Naddaf, Y., Veness, J. & Bowling, M. The arcade learning environment : an evaluation platform for general agents. *J. Artif. Intell. Res.* **47**, 253-279 (2013).
- [74] Wellman, M. P. et al. Designing the market game for a trading agent competition. *IEEE Internet Comput.* **5**, 43-51 (2001).
- [75] Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I. & Osawa, E. RoboCup : the robot world cup initiative. In *Proc. 1st International Conference on Autonomous Agents* 340-347 (ACM, 1997).
- [76] Russakovsky, O. et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211-252 (2015).
- [77] Lin, T.-Y. et al. Microsoft COCO : common objects in context. In *Proc. European Conference on Computer Vision* (eds Fleet, D. et al.) **8693**, 740-755 (Springer International Publishing, 2014).
- [78] Davis, J. & Goadrich, M. The relationship between precision-recall and ROC curves. In *Proc. 23rd International Conference on Machine Learning* 233-240 (ACM, 2006).
- [79] van de Sande, K. E. A., Gevers, T. & Snoek, C. G. M. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 1582-1596 (2010).
- [80] Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU : a method for automatic evaluation of machine translation. In *Proc. 40th Annual Meeting on Association for Computational Linguistics* 311-318 (Association for Computational Linguistics, 2002).
- [81] Zhou, Z., Zhang, W. & Wang, J. Inception score, label smoothing, gradient vanishing and -log(D(x)) alternative. Preprint at <https://arxiv.org/abs/1708.01729> (2017).
- [82] Epstein, Z. et al. Closing the AI knowledge gap. Preprint at <https://arxiv.org/abs/1803.07233> (2018).
- [83] Tinbergen, N. On aims and methods of ethology. *Ethology* **20**, 410-433 (1963).
- [84] Nesse, R. M. Tinbergen's four questions, organized : a response to Bateson and Laland. *Trends Ecol. Evol.* **28**, 681-682 (2013).
- [85] Das, R., Hanson, J. E., Kephart, J. O. & Tesauro, G. Agent-human interactions in the continuous double auction. In *Proc. 17th International Joint Conference on Artificial Intelligence* 1169-1178 (Lawrence Erlbaum, 2001).
- [86] Deng, Y., Bao, F., Kong, Y., Ren, Z. & Dai, Q. Deep direct reinforcement learning for financial signal representation and trading. *IEEE Trans. Neural Netw. Learn. Syst.* **28**, 653-664 (2017).

- [87] Galceran, E., Cunningham, A. G., Eustice, R. M. & Olson, E. Multiplicity decision-making for autonomous driving via changepoint-based behavior prediction : theory and experiment. *Auton. Robots* **41**, 1367-1382 (2017).
- [88] Ribeiro, M. T., Singh, S. & Guestrin, C. Why should I trust you ? Explaining the predictions of any classifier. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135-1144 (ACM, 2016).
- [89] Smilkov, D., Thorat, N., Kim, B., Viegas, F. & Wattenberg, M. SmoothGrad : removing noise by adding noise. Preprint at <https://arxiv.org/abs/1706.03825> (2017).
- [90] Nevmyvaka, Y., Feng, Y. & Kearns, M. Reinforcement learning for optimized trade execution. In *Proc. 23rd International Conference on Machine Learning* 673-680 (ACM, 2006).
- [91] Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L. & Konstan, J. A. Exploring the filter bubble : the effect of using recommender systems on content diversity. In *Proc. 23rd International Conference on World Wide Web* 677-686 (ACM, 2014).
- [92] Dalvi, N. & Domingos, P. Mausam, Sanghai, S. & Verma, D. Adversarial classification. In *Proc. Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 99-108 (ACM, 2004).
- [93] Globerson, A. & Roweis, S. Nightmare at test time : robust learning by feature deletion. In *Proc. 23rd International Conference on Machine Learning* 353-360 (ACM, 2006).
- [94] Biggio, B. et al. Evasion attacks against machine learning at test time. In *Proc. Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 387-402 (Springer, 2013).
- [95] Tramèr, F. et al. Ensemble adversarial training : attacks and defenses. Preprint at <https://arxiv.org/abs/1705.07204> (2017).
- [96] Parkes, D. C. & Wellman, M. P. Economic reasoning and artificial intelligence. *Science* **349**, 267-272 (2015).
- [97] Wagner, A. *Robustness and Evolvability in Living Systems* (Princeton Univ. Press, 2013).
- [98] Edwards, H. & Storkey, A. Censoring representations with an adversary. Preprint at <https://arxiv.org/abs/1511.05897> (2015).
- [99] Zemel, R., Wu, Y., Swersky, K., Pitassi, T. & Dwork, C. Learning fair representations. In *Proc. International Conference on Machine Learning* 325-333 (2013).
- [100] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C. & Venkatasubramanian, S. Certifying and removing disparate impact. In *Proc. 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 259-268 (ACM, 2015).
- [101] Cully, A., Clune, J., Tarapore, D. & Mouret, J.-B. Robots that can adapt like animals. *Nature* **521**, 503-507 (2015). Cette étude caractérise un robot piloté par un algorithme adaptatif qui imite l'adaptation et les comportements des animaux.
- [102] Bongard, J., Zykov, V. & Lipson, H. Resilient machines through continuous self-modeling. *Science* **314**, 1118-1121 (2006).
- [103] Leibo, J. Z. et al. Psychlab : a psychology laboratory for deep reinforcement learning agents. Preprint at <https://arxiv.org/abs/1801.08116> (2018). Dans cette étude, les auteurs utilisent des outils comportementaux des sciences de la vie dans l'étude des comportements des machines.
- [104] Subrahmanian, V. S. et al. The DARPA Twitter bot challenge. Preprint at <https://arxiv.org/abs/1601.05140> (2016).
- [105] Carrascosa, J. M., Mikians, J., Cuevas, R., Erramilli, V. & Laoutaris, N. I always feel like somebody's watching me : measuring online behavioural advertising. In *Proc. 11th ACM Conference on Emerging Networking Experiments and Technologies* 13 (ACM, 2015).
- [106] Datta, A., Tschantz, M. C. & Datta, A. Automated Experiments on Ad Privacy Settings. *Proc. Privacy Enhancing Technologies* **2015**, 92-112 (2015).
- [107] Giusti, A. et al. A machine learning approach to visual perception of forest trails for mobile robots. *IEEE Robot. Autom. Lett.* **1**, 661-667 (2016).
- [108] Berdahl, A., Torney, C. J., Ioannou, C. C., Faria, J. J. & Couzin, I. D. Emergent sensing of complex environments by mobile animal groups. *Science* **339**, 574-576 (2013).
- [109] Couzin, I. D. et al. Uninformed individuals promote democratic consensus in animal groups. *Science* **334**, 1578-1580 (2011).

- [110] Rubenstein, M., Cornejo, A. & Nagpal, R. Programmable self-assembly in a thousand-robot swarm. *Science* **345**, 795-799 (2014).
- [111] Kernbach, S., Thenius, R., Kernbach, O. & Schmickl, T. Re-embodiment of honeybee aggregation behavior in an artificial micro-robotic system. *Adapt. Behav.* **17**, 237-259 (2009).
- [112] Bak, P., Chen, K., Creutz, M. & Tsujikawa, K. *The Game of Life* (World Scientific, 1989).
- [113] Nunes, E., Menezes, G., Silva, A. & Rocha, E. Wikibots : a study of bot editing in Wikipedia. In *Proc. 11th International Symposium on Open Collaboration* 1-7 (ACM, 2015).
- [114] Mordatch, I. & Abbeel, P. Emergence of grounded compositional language in multi-agent populations. Preprint at <https://arxiv.org/abs/1703.04908> (2017).
- [115] Menkveld, A. J. The economics of high-frequency trading : taking stock. *Annu. Rev. Financ. Econ.* **8**, 1-24 (2016).
- [116] Menczer, F., Sapiezynski, P., Ferrara, E. & Lehmann, S. Evidence of complex contagion of information in social media : an experiment using Twitter bots. *PLoS ONE* **12**, e0184148 (2017). Cette étude présente une intervention expérimentale sur Twitter utilisant des bots et fournit des preuves que la diffusion de l'information est décrite le plus précisément par une contagion complexe.
- [117] Bainbridge, L. Ironies of automation. *Automatica* **19**, 775-779 (1983).
- [118] Jeong, S., Breazeal, C., Logan, D. & Weinstock, P. Huggable : the impact of embodiment on promoting socio-emotional interactions for young pediatric inpatients. In *Proc. 2018 CHI Conference on Human Factors in Computing Systems* 495 (ACM, 2018).
- [119] Salisbury, E., Kamar, E. & Morris, M. R. Toward scalable social alt text : conversational crowdsourcing as a tool for refining vision-to-language technology for the blind. *Proc. 5th AAAI Conference on Human Computation and Crowdsourcing* (2017).
- [120] Awad, E. et al. The Moral Machine experiment. *Nature* **563**, 59-64 (2018).
- [121] Dietvorst, B. J., Simmons, J. P. & Massey, C. Algorithm aversion : people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* **144**, 114-126 (2015).
- [122] Gray, K. & Wegner, D. M. Feeling robots and human zombies : mind perception and the uncanny valley. *Cognition* **125**, 125-130 (2012).
- [123] Brynjolfsson, E. & Mitchell, T. What can machine learning do ? Workforce implications. *Science* **358**, 1530-1534 (2017).
- [124] Christiano, P. F. et al. Deep reinforcement learning from human preferences. In *Proc. Advances in Neural Information Processing Systems* **30** (eds Guyon, I. et al.) 4299-4307 (Curran Associates, 2017).
- [125] Tsvetkova, M. et al. Understanding human-machine networks : a cross-disciplinary survey. *ACM Comput. Surv.* **50**, 12 :1-12 :35 (2017).
- [126] Hilbert, M., Ahmed, S., Cho, J., Liu, B. & Luu, J. Communicating with algorithms : a transfer entropy analysis of emotions-based escapes from online echo chambers. *Commun. Methods Meas.* **12**, 260-275 (2018).
- [127] Kramer, A. D. I., Guillory, J. E. & Hancock, J. T. Experimental evidence of massive-scale emotional contagion through social networks. *Proc. Natl Acad. Sci. USA* **111**, 8788-8790 (2014).
- [128] Kamar, E., Hacker, S. & Horvitz, E. Combining human and machine intelligence in large-scale crowdsourcing. In *Proc. 11th International Conference on Autonomous Agents and Multiagent Systems* 467-474 (International Foundation for Autonomous Agents and Multiagent Systems, 2012).
- [129] Jackson, M. *The Human Network : How Your Social Position Determines Your Power, Beliefs, and Behaviors* (Knopf Doubleday, 2019).
- [130] Crandall, J. W. et al. Cooperating with machines. *Nat. Commun.* **9**, 233 (2018). Cette étude examine la coopération algorithmique avec les humains et fournit un exemple de méthodes qui peuvent être utilisées pour étudier le comportement des systèmes hybrides homme-machine.
- [131] Wang, D., Khosla, A., Gargeya, R., Irshad, H. & Beck, A. H. Deep learning for identifying metastatic breast cancer. Preprint at <https://arxiv.org/abs/1606.05718> (2016).
- [132] Pentland, A. *Social Physics : How Social Networks Can Make Us Smarter* (Penguin, 2015).
- [133] Lazer, D. et al. Computational social science. *Science* **323**, 721-723 (2009).

- [134] Aharony, N., Pan, W., Ip, C., Khayal, I. & Pentland, A. Social fMRI : investigating and shaping social mechanisms in the real world. *Pervasive Mobile Comput.* **7**, 643-659 (2011).
- [135] Ledford, H. How to solve the world’s biggest problems. *Nature* **525**, 308-311 (2015).
- [136] Bromham, L., Dinnage, R. & Hua, X. Interdisciplinary research has consistently lower funding success. *Nature* **534**, 684-687 (2016).
- [137] Kleinberg, J. & Oren, S. Mechanisms for (mis)allocating scientific credit. In *Proc. 43rd Annual ACM Symposium on Theory of Computing* 529-538 (ACM, 2011).
- [138] Kannel, W. B. & McGee, D. L. Diabetes and cardiovascular disease. The Framingham study. *J. Am. Med. Assoc.* **241**, 2035-2038 (1979).
- [139] Krafft, P. M., Macy, M. & Pentland, A. Bots as virtual confederates : design and ethics. In *Proc. 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* 183-190 (ACM, 2017).
- [140] Meyer, M. N. Two cheers for corporate experimentation : The A/B illusion and the virtues of data-driven innovation. *Colorado Technol. Law J.* **13**, 273 (2015).
- [141] Xing, X. et al. Take this personally : pollution attacks on personalized services. In *Proc. 22nd USENIX Security Symposium* 671-686 (2013).
- [142] Patel, K. Testing the limits of the First Amendment : how a CFAA prohibition on online antidiscrimination testing infringes on protected speech activity. *Columbia Law Rev.* <https://doi.org/10.2139/ssrn.3046847> (2017).