Des Agents Virtuels pour un Monde Meilleur

Magalie Ochs
Equipe R2I, Laboratoire
d'Informatique et des Systèmes,
Aix Marseille Université

ABSTRACT

Les agents artificiels (robots humanoïdes et personnages virtuels), peuvent être exploités pour un monde meilleur : pour augmenter l'empathie entre les gens, améliorer leur bien-être, promouvoir les coopérations, encourager les comportements prosociaux, réduire les inégalités sociales par l'apprentissage, etc. S'intégrant dans le récent courant de recherche appelé *l'Informatique Prosociale*, un certain nombre de travaux dans le domaine des agents artificiels sont aujourd'hui menés avec cet objectif. Outre ces recherches dédiées, quelque que soit le domaine applicatif de nos travaux de recherche, il est impératif de veiller à la conception de nos agents artificiels. En particulier, l'apparence et le comportement de ces derniers peuvent avoir une influence inattendue et non souhaitable sur le comportement de l'utilisateur et ses croyances, non seulement au sein de l'environnement virtuel mais se poursuivant aussi ensuite dans le monde réel. Dans cet article, ces problématiques sont discutées à la lumière des résultats de recherche récents autour de l'influence des agents artificiels. De solutions sont discutés pour veiller à ce que nos agents artificiels puissent être utilisés pour tendre vers un monde meilleur.

1 Améliorer l'humain à travers les machines : *l'informatique prosociale*

Aujourd'hui, les films de sciences fictions et les médias grand public mettent en évidence principalement le *côté obscur* de l'intelligence artificielle, suscitant une anxiété et une peur générales face à un avenir peuplé de machines qui manipuleront la population et prendront le contrôle de nos vies. Cependant, l'Intelligence Artificielle, et en particulier les agents artificiels (robots humanoïdes et personnages virtuels), peuvent également être exploités pour un monde meilleur, pour augmenter l'empathie entre les gens, promouvoir les coopérations, pour encourager les comportements prosociaux, pour réduire les inégalités sociales par l'apprentissage, etc.

Les questionnements sur l'utilisation de l'Intelligence Artificielle pour le bien de la société sont récents. Différents évènements sont organisés dans ce sens au niveau mondial ; par exemple, le workshop AAAI WS : « AI and Operations Research for Social Good » ou le congrès annuel « AI for Good Global Summit » à Genève. En 2018, un nouveau courant de recherche émerge dans notre communauté : l'*informatique prosociale*, qui se définit comme « l'informatique visant à soutenir et à promouvoir des actions qui profitent à la société et aux autres » (Paiva et al., 2018). Le comportement des systèmes interactifs humanoïdes peut en effet influencer le comportement humain. Par exemple, des travaux de recherche dans (Burnham, et Hare, 2007) montrent que simplement le regard du robot Kimset déclenche un comportement plus altruiste chez l'humain dans un jeu de biens public (« public good games ») en comparaison à l'absence du robot les regardant. L'empathie, un élément essentiel au comportement prosocial, a été largement exploré par plusieurs chercheurs en Informatique Affective afin d'intégrer cette composante dans les systèmes interactifs humanoïdes (Ochs et al., 2008 ; Paiva et al., 2017). L'informatique prosociale inclut les systèmes interactifs permettant la formation des individus.

Dans le domaine de l'e-éducation et de la formation, un intérêt croissant est apparu autour de la « formation par la simulation » dans des environnements virtuels. De nombreux environnements de réalité virtuelle ou mixte ont été développés pour le développement de compétences techniques. Plus récemment, un certain nombre de travaux de recherche ont été menés autour de dispositifs permettant de simuler une interaction sociale avec un agent conversationnel animé (ACA) et ainsi entrainer ses propres compétences sociales (« Virtual Agents for Social Skills Training (VASST) » (Bruijnes, et al., 2019)). Les compétences sociales, partie intégrante de l'intelligence sociale, se définissent comme la capacité de gérer son comportement, verbal et non-verbal, pour construire une relation sociale avec autrui (Albrecht, 2006).
Les recherches montrent que les systèmes interactifs humanoïdes peuvent permettre d'améliorer ces compétences interactionnelles (Aylett, 2007). Par exemple, dans le projet européen Tardis (Anderson et al., 2013), un personnage virtuel jouant le rôle d'un recruteur est utilisé pour former les jeunes adultes aux entretiens d'embauche. Le système permet de détecter en temps réel le comportement verbal et

non-verbal du candidat. Cette détection permet de calculer des indices comportementaux pouvant refléter la qualité de l'entretien (e.g. direction du regard, posture) et ainsi lui donner un retour sur l'entretien à la fois sur le contenu verbal et non-verbal. Dans le projet eCUTE (Hall et al., 2011), les enfants et les jeunes adultes sont confrontés à différents scénarios d'interactions avec des personnages virtuels pour les sensibiliser aux différences culturelles et développer leur empathie. La plateforme FearNot ! (Aylett, 2007) a été développée pour lutter contre les comportements d'harcèlement scolaire à l'école. Elle repose sur la propension des individus à ressentir de l'empathie envers les systèmes interactifs humanoïdes. En jouant le rôle de la victime à travers un personnage virtuel et en développant ainsi de l'empathie pour les victimes d'harcèlement scolaire, les harceleurs changent leur comportement. Dans (Chollet et al., 2018), une audience virtuelle permet de s'entrainer à la prise de parole en public. L'audience virtuelle s'adapte au comportement verbal et non-verbal de l'orateur. Un comportement inadapté de l'orateur engendrera ainsi une audience virtuelle simulant l'ennui à travers un ensemble de postures. A l'inverse, un comportement engageant de l'orateur sera reflété sur le comportement simulé de l'audience virtuelle. Dans le projet VICTEAMS (Huguet et al., 2016), un environnement de réalité virtuelle peuplé d'ACAs est développé afin de former les leaders d'équipes médicales dans des situations de crise. Dans nos travaux, nous avons développé un environnement de réalité virtuelle pour permettre aux médecins de s'entrainer à l'annonce d'évènements indésirables graves avec un patient virtuel (Ochs et al., 2019).

Loin d'être exhaustif, ces exemples d'applications orientés vers la formation des individus montrent qu'ils peuvent permettre de réduire les inégalités sociales, promouvoir des comportements d'empathie au centre des comportements prosociaux et améliorer le bien-être des individus.

Les travaux de recherche en informatique prosociale impliquant des systèmes interactifs humanoïdes restent cependant peu nombreux. Chaque domaine d'application nécessite plusieurs années de recherche pour modéliser les compétences sociales et les environnements de simulation correspondant. Aujourd'hui, de nombreux domaines d'applications restent encore à explorer.

Il existe bien sûr des recherches comme présentées ci-dessus ayant une visée spécifique de formation ou de développement des comportements prosociaux des individus. Outre ces recherches dédiées, quelque que soit le domaine applicatif de nos travaux de recherche, pour une informatique prosociale, il est impératif de veiller à la conception de nos agents artificiels (robots humanoïdes ou ACAs). En particulier suivant l'apparence et le comportement de ces derniers, des stéréotypes peuvent être véhiculés ayant une influence directe sur le comportement de l'utilisateur et sur ses croyances, non seulement au sein de l'environnement virtuel mais se poursuivant aussi ensuite dans le monde réel.

2 L'évidence des stéréotypes associés aux agents virtuels

Plusieurs travaux de recherches ont exploré le paradigme CASA (« Computer are Social Actors ») sous le prisme des stéréotypes. En d'autres termes, appliquons-nous des stéréotypes de la même manière sur les agents virtuels que sur les humains ?

Les stéréotypes sociaux sont un ensemble de croyances à propos des autres, reflétant une généralisation partagée à propos de membres d'un groupe social. Les stéréotypes sont à la fois descriptifs (comment un membre d'un groupe est) et perspectif (quel comportement est attendu d'un membre d'un groupe social donné). Les stéréotypes négatifs, même inconscients, sont à l'origine de comportements de discrimination et d'hostilité.

Plus particulièrement, les recherches sur les *stéréotypes liés aux genres* mettent en évidence les traits associés aux hommes décrits comme plus agressifs, forts, compétents, indépendants alors que les femmes sont associées à des traits d'individu gentils, aidantes, chaleureuses et communicantes (Fiske, 1998 ; Haines et al., 2016). Le comportement des femmes et des hommes sont contraints par des « script sociaux », appris implicitement tout au long de la vie, et amenant chaque individu à avoir un ensemble d'attentes quant aux comportements de son interlocuteur suivant le genre de ce dernier. On s'attend ainsi à ce que les hommes prennent des positions plus dominantes et soit plus compétents que les femmes.

Des recherches montrent que les stéréotypes liés aux genres persistent dans les interactions humain-machine. Par exemple, le comportement des utilisateurs n'est pas le même suivant l'apparence genrée des ACAs. Des comportements sexistes se retrouvent lors d'interaction humain-ACA féminine (De Angei et al., 2006). L'incarnation n'est pas nécessaire, une étude de (Nass et al. 1007) montre en effet

que les voix de synthèses masculines sont perçues comme plus compétentes que les voix de femmes. Outre la perception, le genre de l'ACA a une influence directe sur le comportement de l'utilisateur. Par exemple Lee (2003) a montré que les utilisateurs suivent plus les conseils de personnages virtuels masculin lorsque ces derniers sont liés à une thématique stéréotypées masculin (e.g. sport) et plus les conseils des ACA féminins lorsqu'ils sont liés à des sujets stéréotypés féminins (e.g. les cosmétiques). Dans un système tutoriel intelligent, le genre du tuteur virtuel semble avoir un impact direct sur les performances d'apprentissage de l'utilisateur, un ACA d'apparence masculine étant plus en adéquation avec le stéréotype du professeur (Moreon et al., 2002).

Aux stéréotypes de genre s'ajoute le *stéréotype d'attractivité* selon lequel un individu attractif est perçu comme ayant plus de compétences sociales et comme étant plus extraverti que des individus non attractifs. Dans le domaine des agents virtuels, l'attractivité de l'agent à une influence directe sur son comportement. Par exemple, un agent virtuel vendeur semble plus persuasif pour la vente lorsqu'il est attractif (Holzwarth et al., 2006). Globalement, les ACA attractifs sont perçus comme plus compétent socialement et intellectuellement que les ACA moins attractifs (Khan et al., 2009). Ce stéréotype de l'attractivité dépend de l'apparence de l'ACA. En effet, quelques études mettent en évidence l'impact des agents virtuels ou avatars *sexualisés* sur les attitudes et les croyances de l'utilisateur. Par exemple, l'incarnation d'un avatar sexualisé dans un jeu vidéo diminue les croyances des femmes sur leur propres capacités à atteindre un but dans le *monde réel et diminue les croyances des hommes sur les capacités cognitives des femmes dans le monde réel* (Behm et al., 2009). Dans le domaine des jeux vidéo, une étude démontre que jouer à des jeux vidéo sexistes renforcent les attitudes sexistes des individus masculin dans le *monde réel* (Stermer et al., 2015). Tout porte à croire, que plus nous développerons des ACAs sexualisés plus nous renforcerons les stéréotypes et les attitudes sexistes dans *le monde réel*.

3 Des stéréotypes des agents virtuels à la discrimination dans le monde réel

Les systèmes interactifs humanoïdes, et plus globalement les algorithmes en Intelligence Artificielle, sont créés et optimisés pour résoudre une tâche particulière ; par exemple pour la reconnaissance automatique de la parole, la détection d'objet ou la classification de document. Cependant, comme le soulignent (Rahwan et al., 2019), les comportements de ces algorithmes peuvent induire des effets sociétaux inattendus à la fois positifs et négatifs ; des conséquences non anticipées par leur créateur.

La question de l'impartialité des algorithmes d'intelligence artificielle (« Fainess AI ») et plus globalement de leur effet sur la société doit être au centre des préoccupations avant l'intégration de ces outils dans notre quotidien. Ces derniers peuvent, de manière insidieuse, renforcer des stéréotypes et des discriminations au sein de notre société. Un exemple frappant, développé ci-dessous, est celui *des stéréotypes et discriminations liés au genre.*

Le renforcement des stéréotypes. Les assistants vocaux, des systèmes interactifs, comme l'assistant Siri d'Apple ou Alexa d'Amazon, ont pris une place grandissante dans notre quotidien. Les utilisateurs interagissent en langage naturel avec ces assistants à travers un large éventail de requêtes possibles (recherche sur internet, changer de chanson, etc..). Ces assistants vocaux sont souvent non incarnés (dénoués de corps virtuel ou physique). Force est de constater que ces assistants vocaux sont genrés avec des voix principalement de jeunes femmes. Ce choix est motivé par les entreprises par des préférences sociétales spécifiant que les voix de femmes sont préférées à celles des hommes. Les résultats des recherches menées sur ce sujet sont pourtant loin d'être tranchés : les préférences des voix dépendant de la tâche et du genre de l'utilisateur (Mitchell et al., 2011 ; Stromberg, 2013). Il n'en reste pas moins que les voix de femme sont perçues comme plus aidantes et coopératives que les voix d'homme perçues comme plus autoritaires (Nass et Brave, 2005). Ils semblent néanmoins que les voix de femme sont perçues comme moins intelligente et amènent alors à plus de tolérances lors d'erreur de reconnaissance vocale que des voix d'hommes (Nass, 2010). Elles semblent donc effectivement plus adaptées au rôle d'assistant vocal où la reconnaissance de parole peut échouer régulièrement. Ces assistant*e*s vocales reflètent, renforcent et perpétuent les stéréotypes sociétaux, retranscrits dans des produits de nouvelles technologies. Ainsi comme le soulignent des chercheurs de l'université d'Harvard travaillant sur les biais inconscients (Lai et Mahzarin, 2018), plus la société reflètera une association entre femme et assistante, même dans un monde virtuel, plus les femmes, dans le monde réel, seront associées à des rôles d'assistantes et pénalisées si elles n'adoptent pas ce rôle. Le biais genré des systèmes interactifs peut donc nous seulement perpétuer les stéréotypes discriminants mais aussi les renforcer et les étendre.

Le harcèlement sexuel favorisé par les agents virtuels. Outre la persistante de stéréotypes, une autre problématique est le comportement dialogique des assistantes virtuelles personnelles face à l'harcèlement sexuel des utilisateurs, représentant plus de 5 % des interactions (Coren, 2016). La plupart d'entre elles répondent de manière évasive ou positive, montrant une grande tolérance à ce genre de comportement. Comme le souligne les auteurs du rapport de l'Unesco pour promouvoir la parité dans le numérique (West et al., 2019), les comportements de ces assistantes virtuelles renforcent les stéréotypes d'assistantes serviles et l'idée selon laquelle ces réponses ambiguës, voir positives, sont appropriées aux harcèlements sexuels. Ces éléments sont d'autant plus problématiques que les assistantes virtuelles vocales ou incarnées (i.e. dotées d'une représentation physique) sont de plus en plus réalistes en termes de voix et d'apparence, donnant de plus en plus l'impression aux utilisateurs qu'ils interagissent avec une humaine, mais stéréotypés, docile, offrant des réponses simplistes, et parfois incompétente à réaliser des tâches très simples (Clark, 2018). Un certain nombre de recherches montrent de plus que la représentation sexualisée des systèmes interactifs induit des comportements sexistes des utilisateurs qui perdurent dans le monde réel *après* leur interaction dans les mondes virtuels (comme évoqué Section 2). Les conséquences d'une interaction avec un personnage virtuel adoptant un comportement de soumission (par exemple à travers le regard) vont jusqu'à une influence du mythe de l'acceptation du viol (mythe selon laquelle beaucoup de femmes ont le désir inconscient d'être violée - Rape mythe Acceptance – RMA) (Fox et al., 2009). Les stéréotypes modélisés à travers les systèmes interactifs, quel que soit leur type, engendrent donc des comportements qui dépassent la sphère du virtuel, en véhiculant une image nuisible de la femme.

Reproduction des discriminations par l'apprentissage sur des données discriminatoire. Les femmes représentent aujourd'hui 12% des chercheurs travaillant sur les algorithmes d'apprentissage automatique (Mantha et al., 2018 ; West et al., 2019). L'absence de parité dans ce domaine est un réel problème étant donnée la place que prennent les algorithmes d'intelligence artificielle dans notre quotidien. En particulier, la plupart de ces algorithmes sont fondés sur des masses de données qui contiennent des stéréotypes liés au genre. Entrainés sur ces données, les algorithmes répliquent ces stéréotypes (West et al, 2019 ; Bernheim et al., 2019). Par exemple, l'algorithme d'Amazon permettant d'attribuer une note automatiquement entre 0 et 5 aux CV des candidats discrimine les CV contenant le mot « femme », comme « président du club d'échec féminin » car l'algorithme a été entrainé sur des données principalement de candidats masculins (Dastin, 2018). D'autres travaux de recherches sur des algorithmes d'analyse automatique des textes du site d'actualité google montrent que ces algorithmes entrainés sur ces données construisent un modèle sexiste concernant la carrière des individus, associant les hommes à des métiers de programmeurs et les femmes à des activités de femmes au foyer (Bolukbasi, 2016). De la même manière, le module de reconnaissance vocale de Google, principalement entrainé sur des voix masculines, est nettement plus performant à reconnaitre une voix d'homme que de femme (Tatman, 2016). Des algorithmes d'intelligence artificielle ne véhiculant pas de stéréotypes ou de discrimination de genre sont possibles. Une réflexion est aujourd'hui menée pour veiller, à travers un ensemble de dispositifs d'évaluation de ces algorithmes et à travers des équipes paritaires de développement d'algorithmes d'apprentissage automatique, aux biais sexistes de ces derniers (West et al., 2019 ; Bernheim et al., 2019).

4 Vers des solutions pour un monde meilleur

Pour lutter contre la discrimination émanant des interactions avec les agents artificiels, plusieurs solutions peuvent être envisagées.

Des agents virtuels androgynes ou un renversement des stéréotypes dans le monde virtuel. Aujourd'hui, un certain nombre d'entreprises développant des assistants vocaux ou des agents virtuels ont ajouté des voix masculines en option, ou supprimé la voix féminine par défaut, ou ont opté pour une incarnation non-humanoïde (e.g. des animaux) et ont supprimé les comportements d'excuses et de flirt en réponse au harcèlement sexuel des utilisateurs.

Bien sûr, la première solution évidente semble être de ne pas développer de systèmes interactifs stéréotypés, éventuellement même des agents artificiels androgynes. Cependant, les systèmes interactifs humanoïdes genrés peuvent aussi être utilisés pour briser des stéréotypes, par exemple en utilisant des assistantes personnelles féminines pour accomplir des tâches difficiles.

Intégrer des valeurs morales dans les entreprises et dans nos agents virtuels. La prédominance des assistantes vocales peut s'expliquer par la prédominance masculine

dans les équipes de conception de tels assistants (West et al., 2019, Campolo et al., 2017). Outre la parité des équipes de conception, des algorithmes permettant d'implémenter des valeurs morales doivent être intégrés dans les assistants pour éviter des dérives telles que celles rencontrées lors de la mise en service du chatbot de Microsoft qui après une journée d'apprentissage sur les post de twitter a associé l'égalité des genres au féminisme et le féminisme à un cancer (Caliskan, 2017). Dans cette perspective, dans une édition pour enfant d'assistante vocale, Amazon propose d'imposer que les requêtes soient exécutées uniquement si elles sont posées avec politesse.

Lutter contre la menace du stéréotype. De nombreuses recherches en cognition sociale ont montré que la peur d'être stéréotypée négativement dans un domaine d'aptitude peut nuire à l'apprentissage et aux performances des individus dans ce domaine. Ce phénomène est appelé « la menace du stéréotype » (Steele, 1997). La menace du stéréotype augmente chez les individus les pensées préoccupantes concernant des jugements stéréotypés ce qui diminue les ressources cognitives de ces derniers (Schmader & Johns, 2003) et entraîne ainsi une sous-performance (Schmader et al., 2008). Plusieurs recherches ont montré que ce phénomène explique les performances plus faibles des filles en Sciences, en comparaison aux garçons, et *vice et versa* dans le domaine de la lecture, conformément aux stéréotypes. Par ailleurs, des solutions existent pour réduire voir supprimer ce phénomène, par exemple en présentant différemment les exercices et les tests (Regner et al., 2010). Dans le domaine des agents virtuels, certains travaux, encore très peu nombreux, ont exploré l'utilisation d'ACA ou d'avatars pour réduire la menace du stéréotype et ainsi augmenter les performances de l'apprenant. En particulier, les personnages virtuels sont utilisés pour incarner des modèles sociaux et ainsi modifier les attitudes de l'apprenant et sa motivation. Par exemple, dans (Rosesenberg-Kia et al., 2008), les chercheurs ont montré l'efficacité d'un personnage virtuel féminin, ressemblante aux apprenantes, utilisé pour présenter des modèles de femmes en sciences et ainsi modifier leurs attitudes vis-à-vis des Sciences. D'autres travaux montrent que l'apparence genrée de l'avatar contrôlé par un apprenant dans un environnement virtuel pouvait permettre de réduire la menace du stéréotype (Ratan and Sah, 2015).

Dans d'autres domaines d'application, des ACAs avec un discours féministe sont étudiés (Shi et al., 2015). Cette première étude montre que le féminisme d'un ACA n'est pas approprié face à une population non-féministe.

Equité des agents virtuels. La question de l'impartialité/l'équité des algorithmes d'intelligence artificielle (« Fainess AI ») se pose concernant les algorithmes de prise de décision des agents virtuels. Des travaux proposent des algorithmes permettant des décisions impartiales des agents en se fondant sur des fonctions d'utilité équitables (De Jong et al., 2008). Pour pallier la problématique des algorithmes d'apprentissage automatique appris sur des données potentiellement biaisées et répliquant des décisions discriminantes, des mesures sont proposées très récemment pour évaluer l'impartialité des algorithmes et en particulier le caractère discriminant (direct ou indirect) des décisions issues de l'apprentissage (Corbett-Davies et al., 2018 ; Zhang et al., 2018). Ces mesures devraient s'ajouter systématiquement à celles communément utilisées pour mesurer les performances des algorithmes d'apprentissage (e.g. F-score, rappel). La performance d'un algorithme d'apprentissage se juge en effet dans sa capacité à être équitable pour ne pas reproduire les discriminations sociétales.

4. Conclusion et perspectives

Nos agents conversationnels animés et nos robots humanoïdes sont aujourd'hui conçus en s'inspirant de l'humain et en se nourrissant de données humaines. Comme le soulignent (Rahwan et al., 2019), à la fois *l'humain façonne la machine mais la machine façonne aussi l'humain.* Ainsi, dans une vision optimiste de l'intelligence artificielle, est apparu de nouveaux courants de recherche, comme *l'informatique prosociale*, visant à utiliser le potentiel de l'intelligence artificielle pour améliorer les comportements humains. Plusieurs applications sont développées dans les laboratoires en France et à l'international pour aider les individus et améliorer le bien-être sociétal.

Dans une vision plus pessimiste de l'Intelligence Artificielle, il n'en reste pas moins que les effets sociétaux de l'utilisation de certains algorithmes ou de systèmes interactifs peuvent être complètement inattendus et non souhaitables. Le comportement de nos agents artificiels doit être étudié en dehors de la tâche pour laquelle ils ont été conçus pour évaluer les implications sociétales de l'introduction des agents artificiels dans notre société. L'étude scientifique dans une approche interdisciplinaire du comportement des machines est essentielle pour que les agents artificiels restent bénéfiques à notre société (Rahwan et al., 2019).

Le design des apparences et des comportements des agents artificiels ne doit pas être guidé par une préférence des consommateurs. Cette vision consumériste aujourd'hui des entreprises développant des assistants vocaux ou des ACAs renforcent dans notre société les stéréotypes et les discriminations dans le monde réel. En tant que chercheurs, non pas face à des consommateurs mais face à des utilisateurs, nous devrions nous questionner sur la prise en compte des préférences des utilisateurs quant au comportement et à l'apparence des agents virtuels. Ces éléments ne doivent-ils pas être guidé par notre vision d'un monde égalitaire plutôt que par la réplication d'une société aujourd'hui inégalitaire au sein du virtuel ?

Notre rôle en tant que chercheurs de la communauté ACAI est essentiel. Nous orientons par nos recherches les systèmes interactifs de demain et c'est donc de notre responsabilité de créer un monde virtuel dénué de stéréotypes pour un monde réel meilleur.

REFERENCES

Albrecht, K. (2006). Social intelligence: The new science of success. John Wiley & Sons.

Anderson, K., André, E., Baur, T., Bernardini, S., Chollet, M., Chryssadou, E., Damian, I., Ennis, C.,
Egges, A., Gebhard, P., *et al.* (2013). The tardis framework: intelligent virtual agents for social coaching in job interviews. In Advances in computer entertainment, pages 476-491. Springer.

Aylett, R., Vala, M., Sequeira, P., & Paiva, A. (2007). Fearnot!–an emergent narrative approach to virtual dramas for anti-bullying education. In *International Conference on Virtual Storytelling* (pp. 202-205). Springer, Berlin, Heidelberg.

Behm-Morawitz, E., & Mastro, D. (2009). The effects of the sexualization of female video game characters on gender stereotyping and female self-concept. *Sex roles*, *61*(11-12), 808-823.

Bernheim, A et Vincent F. , L'Intelligence artificielle, pas sans elles !, Belin, collection « Egale à égal » du Laboratoire de l'égalité, 2019.

Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V. and Kalai, A. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Proceedings of the 30th International Conference on Neural Information Processing Systems, pp. 4356–6

Bruijnes, M., Linssen, J., & Heylen, D. (2019). Special issue editorial: Virtual Agents for Social Skills Training. Journal on Multimodal User Interfaces, Volume 13, Issue 1.

Burnham, T. C., & Hare, B. (2007). Engineering human cooperation. *Human nature*, *18*(2), 88-108.

Caliskan, A., Bryson, J., and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. Science, Vol. 365, No. 6334, pp. 183–6.

Campolo, A. et al. 2017. AI Now 2017 Report. New York, AI Now Institute, New York University

Chollet, M., Ghate, P., Neubauer, C., & Scherer, S. (2018, November). Influence of Individual Differences when Training Public Speaking with Virtual Audiences. In Proceedings of the 18th International Conference on Intelligent Virtual Agents (pp. 1-7). ACM.

Clark, P. 2018. The digital future is female – but not in a good way. Financial Times, 17 June 2018.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Coren, M. J. 2016. Virtual assistants spend much of their time fending off sexual harassment. Quartz, Oct, 2016

Dastin, J. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, 9 October 2018.

De Angeli, A., & Brahnam, S. (2006). Sex stereotypes and conversational agents. *Proc. of Gender and Interaction: real and virtual women in a male world, Venice, Italy*.

De Jong, S., Tuyls, K., & Verbeeck, K. (2008, May). Artificial agents learning human fairness. In Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2 (pp. 863-870). International Foundation for Autonomous Agents and Multiagent Systems.

Fiske, S. T. (1998). Stereotyping, prejudice, and discrimination. *The handbook of social psychology*, *2*(4), 357-411.

Fox, J., & Bailenson, J. N. (2009). Virtual virgins and vamps: The effects of exposure to female characters' sexualized appearance and gaze in an immersive virtual environment. *Sex roles*, *61*(3-4), 147-157.

Haines, E. L., Deaux, K., & Lofaro, N. (2016). The times they are a-changing… or are they not? A comparison of gender stereotypes, 1983–2014. *Psychology of Women Quarterly*, *40*(3), 353-363.

Hall, Lynne, et al. "Fostering empathic behaviour in children and young people: interaction with intelligent characters embodying culturally specific behaviour in virtual world simula-tions." INTED2011 Proceedings (2011): 2804-2814.

Holzwarth, M., Janiszewski, C., & Neumann, M. M. (2006). The influence of avatars on online consumer shopping behavior. *Journal of marketing*, *70*(4), 19-36.

L. Huguet, D. Lourdeaux, N. Sabouret, M.-H. Ferrer. Perturbed Communication in a Virtual Environment to Train Medical Team Leaders. Stud Health Technol Inform. 220:146-9. 2016

Khan, R., & De Angeli, A. (2009, August). The attractiveness stereotype in the evaluation of embodied conversational agents. In *IFIP Conference on Human-Computer Interaction* (pp. 85-97). Springer, Berlin, Heidelberg.

Lai, C. and Mahzarin, B. 2018. The Psychology of Implicit Bias and the Prospect of Change. 31 January 2018. Cambridge, Mass., Harvard University

Lee, E. J. (2003). Effects of "gender" of the computer on informational social influence: the moderating role of task type. *International Journal of Human-Computer Studies*, *58*(4), 347-362.

Mantha, Y. and Hudson, S. 2018. Estimating the gender ratio of AI researchers around the world. Medium, 17 August 2018.

Mitchell W. et al. 2011. Does social desirability bias favour humans? Explicit-implicit evaluations of synthesized speech support a new HCI model of impression management. Computers in Human Behavior, Vol. 27, No. 1. pp. 402–12

Moreno, K. N., Person, N. K., Adcock, A. B., Eck, R. N. V., Jackson, G. T., & Marineau, J. C. (2002, November). Etiquette and efficacy in animated pedagogical agents: The role of stereotypes. In *AAAI symposium on personalized agents, Cape Cod, MA*.

Nass, C., Moon, Y., & Green, N. (1997). Are machines gender neutral? Gender-stereotypic responses to computers with voices. *Journal of applied social psychology*, *27*(10), 864-876.

Nass, C. and Brave, S. 2005. Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship.Cambridge, Mass., MIT Press

Nass, C. I., & Yen, C. (2010). *The man who lied to his laptop: What machines teach us about human relationships*. New York:: Current.

Ochs, M., Pelachaud, C., & Sadek, D. (2008, May). An empathic virtual dialog agent to improve human-machine interaction. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 1* (pp. 89-96). International Foundation for Autonomous Agents and Multiagent Systems.

Paiva, A., Leite, I., Boukricha, H., & Wachsmuth, I. (2017). Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *7*(3), 11.

Paiva, A., Santos, F. C. (2018, April). Engineering pro-sociality with autonomous agents. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Jennings, N. R. (2019). Machine behaviour. *Nature*, *568*(7753), 477.

Ratan, R., & Sah, Y. J. (2015). Leveling up on stereotype threat: The role of avatar customization and avatar embodiment. Computers in Human Behavior, 50, 367-374.

Régner, I., Smeding, A., Gimmig, D., Thinus-Blanc, C., Monteil, J. M., & Huguet, P. (2010). Individual differences in working memory moderate stereotype-threat effects. Psychological Science, 21(11), 1646-1648.

Rosenberg-Kima, R. B., Baylor, A. L., Plant, E. A., & Doerr, C. E. (2008). Interface agents as social models for female students: The effects of agent visual presence and appearance on female students' attitudes and beliefs. Computers in Human Behavior, 24(6), 2741-2756.

Shi, L., Bickmore, T., & Edwards, R. (2015, August). A feminist virtual agent for breastfeeding promotion. In *International Conference on Intelligent Virtual Agents* (pp. 461-470). Springer, Cham.

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. Journal of personality and social psychology, 85(3), 440.

Schmader, T., Johns, M., & Forbes, C. (2008). An integrated process model of stereotype threat effects on performance. Psychological review, 115(2), 336.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. American psychologist, 52(6), 613.

Stermer, S. P., & Burkley, M. (2015). SeX-Box: Exposure to sexist video games predicts benevolent sexism. *Psychology of Popular Media Culture*, *4*(1), 47.

Stromberg, J. 2013. Why women like deep voices and men prefer higher ones. Smithsonian Magazine, 24 April 2013

Tatman, R. 2016. Google's speech recognition has a gender bias. Making Noise and Hearing Things, 12 July 2016

West, M., Kraut, R., & Ei Chew, H. (2019). I'd blush if I could: closing gender divides in digital skills through education.

Zhang, J., & Bareinboim, E. (2018, April). Fairness in decision-making—the causal explanation formula. In *Thirty-Second AAAI Conference on Artificial Intelligence*.