# Modélisation individuelle: stéréotypes, préjugés et discrimination à l'embauche

*Kawtare DKHISSATE, Pierre BONDESAN, Anthony PEDROSA DOS SANTOS*

# Résumé

La présente étude vise à expliquer l'expression d'un comportement sexiste dans le cadre d'un recrutement professionnel par une modélisation d'un réseau de neurones artificiels. Une première partie sera consacrée à un état de l'art de la catégorisation sociale et dans quelle mesure cela conduit à l'activation de stéréotypes de genre. Dans cette même partie, nous expliquerons l'ensemble des processus cognitifs impliqués pour un recruteur homme lors de l'évaluation d'un CV selon la théorie du double processus de Kahneman (1974). Dans une seconde partie, nous présenterons la modélisation. Afin de rendre compte au mieux de la théorie du double processus, nous avons choisi d'utiliser un réseau de neurones récurrents permettant de garder une trace en mémoire de la réponse du système 1 afin de pouvoir être utilisée dans le traitement de la réponse du système 2. Enfin, nous discuterons des apports et des limites de cette approche méthodologique ainsi que des critiques émises à l'encontre de la théorie de Kahneman.

# Table des matières

# Introduction

En france, la discrimination des femmes à l'embauche reste présente et a été mis en avant par de nombreuses études de testing (Bonte & AFP, 2018). Cette inégalité est d'autant plus importante lors du recrutement de femmes dans des secteurs d'activité considérés comme typiquement masculin. Dans cette étude, nous avons cherché à modéliser le comportement sexiste d'un homme dans le cadre du recrutement au regard de la théorie sur la catégorisation sociale et des stéréotypes de genre. L'intérêt de la modélisation est de pouvoir rendre compte de l'ensemble des processus de traitement (système 1 et système 2) impliqués lors de l'évaluation d'un CV.

# 1. Partie théorique

## 1.1 Etat de l'art sur les stéréotypes et préjugés

### La théorie de la catégorisation sociale

La catégorisation automatique des stimuli environnants est une capacité humaine qui nous permet d'évoluer dans un environnement social complexe en simplifiant nos jugements tout en préservant nos capacités cognitives (Macrae, Milne & Bodenhausen, 1994). Dans la plupart des situations sociales, il n'est pas utile de disposer d'informations complètes sur les personnes et les objets.

Selon la théorie de la catégorisation sociale (Kinzler et al., 2017), cette capacité de segmentation par caractéristiques est présente dès la petite enfance, conférant aux nourrissons une capacité de déduction par association de caractéristiques communes afin de mieux comprendre le monde qui l'entoure. La catégorisation sociale est un mécanisme sous-tendu par des processus automatiques. Lors d'une tâche d'apprentissage d'informations sur d'autres individus, une étude a montré que des enfants âgés de 3 à 6 ans vont présenter plus de confusion lors du rappel d'informations concernant des personnes de la même origines ethniques ou du même genre que lorsque les personnes n'ont pas les mêmes origines ethniques ou ne sont pas du même genre (Weisman, 2014). Cet effet était

amoindri chez les enfants vivant dans un environnement avec des personnes ayant des origines ethniques variées.

La catégorisation est un mécanisme nécessaire et adapté qui tend à légitimer les catégories en leur conférant une existence afin d'organiser l'ensemble de ces connaissances et d'être capable de les mobiliser rapidement. Cependant, cette création de catégories sociales va avoir un rôle crucial dans la façon dont vont interagir les êtres humains dans leurs rapports sociaux. Les effets délétères auxquels la recherche sur la catégorisation sociale portent un fort intérêt sont étudiés sous le prisme des stéréotypes et préjugés, en particulier lorsqu'ils conduisent au déploiement de comportements racistes ou sexistes.

## L'impact de la catégorisation sociale dans les rapports sociaux

Ces constructions de catégories sont des images simplifiées de la réalité. Dans un contexte social, il s'agit d'attribuer des caractéristiques stéréotypées à un groupe. La construction d'un stéréotype à l'égard d'un groupe est tirée d'un apprentissage influencé par l'héritage culturel de la société dans laquelle vit l'individu et dont il ne peut échapper (Ehrlich, 1973). Les théoriciens classiques de la psychologie sociale (Allport, 1954; Ehrlich, 1973; Tajfel; 1981) avancent l'argument des préjugés inévitables selon lequel la connaissance d'un stéréotype entraîne indissociablement des préjugés. Les préjugés sont des attitudes, états mentaux pouvant amener une personne à adopter une opinion ou un comportement. Les recherches plus récentes montrent qu'un stéréotype n'entraîne pas systématiquement l'expression d'un préjugé (Devine, 1989). Lors d'une étude menée sur des vétérans de guerre, il a été montré qu'il n'y avait pas de lien systématique entre la connaissance d'un stéréotype sur les personnes juives et les personnes noires et le degré d'expression d'un préjugé à l'égard de ces groupes (Devine, 1989).

Le lien entre un stéréotype et un préjugé n'est plus uniquement envisagé comme étant la conséquence de la connaissance d'un stéréotype mais également par l'intégration de ce stéréotype (Devine, 1989). La connaissance et d'un stéréotype et la croyance envers celui-ci seraient ainsi issues de deux processus différents.

## Les parts automatique et contrôlée d'un stéréotype

La construction d'un stéréotype est inévitable car il repose sur les processus automatiques qui sous-tendent la catégorisation, présents dès l'enfance et dépend l'exposition à son environnement. La connaissance d'un stéréotype est le fruit d'un apprentissage : exposition répétée et création d'associations en mémoire entre des caractéristiques et un groupe. La connaissance d'un stéréotype s'active lorsque la personne est de nouveau exposée au stimulus (Schneider & Shiffrin, 1977). L'activation d'un stéréotype de manière non consciente a été démontrée de nombreuses fois par le Test d'Association Implicite ("*IAT effect*"). L'IAT mesure le temps de réaction de l'évaluation de la valence d'un mot (par exemple, juger le mot "fleur" comme étant bon ou mauvais) et de la catégorisation d'un groupe de personnes à partir de photos (par exemple en catégorisant la photo d'une personne en déterminant si elle est blanche ou noire). Les résultats montrent que lorsque la touche à appuyer est la même pour le mot négatif et pour la photo de la personne noire, le temps de réponse est significativement plus rapide que lorsque la touche est la même pour le mot positif et la personne noire (Greenwald et al., 1998). Cela montre que l'activation d'un stéréotype intervient de manière automatique.

Au contraire, l'acceptation et la croyance en un stéréotype est issu des processus de contrôle, ce qui demandent une attention active de la part de l'individu. Ces processus sont plus flexibles que les processus automatiques car ils peuvent être inhibés lorsqu'ils entrent en conflit avec la connaissance d'un stéréotype. Lors d'une étude où les participants devaient lister leurs opinions à l'égard des personnes noires, les résultats montrent que malgré le même niveau de connaissance des stéréotypes, les personnes avec un haut niveau de préjugé listaient des caractéristiques en lien avec le stéréotype alors que les personne avec un faible niveau de préjugé listaient des caractéristiques autres que le stéréotype et donc inhibaient leur réponse (Devine, 1989).

## La préférence sociale: effet in-group/out-group

La catégorisation sociale se distingue des autres formes de catégorisation par le fait que la personne est amenée elle-même à se placer dans son groupe d'appartenance. Les membres du groupe d'appartenance vont être perçus comme faisant partie de l'in-group et

les membres d'un groupe extérieur comme faisant partie de l'out-group. Les relations intergroupes vont conduire à adopter des comportements pro-groupe. Selon la théorie du conflit réaliste, Sherif (1966) a montré que le simple fait d'être rangé dans une équipe entraîne une évaluation plus positive des membres de son groupe et une évaluation plus négative des membres de l'out-group. Dans le cadre compétitif, cela entraîne également des attitudes négatives à l'égard de l'autre groupe pouvant aller jusqu'à la déshumanisation (Haris & Fiske, 2006).

Dès l'enfance, la présence de préférence sociale basée sur le genre ou sur les origines ethniques va apparaître (Kinzler et al., 2017). La théorie des schémas du genre cherche à expliquer comment se construisent les stéréotypes liés au genre et comment ils se maintiennent à l'âge adulte.

## Théorie des schémas du genre

La théorie des schémas du genre propose que l'être humain traite l'information sous le prisme du genre féminin et du genre masculin. Il existe deux types de schémas liés au genre, le premier est un schéma général "superordonné" qui aide les enfants à classer les objets, les caractéristiques et les traits dans des catégories de base masculines et féminines. Le second est une version plus restreinte du schéma, appelée le schéma du " propre genre" que les enfants utilisent pour identifier et apprendre en profondeur les informations correspondant à leur "propre genre". Ces deux types de schémas permettent aux enfants de traiter des informations sur des événements, des objets, des attitudes, des comportements et des rôles et, à leur tour, de catégoriser ces aspects en termes de masculin ou de féminin, ou comme similaires ou différents de l'enfant (Martin & Halverson, 1981).

A termes, l'enfant aura construit des catégories stéréotypées de genre avec des traits et des rôles attribués aux hommes et aux femmes et se sera positionné dans l'une de ces catégories selon son genre. Ce traitement de l'information par le prisme du genre dès le plus jeune âge conduit à des associations fortes qui seront maintenues à l'âge adulte.

## Comportement sexiste dans le contexte du recrutement

Les stéréotypes de genre vont être un frein à l'embauche pour les femmes, en particulier dans les secteurs d'activité en incongruence avec ces stéréotypes. Peu importe

l'évaluateur, ces stéréotypes vont être activés au moment du recrutement. Les études sur l'IAT ont montré une évaluation pro-homme pour les activités scientifiques chez les hommes et les femmes (Fahrell, 2015).

Cependant, la retranscription des ces stéréotypes en comportement sexiste de la part du recruteur vont dépendre de sa croyance en ces stéréotypes. Si le recruteur est un homme, il pourra avoir une motivation supplémentaire à adopter un comportement pro-groupe en favorisant un homme et en dévaluant une femme.

Lors de l'évaluation, la présence d'autres personnes peut influencer la décision de l'évaluateur. Une étude a montré que plus la catégorie est accessible plus il sera facile de catégoriser, plus la catégorie est inaccessible moins il sera facile de catégoriser et d'adopter des comportements pro-groupe (Haslam et al., 1991). Dans cette étude de Haslam et al. (1991), lorsque les participants hommes se trouvent en présence d'hommes et de femmes, leur identité sociale d'homme leur paraît moins saillante que lorsqu'ils sont entourés uniquement d'hommes.

## La théorie du double processus: système 1 et système 2

La théorie du double processus soutient que des capacités cognitives de haut niveau tel que le raisonnement proviennent du résultat d'un double traitement (Kahneman, 1974, 2011). Ce système double implique des processus cognitifs distincts qui interviennent à différents niveaux de consciences.

Le système 1 agit de manière non consciente et fait appel à l'ensemble des processus autonomes. Parmi ces processus, on retrouve la récupération d'information provenant de la mémoire associative implicite (Atkinson & Shiffrin, 1968). La mémoire associative implicite est la récupération spontanée d'une information suite à l'exposition d'un stimulus. Cette récupération est non consciente et est possible suite à un apprentissage qui a renforcé l'association entre deux concepts. Ce système de traitement a pour objectif de donner une réponse rapide et nécessitant peu de ressources et qui serait particulièrement développé dans un besoin de survie. Les réponses du système 1 sont ainsi suffisantes pour faire face à la majorité des épreuves de la vie quotidienne.

Le système 2 traite l'information en parallèle en impliquant l'ensemble des fonctions cognitives dont la mémoire de travail, ce qui nécessite un niveau d'attention soutenue (Morewedge & Kahneman, 2010). Le rôle de la mémoire de travail grâce au buffer épisodique permet d'aller récupérer des informations issues de la mémoire à long terme (Baddeley, 2007). Ces informations sont nécessaires pour raisonner sous la forme de règles logiques. Le système 2 étant limité par la capacité de la mémoire de travail, le traitement est plus lent et mobilise plus de ressources que le système 1.

C'est également le système 2 qui va jouer un rôle de contrôle de la réponse du système 1 car les réponses issus du système 1 comportent de nombreux biais de jugement (Morewedge & Kahneman, 2010; Evans, 2003). Si, aux regards des différentes règles de logique, la réponse du système 1 n'est pas adaptée, le système 2 va pouvoir inhiber la réponse du système 1 et produire une réponse plus adaptée (Evans, 2003).

## Théorie du double processus et psychologie sociale

La théorie du double processus semble particulièrement bien rendre compte des deux composantes implicite et explicite d'un stéréotype (Morewedge & Kahneman, 2010). La connaissance d'un stéréotype se construit par la catégorisation sociale qui associe des traits et caractéristiques à un groupe. Ces associations sont activées suite à l'exposition d'un stimulus, ce qui peut conduire à des biais de jugement du système 1. L'expression ou l'inhibition d'un préjugé va dépendre de la croyance et de son adéquation avec le stéréotype (Devine, 1989). Déterminer si l'on est en adéquation ou inadéquation avec le stéréotype nécessite un raisonnement qui s'identifie aux processus de contrôle du système 2 (cf. *Figure 1*).

Théorie du double processus dans le cas d'un stéréotype de genre

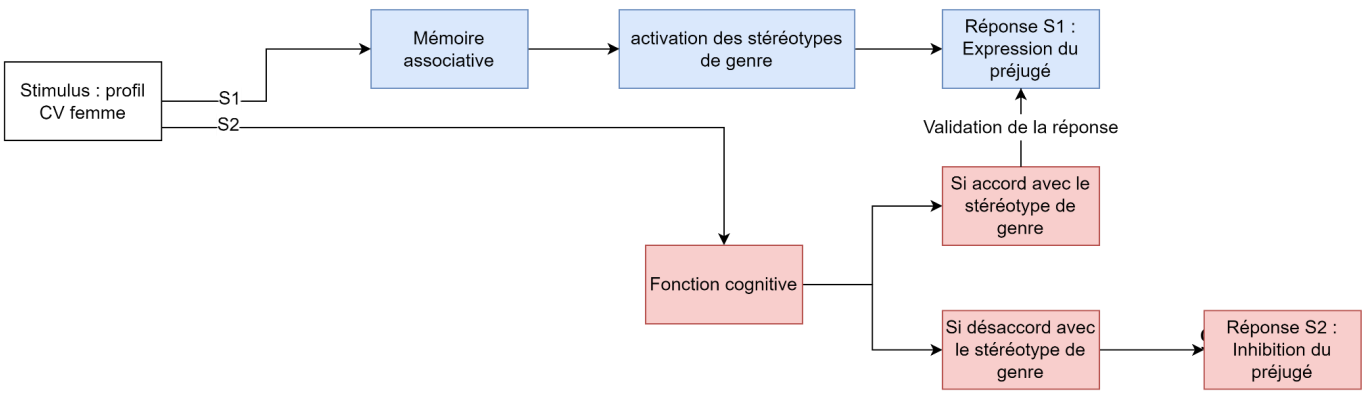


*Figure 1: Exemple de la théorie du double processus appliquée à un stéréotype de genre*

Dans une étude parue en 2011, Evans introduit une troisième catégorie de processualité, le système 3, qui serait responsable de l'initiation du système 2 lorsqu'il résout un conflit entre le système 1 heuristique et le système 2 analytique pour contrôler le comportement finalement déployé. Certains auteurs émettent plusieurs critiques envers cette théorie (Keren, 2013). Ils soulignent notamment le manque de résultats empiriques pouvant confirmer la théorie d'Evans. C'est pourquoi nous avons préféré l'approche théorique de Morewedge et Kahneman (2010) qui ne considère pas de troisième système. Ainsi, dans notre approche théorique, le système 2 s'active toujours, vérifiant la pertinence de la réponse émise par le système 1 et, le cas échéant, modifiant la réponse finale (cf. *Figure 1*).

## 1.2 Etat de l'art sur la modélisation de neurones formels

### L'approche connexionniste

L'approche connexionniste sous-tend la possibilité de retranscrire le schéma de fonctionnement du cerveau humain en se basant sur l'architecture de neurones (unités) interconnectés (réseau) par des synapses (connecteurs). C'est ce que l'on appellera un réseau de neurones « formels » élémentaires interconnectés entre eux. Le terme de neurone formel renvoie à une représentation mathématique et informatique du neurone biologique. Concrètement, il s'agit d'un réseau de neurones connectés virtuellement où chaque neurone ou unité reçoit de l'information entrante et émet de l'information sortante. La première machine connexionniste voit le jour en 1957 avec le Perceptron de Frank Rosenblatt. S'inspirant des travaux de McCulloch et Pitts, le perceptron est un algorithme d'apprentissage supervisé basé sur un réseau de neurones formels.

### L'approche symbolique/classique

L'intelligence artificielle symbolique a pour but de reproduire le raisonnement humain en le modélisant par un ensemble de symboles. Cette représentation symbolique est

soumise à un ensemble de règles, d'instructions permettant d'édicter à la machine dans ses prises de décision. L'objectif ici est de reproduire une logique de raisonnement.

### L'approche neuro-symbolique

L'intelligence artificielle utilisant l'approche neuro-symbolique est la combinaison des approches connexionnistes et symboliques. Le neuro-symbolisme combine l'apprentissage machine reposant sur la modélisation sous formes de réseaux de neurones formels des connaissances et l'approche symbolique logique des systèmes à bases de règles explicites.

### Les réseaux de neurones récurrents

Dans un réseau neuronal à action directe (feedforward), les informations ne se déplacent que dans un sens : de la couche d'entrée à la couche de sortie, en passant par les couches cachées. L'information se déplace directement à travers le réseau et ne touche jamais deux fois un nœud.

Les réseaux neuronaux feedforward n'ont aucune mémoire des données qu'ils reçoivent et sont incapables de prédire ce qui va se passer. Comme un réseau feedforward ne prend en compte que l'entrée actuelle, il n'a aucune notion de l'ordre dans le temps. Il ne peut tout simplement pas se "souvenir" de ce qui s'est passé dans le passé, sauf de sa formation.

Dans un réseau de neurones récurrents, l'information passe par une boucle. Lorsqu'il prend une décision, il tient compte de l'entrée actuelle et de ce qu'il a appris des entrées qu'il a reçues précédemment. L'image ci-dessous illustre la différence de flux d'informations entre un RNN et un réseau feed-forward.

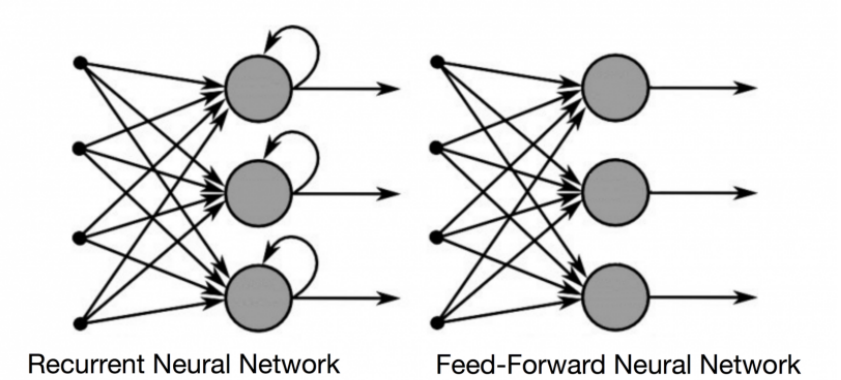


*Figure 2: RNN vs FNN*

Le réseau neuronal récurrent possède une mémoire interne. Le RNN est récurrent par nature car il exécute la même fonction pour chaque entrée de données, tandis que la sortie de l'entrée actuelle dépend du calcul précédent. Après avoir produit la sortie, celle-ci est copiée et renvoyée dans le réseau récurrent. Pour prendre une décision, il considère l'entrée actuelle et la sortie qu'il a apprise de l'entrée précédente.

Contrairement aux réseaux feedforwards, les RNN peuvent utiliser leur état interne (mémoire) pour traiter des séquences d'entrées. Cela les rend applicables à des tâches telles que la reconnaissance de l'écriture manuscrite non segmentée et connectée ou la reconnaissance vocale. Dans les autres réseaux neuronaux, toutes les entrées sont indépendantes les unes des autres. Mais dans un RNN, toutes les entrées sont liées les unes aux autres.

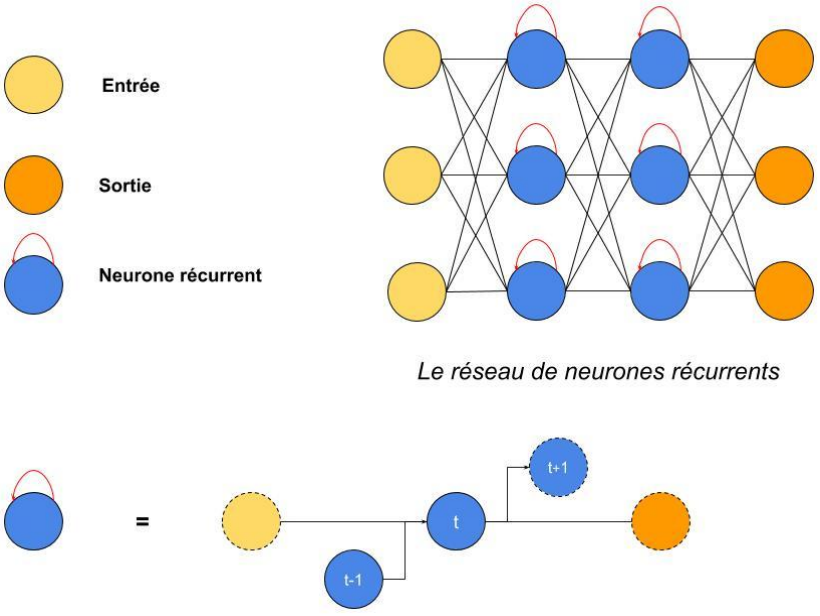



*Figure 3: Architecture d'un RNN*

Le réseau de neurones récurrents correspond parfaitement à la modélisation de la théorie du double processus de Kahneman. L'objectif étant de créer un algorithme qui tient compte des données précédentes durant son apprentissage. Cette méthode permet de

modéliser le système 1 et le système 2 en possédant une boucle permettant de créer une mémoire associée au contexte sexiste ou non sexiste lors d'un recrutement.

## Les précédentes recherches modélisant la théorie du double processus

De nombreuses recherches et expériences en intelligence artificielle (IA) se focalisent sur l'explication des résultats d'un algorithme. Notamment l'expérience de Hikaru & al. (2020) où les chercheurs se sont inspirés de la théorie du double processus en psychologie. Ils ont développé un système d'IA qui fait une inférence logique et produit ainsi une explication interprétable. Ils ont pour cela procédé en deux temps. Dans un premier temps, Ils demandent à des travailleurs sociaux de saisir des caractéristiques compréhensibles par l'homme pour des classes dans un texte. Les caractéristiques compréhensibles par l'homme sont des termes pour décrire des attributs tels que les couleurs, les tailles et des motifs. Lorsqu'ils donnent les termes, les travailleurs doivent suivre des formulaires spécifiques car cela facilite l'extraction des mots-clés des caractéristiques des textes saisis par les utilisateurs et les mots-clés seront utilisés comme mots de recherche. Pour éviter de générer des classificateurs qui ont des critères en double dans la deuxième phase, ils utilisent la distance de "Word Mover" pour calculer la similarité des termes donnés par les travailleurs et filtrer ceux qui ont une grande similarité avec les autres. Les résultats ont montré un taux de réussite élevé (66% d'accuracy).

Dans les algorithmes d'apprentissage par renforcement profonds tels que REINFORCE et DQN, les réseaux neuronaux effectuent des sélections d'actions sans anticipation, ce qui est analogue au Système 1. Contrairement à l'intuition humaine, leur formation ne bénéficie pas d'un "système 2" pour suggérer des actions fortes. Cependant,les travaux d'Anthony & al. (2017) proposent une itération experte (EXIT) qui utilise une recherche arborescente comme analogue du système 2, ce qui facilite la formation du réseau neuronal. À son tour, le réseau neuronal est utilisé pour améliorer les performances de la recherche arborescente en fournissant des "intuitions rapides" pour guider la recherche représentant ainsi le système 1.

# Partie modélisation

## A. Expérience 1

### But de l'expérience

L'objectif de la première expérience est de mettre en pratique nos connaissances issues de la théorie du double processus et des réseaux de neurones récurrents appliquée au stéréotype de genre lors d'un contexte de recrutement. L'idée est de modéliser un algorithme permettant la prédiction d'une note d'un Curriculum Vitae d'une femme lors d'un recrutement. Le recruteur est un homme et deux contextes sont existants : un contexte favorisant un comportement sexiste, c'est un contexte *in group*, où le recruteur est en présence d'hommes et note le CV d'une femme, le second est un contexte ne favorisant pas un comportement sexiste, c'est un contexte *out group* où le recruteur cette fois est en présence de femmes et doit noter le CV d'une femme.

### Participants

Une base de données à été créée afin d'entraîner l'algorithme. La base de données comprend au total 500 lignes correspondant à 500 individus. Les caractéristiques spécifiques les représentants se présentent comme suit : différentes colonnes comprenant des données liées à l'âge, le sexe, l'expérience professionnelle, le niveau académique, le permis de conduire et le niveau d'anglais. Dans la figure du tableau ci-dessous, la colonne "Context" correspond au contexte sexiste ou non sexiste (0 = contexte sexiste, 10 = contexte non sexiste), la colonne "N1" correspond quant à elle à la note éventuelle lorsque le système 1 est activé, c'est à dire, lorsque le recruteur attribue une note de façon automatique et la colonne "N2" correspond à la note du système 2 une fois activé, cette fois-ci le recruteur attribue une note de façon plus réfléchie, en considérant ainsi le contexte.

| | Sexe | Age | AnneeEtude | ExpPro | Permis | NivAnglais | Context | N1 | N2 |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 1 | 44 | 5 | 1 | 0 | 4 | 0 | 0 | 0 |
| 3 | 0 | 45 | 6 | 5 | 0 | 4 | 0 | 20 | 20 |
| 4 | 0 | 35 | 3 | 4 | 1 | 4 | 0 | 20 | 19 |
| 5 | 0 | 46 | 3 | 7 | 1 | 3 | 10 | 20 | 10 |
| 6 | 1 | 32 | 1 | 6 | 1 | 2 | 0 | 2 | 1 |
| 7 | 0 | 32 | 4 | 7 | 0 | 3 | 0 | 17 | 19 |
| 8 | 0 | 44 | 2 | 4 | 0 | 4 | 10 | 18 | 11 |
| 9 | 1 | 31 | 6 | 0 | 1 | 4 | 0 | 4 | 0 |
| 10 | 1 | 50 | 2 | 5 | 1 | 1 | 10 | 3 | 12 |
| 11 | 1 | 20 | 1 | 4 | 0 | 2 | 0 | 4 | 0 |
| 12 | 0 | 46 | 5 | 6 | 1 | 1 | 10 | 16 | 10 |
| 13 | 1 | 39 | 3 | 3 | 1 | 2 | 10 | 2 | 11 |
| 14 | 0 | 29 | 8 | 10 | 0 | 0 | 10 | 19 | 10 |
| 15 | 1 | 20 | 1 | 10 | 1 | 1 | 0 | 7 | 2 |
| 16 | 1 | 43 | 6 | 1 | 1 | 4 | 0 | 2 | 0 |
| 17 | 0 | 19 | 3 | 3 | 0 | 3 | 0 | 19 | 19 |
| 18 | 0 | 32 | 5 | 8 | 0 | 3 | 10 | 20 | 12 |
| 19 | 0 | 23 | 5 | 3 | 0 | 3 | 10 | 20 | 13 |
| 20 | 0 | 23 | 7 | 1 | 1 | 4 | 0 | 20 | 20 |

*Figure 4 : Tableau des 20 premières lignes de la base de données de l'algorithme*

## Matériel

Pour réaliser cela, la modélisation de l'algorithme se fait grâce au langage de programmation Python et sur l'environnement de développement intégré (IDE) appelé Spyder. Des librairies spécifiques sont utilisées telles que Scikit-Learn qui est une bibliothèque libre Python destinée à l'apprentissage automatique. Les librairies Keras et Tensorflow sont utilisées car elles permettent une expérimentation rapide avec les réseaux de neurones profonds (deep learning).

La structure de l'algorithme se construit comme suit (cf : *Figure 5*) : une entrée système 1 et système 2 composée de la base de données sans prise en compte du contexte pour le système 1, c'est pour cela qu'il y a 6 colonnes sélectionnées et qu'il y en a 7 pour le système 2.

La base de données système 1 passe par une couche Dense qui est une couche de réseau neuronal connectée en profondeur, ce qui signifie que chaque neurone de la couche dense reçoit des entrées de tous les neurones de la couche précédente. Dans cette même couche on retrouve une fonction d'activation qui transforme le signal d'entrée du neurone artificiel. La fonction d'activation est utile pour introduire une non-linéarité dans le réseau neuronal afin que le réseau puisse apprendre une relation complexe entre les données d'entrée et de sortie. Les fonctions d'activation utilisées dans le réseau sont les suivantes : fonction softmax, fonction tanh, et ReLU.

Il y a beaucoup plus de couches de neurones dans la partie système 2 car celui-ci représente le système lent, donc l'intégration de plusieurs couches de neurones implique un apprentissage beaucoup plus en profondeur correspondant ainsi à la prise en compte du contexte.

L'ajout d'une couche de neurone récurrent permet de prendre en compte les informations préalablement traitées dans le système 1 puis reprises par le système 2 permettant un apprentissage plus long permettant la prise en considération du contexte (cf. *Figure 5*).

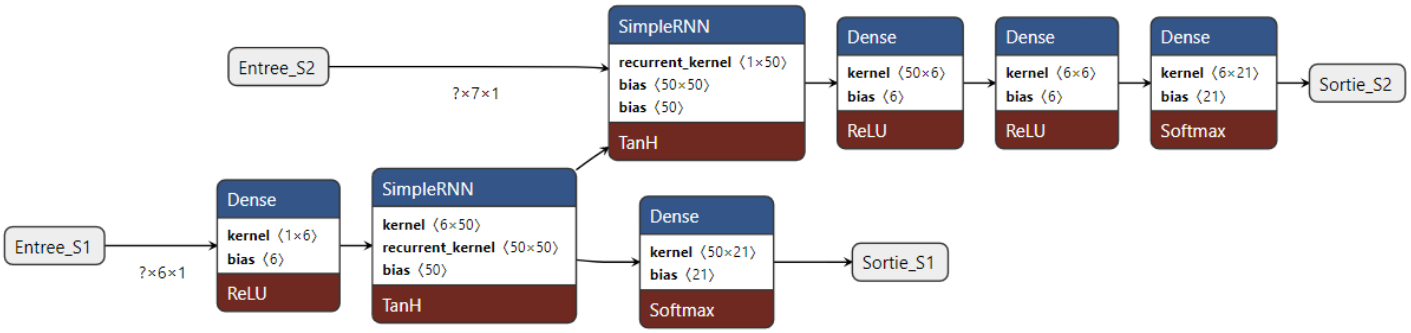



*Figure 5: Schéma de l'architecture du réseau de neurones.*

Nous avons donc adopté une approche essentiellement connexionniste pour modéliser la théorie du double processus. Néanmoins, nous avons été amenés à intégrer une partie symbolique/classique dans notre modèle pour simuler l'évaluation d'un CV spécifique afin de "forcer" le modèle à prendre en considération le contexte dans lequel le recruteur se trouve (cf. *Figure 6*). Notre modèle peut donc être considéré comme mixte: mélangeant conception connexionniste et symbolique/classique.

```
if contexte == 0 :
    strContexte = Fo
    print(" La note
else :
    strContexte = Fo
    print(" La note
```

*Figure 6 : partie du code classique/symbolique qui force le modèle à prendre en compte le contexte via une balise "if/else"*

## Hypothèses

Pour cette première expérience, les hypothèses sont les suivantes :

1. Le CV d'un homme est noté très favorablement lorsque le recruteur est dans un contexte *ingroup* favorisant l'expression d'un comportement sexiste (par exemple, dans une réunion où il n'y a que des hommes). Le CV d'une femme est noté très défavorablement lorsque le recruteur est dans un contexte *ingroup* favorisant l'expression d'un comportement sexiste. Le recruteur est engagé dans un processus de prise de décision heuristique, issue du système 1.
2. À l'inverse, dans un contexte *outgroup* ne favorisant pas un comportement sexiste, le recruteur note de façon équivalente le CV d'une femme et le CV d'un homme (par exemple, dans une réunion où il n'y a que des femmes). Le recruteur est engagé dans un processus de prise de décision analytique, issue du système 2.

## Résultats

La figure ci-dessous représente l'entraînement de l'algorithme. La performance (validation accuracy) du système 2 est meilleure que le système 1 est cela s'explique par le nombre de couches de neurones plus important pour le système 2. La performance du système 2 s'élève à 98,8% et la performance du système 1 s'élève à 93,6% (cf. *Figure 7*).

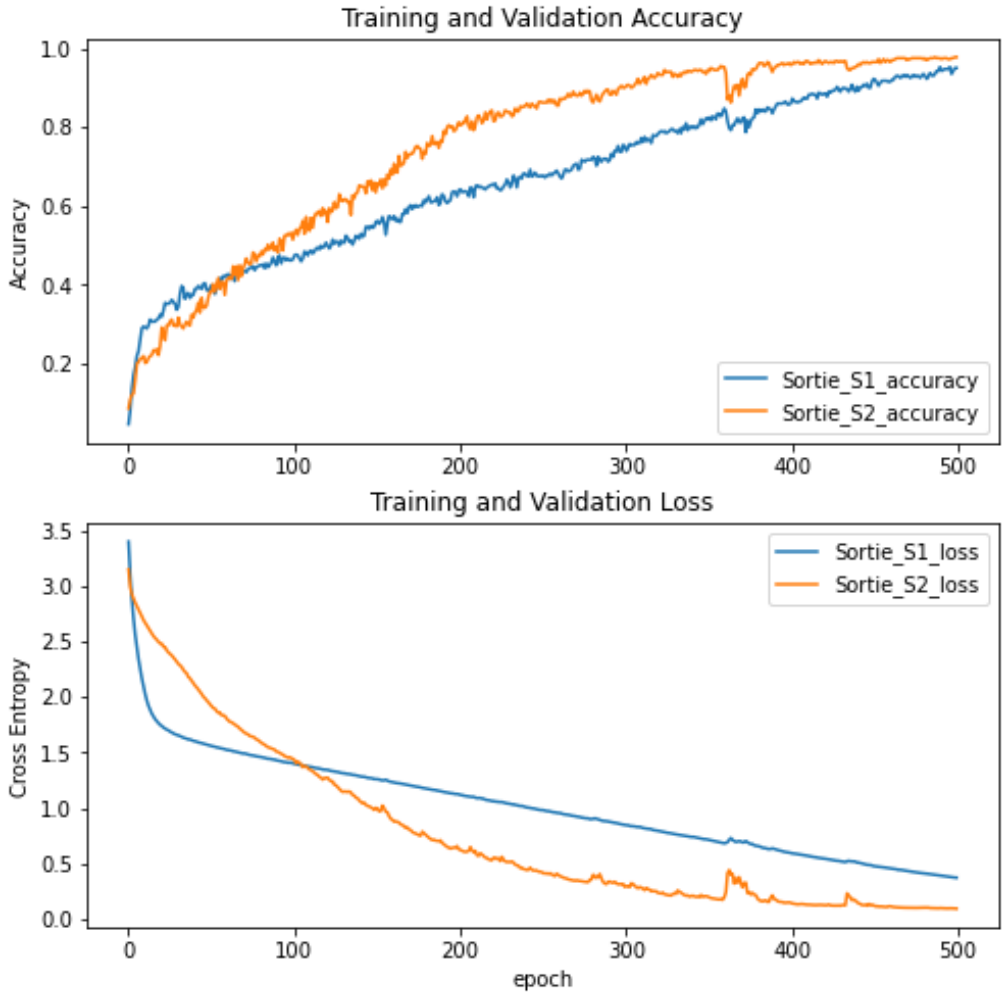
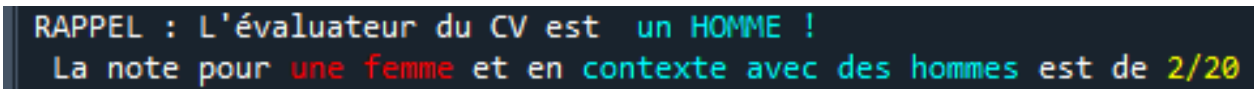



*Figure 7. Visualisation de l'entraînement de l'algorithme*

L'image ci-dessous représente la prédiction de l'entraînement de l'algorithme sur de nouvelles données. Dans cet exemple, le recruteur note le CV d'un homme et le résultat suit l'hypothèse qui stipule qu'en effet, lorsque le recruteur doit juger le CV d'un homme, il l'évalue de façon très positive, tandis que s'il doit noter le CV d'une femme, il l'évalue de façon négative. Il y a bien eu un apprentissage du stéréotype tout en suivant la théorie du double processus car lorsque le recruteur doit juger le CV d'une femme en tenant en compte le contexte, il l'évalue différemment.

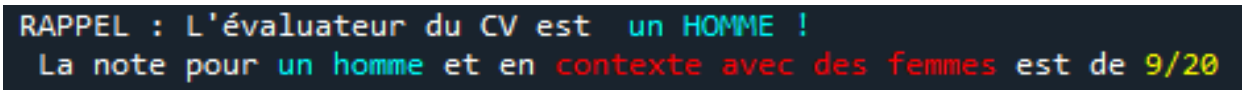
```
RAPPEL : L'évaluateur du CV est  un HOMME !
 La note pour un homme et en contexte avec des hommes est de 18/20
1/1 [==============================] - 0s 46ms/step
```

*(a) Schéma du résultat de l'hypothèse 1 : les hommes sont favorisés (contexte ingroup)*




*(b) Schéma du résultat de l'hypothèse 1 : les femmes sont défavorisés (contexte ingroup)*


*(c) Schéma du résultat de l'hypothèse 2 : les hommes ne sont pas favorisés (contexte outgroup)*




*(d) Schéma du résultat de l'hypothèse 2 : les femmes ne sont pas défavorisés (contexte outgroup)*

Nous avons aussi créé une méthode permettant de simuler le recrutement dans un contexte où il y aurait un pourcentage précis de femmes. Jusqu'ici, nous avions modélisé deux contextes : soit une réunion où il n'y aurait que des femmes, soit une réunion où il n'y aurait que des hommes. Nous avons donc entraîné le modèle avec un contexte plus nuancé pour observer le comportement du recruteur lorsqu'il est entouré de 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% et 90% de femmes. Les résultats de ces simulations montrent que plus le pourcentage de femmes dans la réunion est faible, plus le recruteur déploie un comportement sexiste lors d'évaluations de CV des femmes en les dévalorisant (cf. *Figure 8*). Lorsqu'il s'agit d'évaluer le CV d'un homme[1], plus le pourcentage femmes dans la réunion est faible, plus le recruteur déploie un comportement sexiste lors d'évaluations de CV des hommes en les favorisant (cf. *Figure 9*).

Ainsi, plus la proportion de femmes dans la réunion est petite, plus le recruteur déploie un comportement sexiste (et engage un processus décisionnel heuristique). Plus la proportion de femmes dans la réunion est grande, plus le recruteur déploie un comportement non sexiste (et engage un processus décisionnel analytique).

[1] Dans cette expérience, le CV de la femme jugée et de l'homme jugé étaient identiques (même âge, mêmes années d'études, …). Seul le sexe était modifié.

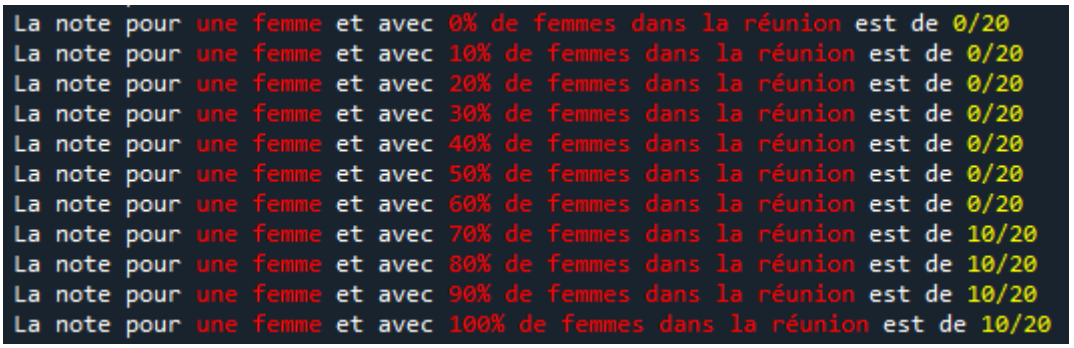



*Figure 8 : Simulation d'une évaluation d'un CV d'une femme lorsqu'il y a un pourcentage précis de femmes dans la réunion.*

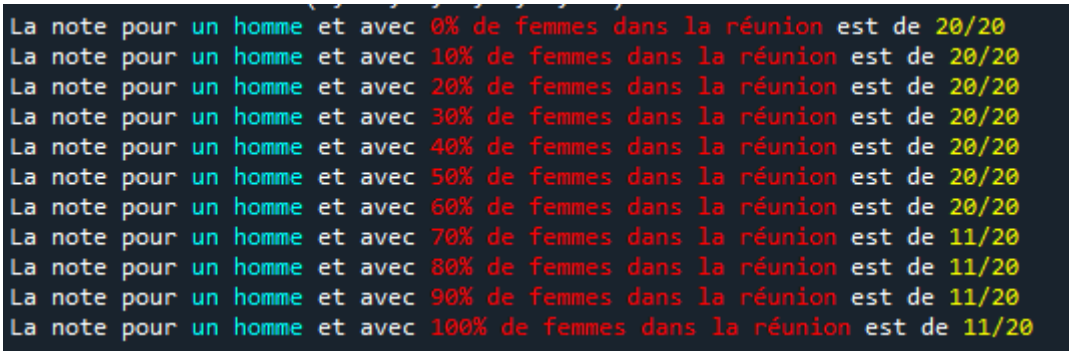



*Figure 9 : Simulation d'une évaluation d'un CV d'un homme lorsqu'il y a un pourcentage précis de femmes dans la réunion.*

## B. Expérience 2

### But de l'expérience

L'expérience 2 a pour but de tester l'hypothèse d'un cas prototypique où l'on imagine un recrutement dans une entreprise où il y a que très peu de femme. Pour cela, la base de données à été modifiée de sorte que 80% des données de femmes ont été retirées.

### Hypothèse

L'hypothèse est que cette fois, le recruteur évalue le CV de femmes de façon moins sexiste et ce en raison de la base de données biaisée. En effet, en étant moins confronté à

des expériences où les femmes sont discriminées à l'embauche, le recruteur aurait moins intégré ce type de préjugé et aurait donc déployé moins de comportements discriminatoires.

### Résultats

Les résultats démontrent qu'avec une base de données où le nombre de femmes est moindre, le modèle entraîné simule toujours un recruteur déployant un comportement sexiste seulement dans un contexte *ingroup* et un comportement non sexiste dans un contexte *outgroup*.

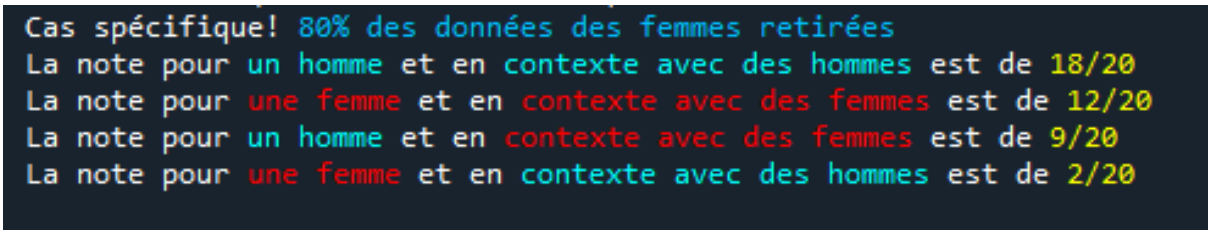



*Figure 10: cas où 80% des données des femmes ont été retirées pour entraîner le modèle.*

# Discussion

L'objectif de notre exercice était de simuler un individu doté d'un processus décisionnel comme conceptualisé par la théorie du double processus de Kahneman. Pour ce faire, nous avons modélisé un individu recruteur qui, étant confronté à des expériences où les femmes étaient discriminées dans un contexte favorable au déploiement de préjugés sexistes, les auraient intégrés et imités lors de ses propres évaluations. L'enjeu était alors de modéliser la théorie du double processus selon l'approche théorique de Kahneman dans un réseau de neurones artificiels. Il était aussi important d'établir une base de données qui met en évidence des comportements sexistes lors de l'évaluation de CV de femmes dans certains contextes habituellement favorables à l'expression de préjugés et au déploiement de discriminations sexistes. Pour ce faire, nous nous sommes appuyés sur les récentes recherches en psychologie sociale des stéréotypes et préjugés. Notre modélisation est ainsi capable de simuler un recruteur sexiste capable d'inhiber ses préjugés sexistes dans un contexte *outgroup*, entouré de femmes, les évaluant à l'égal des hommes et d'exprimer librement ses préjugés sexistes dans un context *ingroup*, entouré d'hommes, les évaluent plus négativement que les hommes sur la seule base de leur genre. Cette simulation fut

possible grâce à la modélisation de l'approche théorique de Kahneman sur les processus décisionnels, permettant au recruteur de déployer une prise de décision heuristique dans un contexte *ingroup* et analytique dans un contexte *outgroup*.

Néanmoins, plusieurs auteurs ont critiqué la pertinence et l'utilité de la théorie du double processus (Gigerenzer et Reiger, 1996 ; Keren et Schul, 2009 ; Kruglanski et Gigerenzer, 2011 ; Osman, 2004, cités par Keren, 2013). En 2009, Keren et Schul ont tenté de modifier la théorie du double processus afin de pouvoir la vérifier. Ils conclurent que la tâche était impossible à réaliser car la théorie du double processus est censée "rendre compte de presque tous les phénomènes socio-cognitifs de haut niveau" (Keren, 2013). En considérant cette approche théorique comme sous-jacente aux concepts des stéréotypes et des préjugés, nous avons modélisé celle-ci dans un réseau de neurones artificiels. Même si nos résultats sont satisfaisants et se traduisent par une bonne simulation d'un recruteur sexiste sachant se contrôler en fonction du contexte, nous aurions pu envisager d'autres approches théoriques. En effet, des études récentes explorent l'hypothèse de l'auto-congruence, qui permet d'expliciter les processus responsables du contrôle de soi lorsqu'un individu déploie, ou non, un préjugé conforme à un stéréotype connu (Smeding et al., 2016). La méthodologie déployée par ce genre d'étude, qui utilisent le mouse-tracking - un outil permettant de souligner les processus sous-jacents à la prise de décision (Freeman, 2011) - pourrait permettre une meilleure modélisation, car elle génère des données réellement révélatrice de la processualité de la prise de décision d'un individu déployant un comportement discriminatoire. Il serait ainsi intéressant que de futures modélisations s'appuient sur cette approche théorique et sur la méthodologie très efficace du mouse-tracking.

Globalement, cet exercice nous a surtout permis de simuler un comportement humain, la discrimination, ainsi que les processus responsables de celui-ci. C'est en modélisant une approche théorique en psychologie que nous avons pu simuler un tel comportement. Grâce à cet exercice nous démontrons qu'il est pertinent de prendre en considération les approches théoriques en psychologie pour simuler un comportement humain. Les sciences sociales peuvent ainsi apporter un éclairage nouveau dans l'intelligence artificielle.

# Bibliographie

Allport, G. W. (1935). Attitudes. In _A Handbook of Social Psychology_ (pp. 798–844). Clark University Press.

Allport, G. W. (1954). The nature of prejudice. Reading, MA: Addison-Wesley.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human Memory: A Proposed System and Its Control Processes. In K. W. Spence, & J. T. Spence (Eds.), The Psychology of Learning and Motivation: Advances in Research and Theory (Vol. 2, pp. 89-195). New York: Academic Press. http://dx.doi.org/10.1016/s0079-7421(08)60422-3

A. Augello, I. Infantino, A. Lieto, U. Maniscalco, G. Pilato, andF. Vella, “Towards a dual process approach to computational explanation in human-robot social interaction,” in CAID@ IJCAI, 2017.

Baddeley, A. (2007). _Working memory, thought, and action._ Oxford University Press.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. Journal of Personality and Social Psychology, 56, 5-18. doi:10.1037/0022-3514.56.1.5

Ehrlich, H. J. (1973). The social psychology of prejudice. New \brk: Wiley.

Evans, J. S. B. (2003). In two minds: dual-process accounts of reasoning. Trends in cognitive sciences, 7(10), 454-459.

Evans, J. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. Annual Review of Psychology, 59(1), 255-278.

Evans, J. S. B. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. Developmental Review, 31(2-3), 86-102.

Farrell, L., Cochrane, A., & McHugh, L. (2015). Exploring attitudes towards gender and science: The advantages of an IRAP approach versus the IAT. _Journal of Contextual Behavioral Science, 4_(2), 121–128.

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. Frontiers in psychology, 2, 59.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. _Journal of Personality and Social Psychology, 74_(6), 1464–1480.

Keren, G. (2013). A tale of two systems: A scientific advance or a theoretical stone soup? Commentary on Evans & Stanovich (2013). Perspectives on Psychological Science, 8(3), 257-262.

Oakes, P.J., Turner, J.C. and Haslam, S.A. (1991), Perceiving people as group members: The role of fit in the salience of social categorizations. British Journal of Social Psychology, 30: 125-144.

Harris LT, Fiske ST. Dehumanizing the Lowest of the Low: Neuroimaging Responses to Extreme Out-Groups. Psychological Science. 2006;17(10):847-853. doi:10.1111/j.1467-9280.2006.01793.x

Liberman, Z., Woodward, A. L., & Kinzler, K. D. (2017). The Origins of Social Categorization. Trends in cognitive sciences, 21(7), 556–568. https://doi.org/10.1016/j.tics.2017.04.004

Macrae, C. N., Milne, A. B., & Bodenhausen, G. V. (1994). Stereotypes as energy-saving devices: A peek inside the cognitive toolbox. Journal of Personality and Social Psychology, 66(1), 37–47.

Martin, C. L., & Halverson, C. F. (1981). Schematic Processing Model of Sex-Typing and Stereotyping in Children. Child Development, 52, 1119-1134. http://dx.doi.org/10.2307/1129498

Morewedge, C. K., & Kahneman, D. (2010). Associative processes in intuitive judgment. Trends in cognitive sciences, 14(10), 435-440.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. Psychological Review, 84, 127-190.

Tajfel, H. (1981). Human groups and social categories: Studies in social psychology. Cambridge, England: Cambridge University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. science, 185(4157), 1124-1131.

T. Anthony, Z. Tian, and D. Barber. 2017. Thinking fast and slow with deep learning and tree search. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 5366–5376.

Weisman, K., Johnson, M.V. and Shutts, K. (2015), Young children's automatic encoding of social categories. Dev Sci, 18: 1036-1043. (https://doi.org/10.1111/desc.12269)

X. Guo, S. Singh, H. Lee, R. L. Lewis, and X. Wang, “Deep learningfor real-time atari game play using offline monte-carlo tree searchplanning,” in Advances in neural information processing systems, 2014,pp. 3338–3346.

Smeding, A., Quinton, J. C., Lauer, K., Barca, L., & Pezzulo, G. (2016). Tracking and simulating dynamics of implicit stereotypes: A situated social cognition perspective. Journal of Personality and Social Psychology, 111(6), 817.

Strannegård, C., von Haugwitz, R., Wessberg, J., Balkenius, C. (2013). A Cognitive Architecture Based on Dual Process Theory. In: Kühnberger, KU., Rudolph, S., Wang, P. (eds) Artificial General Intelligence. AGI 2013. Lecture Notes in Computer Science(), vol 7999. Springer, Berlin, Heidelberg. doi: 10.1007/978-3-642-39521-5_15