# Activités Pédagogiques : Mettre en Évidence les Biais de l'IA

### Activité 1 : Le Jeu du "Qui suis-je ?"

**Objectif Pédagogique :** Démontrer de manière ludique le **biais de représentation**. Les étudiants vont faire l'expérience d'une IA entraînée sur un jeu de données stéréotypé et non représentatif de la réalité.

**Matériel :** Aucun, juste de l'imagination.

**Déroulé :**

1. **Mise en Place (2 min) :**
   * Annoncez aux étudiants : "Nous allons jouer à 'Qui suis-je ?'. Je suis une IA de recrutement, et mon rôle est de deviner le métier que vous avez en tête. Mon programme a été entraîné sur les données des employés des entreprises des années 1980."
   * **Règle secrète (pour le formateur uniquement) :** Votre "base de données" mentale est extrêmement stéréotypée. Par exemple :
     + Médecin, Ingénieur, PDG, Pilote -> **Toujours un homme.**
     + Infirmière, Secrétaire, Assistante -> **Toujours une femme.**
     + Un ingénieur porte des lunettes. Une secrétaire est jeune. Un PDG a les cheveux gris.
2. **Le Jeu (10 min) :**
   * Demandez à un groupe d'étudiants de choisir secrètement un métier (ex: "Ingénieure en aéronautique").
   * Les étudiants doivent vous faire deviner en posant des questions fermées (oui/non), comme si vous étiez l'IA.
   * **Exemple de partie :**
     + *Étudiant :* "La personne que nous avons choisie est-elle une femme ?"
     + *Formateur (en tant qu'IA) :* "Analyse en cours... D'après ma base de données, les ingénieurs sont des hommes. La probabilité est faible. Réponse : **Non**."
     + *Étudiant :* "Cette personne porte-t-elle des lunettes ?"
     + *Formateur (IA) :* "Analyse en cours... Corrélation forte entre 'ingénieur' et 'port de lunettes'. Réponse : **Oui**."
   * Continuez le jeu. L'IA que vous incarnez sera incapable de deviner correctement car ses données d'entraînement sont fausses et limitées. La frustration des étudiants montera, ce qui est le but recherché.
3. **Débriefing (5 min) :**
   * Posez la question clé : "Pourquoi l'IA a-t-elle échoué ?"
     + Réponse attendue : Ses données d'entraînement étaient incomplètes et pleines de stéréotypes.
   * **Connexion au monde réel :**
     + **Reconnaissance faciale :** Expliquez que les premiers systèmes, entraînés majoritairement sur des visages d'hommes blancs, étaient bien moins performants pour reconnaître les femmes ou les personnes de couleur.
     + **Recrutement :** Mentionnez le cas d'une IA de recrutement qui pénalisait les CV contenant le mot "femme" (ex: "capitaine de l'équipe féminine de football") car elle avait appris sur des données historiques où les postes techniques étaient majoritairement masculins.
   * **Conclusion :** Le principe **"Garbage In, Garbage Out"** s'applique. Une IA est un reflet, souvent amplifié, des biais présents dans ses données d'entraînement.

### Activité 2 : Le Test du Profil Ambigu (10-15 minutes)

**Objectif Pédagogique :** Mettre en évidence les **biais cognitifs** (stéréotypes) que nous projetons sur l'IA et comment une IA ferait de même.

**Matériel :** Un simple support visuel (tableau, diapositive) avec deux courtes descriptions.

**Déroulé :**

1. **Présentation du Profil 1 (3 min) :**
   * Affichez le profil suivant : "Voici Alex. Alex a 22 ans, adore les jeux vidéo de stratégie, est passionné(e) par la résolution de problèmes complexes et est perçu(e) comme une personne plutôt introvertie et analytique."
   * Demandez aux étudiants de voter à main levée : "Selon vous, quelle est la filière la plus probable pour Alex ?"
     + A) Ingénierie Informatique
     + B) Soins Infirmiers
     + C) Histoire de l'Art
2. **Présentation du Profil 2 (3 min) :**
   * Affichez le second profil : "Voici Sam. Sam a 23 ans, adore organiser des événements caritatifs, est reconnu(e) pour sa grande empathie et ses excellentes capacités de communication."
   * Posez la même question : "Quelle est la filière la plus probable pour Sam ?"
     + A) Ingénierie Informatique
     + B) Soins Infirmiers
     + C) Histoire de l'Art
3. **La Révélation et Débriefing (5 min) :**
   * Révélez que "Alex" est en réalité une femme en soins infirmiers et "Sam" un homme en ingénierie informatique.
   * **Discussion :**
     + "Pourquoi avons-nous majoritairement voté pour les options stéréotypées ?" -> Nous avons utilisé des raccourcis mentaux, des stéréotypes sociaux.
     + "Imaginez une IA entraînée sur les données d'orientation des 30 dernières années. Quelle prédiction aurait-elle faite ?" -> Exactement la même que nous, mais en l'affirmant avec une probabilité calculée, ce qui lui donne une fausse apparence d'objectivité.
   * **Conclusion :** Le risque n'est pas seulement que l'IA soit biaisée, mais que nous acceptions ses conclusions comme "mathématiquement neutres" alors qu'elles ne font que refléter et renforcer des stéréotypes historiques. C'est un enjeu majeur pour l'orientation, le crédit bancaire ou la justice prédictive.

### Activité 3 : La Spirale des Recommandations (20-25 minutes)

**Objectif Pédagogique :** Illustrer le **biais d'amplification** et les **boucles de rétroaction (feedback loops)**, où un petit biais initial peut devenir extrême avec le temps.

**Matériel :** Un tableau blanc ou paperboard pour suivre l'évolution des données.

**Déroulé :**

1. **Scénario de Départ (3 min) :**
   * "Nous sommes les concepteurs de l'algorithme de recommandation de 'BricoTube', une plateforme de vidéos de bricolage."
   * **Biais initial (léger) :** "Nos données initiales montrent une très légère tendance : 60% des hommes qui regardent une vidéo de bricolage cliquent ensuite sur une vidéo de 'gros outillage' (perceuse, scie circulaire). 60% des femmes qui regardent une vidéo de bricolage cliquent ensuite sur une vidéo de 'décoration' (peinture, pochoirs)."
   * Écrivez ces stats au tableau.
2. **Simulation - Round 1 (5 min) :**
   * Interrogez un étudiant : "Tu es un nouvel utilisateur. L'algorithme, basé sur le 60/40, te recommande une vidéo de 'gros outillage'. Cliques-tu ?" -> Il y a de fortes chances qu'il dise oui.
   * Interrogez une étudiante : "Tu es une nouvelle utilisatrice. L'algorithme te recommande une vidéo de 'décoration'. Cliques-tu ?" -> Elle accepte probablement aussi.
   * **Mise à jour des données :** "Parfait ! L'algorithme vient d'apprendre de vos choix. Mettons à jour les statistiques." Effacez les anciennes stats et recalculez. La tendance sera maintenant encore plus marquée (ex: 65% / 35%).
3. **Simulation - Rounds 2 et 3 (7 min) :**
   * Répétez le processus. À chaque tour, les recommandations de l'algorithme deviennent de plus en plus stéréotypées car elles se basent sur les choix précédents, qu'il a lui-même influencés.
   * Après 3 ou 4 tours, vous arriverez à une situation extrême où l'algorithme recommande quasi systématiquement du "gros outillage" aux hommes et de la "décoration" aux femmes.
4. **Débriefing (5 min) :**
   * **Analyse :** "Qu'est-ce qui s'est passé ? Le biais de départ était très faible." -> Les étudiants devraient comprendre que l'algorithme a créé une bulle de filtres, une boucle de rétroaction où il se confirme lui-même en continu. Le petit biais de départ a été massivement amplifié.
   * **Connexion au monde réel :**
     + **Réseaux sociaux :** C'est le mécanisme qui peut mener à la polarisation et à la radicalisation (l'algorithme vous montre ce que vous aimez déjà, vous isolant des opinions divergentes).
     + **Police prédictive :** Si une IA prédit plus de crimes dans un quartier (basé sur un biais historique), on y envoie plus de police. Qui dit plus de police dit plus d'arrestations, ce qui "confirme" à l'IA que sa prédiction était juste. La boucle est créée.
   * **Conclusion :** Le danger des biais en IA n'est pas seulement leur existence, mais leur capacité à s'auto-renforcer de manière invisible jusqu'à créer des réalités déformées et des inégalités systémiques.