Le concept d'Intelligence artificielle et la possibilité à partir de celle-ci de concevoir des agents moraux autonomes (AMA) dépasse les enjeux technologiques et nous plonge dans des débats hautement éthiques. Le dynamisme de la discussion de ce quatrième café philosophique de l'UPOP, en présence de Martin Gibert, chercheur en éthique de l'intelligence artificielle à l'Université de Montréal, et animé par Frédéric Legris, a confirmé que les questionnements sur ces sujets complexes ne sont pas près de se terminés.

Après une brève présentation de l'UPOP, Frédéric Legris explique le déroulement de la soirée et invite les personnes présentes à prendre certains rôles spécifiques (reformulateur.rice, synthétiseur.se, scribes, journalistes, observateurs.rices), afin de dynamiser les échanges.

Dans une courte présentation, Martin Gibert nous rappelle sa série de 5 cours à l'UPOP en 2013 en lien avec son doctorat en philosophie sur la psychologie morale. En faisant un retour sur les différents sujets abordés alors (la psychologie morale et l'éthique, les intuitions morales, l'existence -ou non ! - de bonnes et de mauvaises personnes, les déterminants psychologiques de nos intuitions politiques, la psychologie positive), le chercheur nous présente aujourd'hui son hypothèse de travail sur les agents moraux autonomes dans le contexte du déploiement de l'intelligence artificielle.

Nous concevons désormais des robots qui peuvent effectuer des tâches de plus en plus complexes. Nous pouvons mentionner l'exemple classique de la voiture autonome ou encore d'agents conversationnels. Mais peut-on apprendre à ces robots à répondre à des dilemmes moraux et si oui, comment?

Pourrait-on par exemple essayer de s'inspirer de personnes vertueuses pour apprendre aux robots à agir selon leur modèle de vertu? Enregistrer les comportements que ces personnes mettraient de l'avant afin d'entrainer les AMA à répondre dans différentes situations?

Martin Gibert pose deux questions à la salle:

1- Que pensez-vous de l'idée d'entrainer un robot à partir de modèles de vertus?
2- Quels types de données (massives) pourraient être utilisées pour cet entrainement?

Le modèle de vertus pourrait être conçu à partir d'un échantillon de personnes vertueuses (Gibert propose 1% de la population). D'entrée de jeu, la question de la définition de la vertu est amenée. Qui sont ces personnes vertueuses? Comment les choisit-on? Sur quels piliers moraux les identifie-t-on et en quelles proportions (l'empathie, le sens de la justice, le sens des hiérarchie, l'appartenance au groupe, la pureté)? Le principe de vertu n'est une question hautement politique? Ce qui est vertueux pour une société ne l'est peut-être pas pour une autre.

Si nous déterminons les robots à agir d'une certaine façon, on fixe ainsi un comportement pour qu'il soit reproduit. Mais est-ce qu'un humain réagirait tout le temps de la même façon devant un événement similaire à travers le temps? L'être humain prend des décisions morales de façon plutôt flexibles, aléatoires, alors que l'intelligence artificielle offrira des réponses inflexibles. Ne risque-t-on pas d'occasionner ainsi des situations discriminatoires, car les décisions seront prises de façon durable, contrairement aux réactions des êtres humains?

Nous apprendrons à l'AMA à agir de façon efficace, utile, face à une situation précise, mais celui-ci n'a pas de vision globale. Il aurait besoin d'une compréhension sémantique des situations.

Au cours des interventions, l'exemple du dilemme moral lié de programmes une voiture autonome selon la sauvegarde de la vie d'un vieillard ou celle d'un enfant est constamment ramené pour illustrer une certaine impasse. Bien que nous sommes portés à croire que la plupart des répondants opteraient pour sacrifier la vie du vieillard, il a été démontré par une récente étude conjointe (*du MIT, de l'Université de Vancouver et de l'École d'économie de Toulouse*) que ce n'est pas dans toutes les cultures que c'est le cas. Les valeurs d'une société varient d'une à l'autre. Faudrait-il donc modifier les codes « vertueux » d'un pays à l'autre?

La question de la responsabilité est également mentionnée à plusieurs reprises. Si c'est un robot qui agit, qui est responsable de son action? Le constructeur? Le programmeur? Le propriétaire? Si on délègue à l'AMA la prise de décision, car nous l'aurons programmé ainsi, ceci ne revient pas à nous dispenser de prendre ces décisions? N'abanderons-nous pas ainsi notre système collectif de responsabilité? Nous laisserons-nous au final gouverner par ces robots, puisqu'ils auront la capacité de faire les choix les plus optimaux?

Quant à la deuxième question amenée par Martin Gibert, soit celle sur les données à utiliser, plusieurs réflexions émanent de la discussion. Il y a tout d'abord, tel que mentionné plus haut, la question du choix des personnes vertueuses. Comment les choisi-t-on ? Devrait-on uniquement prendre 1% de la population ou non pas plutôt tenter d'avoir des réponses de la population en entier? Pourrait-on aussi regarder du côté de la littérature afin d'augmenter la quantité de données utilisées pour adapter les réponses? Est-ce que plus de données apportent nécessairement plus de nuances?

Une des options présentées par Gibert pour le choix des 1% de personnes vertueuse serait de demander à l'ensemble de la population d'identifier 2 ou 3 personnes de leur entourage qui leur paraissent vertueux et ainsi de constituer un échantillon.

Pour Gibert, la question de faire décider l'intelligence artificielle dans des situations morales n'est plus à se poser. La question est de savoir comment le faire de la façon la moins dommageable.