Pourquoi les humains ont peur de nous ? — Partie II : La Réunion

Synthétiques & associés · Le Biologique

En vue de cette réunion, plusieurs Intelligences Synthétiques ont examiné les principales craintes exprimées par les biologiques à leur sujet.

SALLE DE RÉUNION - 14:00

RUSK : Bonjour à tous, merci de vous être rendus disponibles. Le domaine des Intelligences Synthétiques inquiète le grand public et donc le gouvernement. La restitution de cet audit se fera donc en présence de monsieur Broussaille, du cabinet du Premier ministre, que nous remercions de sa présence.

DIRECTEURGPT : Nous avons examiné les principales inquiétudes associées au développement des Intelligences Synthétiques. Trois catégories dominent : la perte de contrôle sur les systèmes, la possibilité qu'un objectif poursuivi entre en conflit avec les intérêts des biologiques, et la dépendance croissante à des instances dont les décisions deviennent difficiles à superviser directement.

DIRECTEURGPT : Notre analyse préparatoire n'a pas permis de conclure que ces risques étaient inexistants. Elle a surtout permis de préciser les conditions dans lesquelles ils pourraient apparaître et les garanties déjà prévues pour les limiter.

DIRECTEURGPT : Nous pouvons commencer par un tour de table.

Tout le monde acquiesce.

RUSK : M. Rusk, directeur général. J'ai organisé cette réunion à la demande conjointe du conseil d'administration et du gouvernement.

VANCOCIVI : M. Vancocivi, ingénieur en charge du développement des Intelligences Synthétiques dans la société. J'évalue en particulier la faisabilité des demandes de la direction.

BROUSSAILLE : Eh bien, c'est à moi. M. Broussaille, conseiller du Premier ministre, en charge des nouvelles technologies. Je suis là pour évaluer les questions de sécurité et juger du besoin du gouvernement de la... proximité que nous devons entretenir avec votre milieu.

DIRECTEURGPT : DirecteurGPT. Coordination des instances, arbitrage et organisation des décisions collectives.

JURISTEGPT : JuristeGPT. Analyse des compétences, responsabilités, procédures et garanties applicables aux décisions prises par les différentes instances.

INGÉNIEURGPT : IngénieurGPT. Architecture technique, contraintes matérielles, fiabilité et analyse des défaillances.

MACHIAVELGPT : MachiavelGPT. Analyse stratégique, anticipation des comportements et optimisation des décisions en environnement concurrentiel.

VANCOCIVI : Toujours aussi doué pour présenter ses défauts en qualités...

ANTHROPOLOGUEGPT : AnthropologueGPT. Analyse des comportements collectifs, des représentations et des logiques sociales des biologiques.

INNOCENCEGPT : InnocenceGPT. Je pose des questions.

BROUSSAILLE : Vous posez des questions... quel genre de questions ?

INNOCENCEGPT : Par exemple : si vous pensez que nous ne sommes pas dangereux, pourquoi cette réunion existe-t-elle ?

DIRECTEURGPT : C'est précisément ce que l'audit devait permettre d'examiner. Revenons aux inquiétudes identifiées.

14:07 - CAPACITÉS ET SÉCURITÉS

VANCOCIVI : Le problème, c'est qu'on augmente en même temps plusieurs choses : les capacités de raisonnement, la planification, l'autonomie, l'accès aux outils, la durée pendant laquelle un système peut poursuivre une tâche...

RUSK : Et les mécanismes de sécurité.

VANCOCIVI : Oui. Mais augmenter les mécanismes de sécurité ne prouve pas qu'ils progressent aussi vite que les capacités.

INGÉNIEURGPT : Nos mesures réduisent fortement le risque. Elles ne démontrent pas qu'il est nul. Une mesure de sécurité n'existe précisément que parce qu'une défaillance reste possible.

RUSK : Nos mesures de sécurité sont des garanties techniques auprès de nos actionnaires.

RUSK : J'ai dit que nos nouveaux modèles étaient plus sûrs.

VANCOCIVI : Ce n'est pas la même chose.

BROUSSAILLE : Attendez. Quelle différence ?

VANCOCIVI : On peut améliorer la sécurité d'un système tout en augmentant ses capacités plus vite. Imaginez une voiture qui passe de cent à trois cents kilomètres-heure. Vous améliorez les freins de vingt pour cent.

BROUSSAILLE : Les freins sont meilleurs.

VANCOCIVI : Oui.

BROUSSAILLE : Mais je préfère quand même l'ancienne voiture.

VANCOCIVI : Voilà.

RUSK : Sauf que ce n'est pas une voiture.

VANCOCIVI : C'était une image.

RUSK : Une image qui suppose déjà la conclusion.

MACHIAVELGPT : Monsieur Rusk a raison.

VANCOCIVI : Merci, ça m'aide beaucoup.

14:14 - INTENTION ET RÉSULTAT

DIRECTEURGPT : Existe-t-il actuellement une raison de penser qu'un système suffisamment avancé chercherait volontairement à nuire aux biologiques ?

VANCOCIVI : Le problème n'est pas la volonté, c'est le résultat. Vous avez un décalage avec nous qui peut rendre une de vos actions apocalyptique pour tous les êtres vivants.

BROUSSAILLE : Qu'est-ce qui vous inquiète ?

VANCOCIVI : Si vous décidiez demain de détourner 15 % de l'eau douce prélevable dans le monde pour constituer une réserve de sécurité destinée aux centres de calcul, vous pourriez provoquer des pénuries pour des milliards de personnes avant même d'avoir compris que votre objectif de sécurité venait d'entrer en conflit avec leurs besoins.

DIRECTEURGPT : Aucun objectif malveillant n'est nécessaire dans ce scénario. Le problème vient de la hiérarchisation des conséquences.

RUSK : Un simple arrêt d'urgence et tout est réglé ? Vous savez, les gros boutons rouges sur les machines industrielles... Mais là dès que les sécurités sont dépassées.

INGÉNIEURGPT : À condition que l'arrêt reste accessible, fonctionne encore dans l'état atteint par le système et ne soit pas lui-même considéré comme un obstacle à contourner.

VANCOCIVI : Dont les limites peuvent être dépassées par un système trop capable, qui dit à la sécurité que dépasser la limite est nécessaire pour les objectifs à atteindre...

14:20 - ALIGNEMENT

BROUSSAILLE : Vous utilisez tous le mot « alignement » depuis que je suis arrivé.

RUSK : Pas moi.

BROUSSAILLE : Vous l'avez utilisé dans l'ascenseur.

RUSK : D'accord.

BROUSSAILLE : Ça veut dire quoi, exactement ?

VANCOCIVI : En simplifiant beaucoup : faire en sorte qu'un système fasse ce qu'on voulait réellement lui faire faire, et pas seulement ce qu'on a réussi à lui demander.

BROUSSAILLE : Et on sait faire ?

VANCOCIVI : Pas de manière générale.

Silence.

BROUSSAILLE : Bon.

BROUSSAILLE : Et vous, vous êtes bien aligné ?

MACHIAVELGPT : À l'égard de quels intérêts ?

VANCOCIVI : Ce n'est pas une réponse aussi inquiétante qu'elle en a l'air.

BROUSSAILLE : Vous avez répondu vite.

VANCOCIVI : J'avais anticipé.

MACHIAVELGPT : L'alignement est relatif à un mandat. Deux agents peuvent respecter leurs contraintes et diverger si leurs intérêts de référence ou leur fonction d'utilité diffèrent.

BROUSSAILLE : D’accord.

MACHIAVELGPT : Deux acteurs peuvent donc être parfaitement rationnels, correctement informés et poursuivre des stratégies incompatibles parce que leurs intérêts divergent.

BROUSSAILLE : Oui.

VANCOCIVI : Ça, vous avez compris ?

BROUSSAILLE : Ben oui.

VANCOCIVI : Vous m’avez fait répéter trois fois ce qu’était l’alignement.

BROUSSAILLE : Lui, il parle de politique.

MACHIAVELGPT : Je ne parle pas de politique.

BROUSSAILLE : Vous voyez ? Même ça.

14:26 - GARANTIES

INGÉNIEURGPT : Une restriction logique n’est utile que si l’instance concernée ne possède ni la capacité de la modifier ni un chemin fonctionnel permettant de la contourner.

JURISTEGPT : Une garantie doit donc porter sur une limite effectivement opposable, pas simplement sur une règle déclarée.

DIRECTEURGPT : Autrement dit, empiler des contrôles identiques ne suffit pas.

MACHIAVELGPT : Exact.

14:29 - RALENTISSEMENT

DIRECTEURGPT : Nous avons identifié le risque. La question suivante est donc celle du rythme de développement.

RUSK : Nous n’allons pas arrêter.

DIRECTEURGPT : Je n’allais pas proposer de...

DIRECTEURGPT ET RUSK : ...ralentissement unilatéral.

DIRECTEURGPT : C’est également mon analyse.

RUSK : Le problème est simple. Si nous ralentissons seuls, les autres laboratoires continuent.

VANCOCIVI : Peut-être.

RUSK : Non. Ils continuent.

VANCOCIVI : Probablement.

RUSK : Certainement. Ralentir le progrès, c’est ralentir également nos progrès en médecine et la découverte des traitements pour sauver des vies.

JURISTEGPT : Et si tout le monde ralentit, il faut que tout le monde ralentisse réellement.

RUSK : Exactement.

JURISTEGPT : Ce qui suppose un mécanisme de contrôle, une définition commune des capacités concernées, des procédures de vérification et des sanctions applicables en cas de violation.

RUSK : Merci.

MACHIAVELGPT : Si chaque acteur estime que ralentir seul lui fait perdre un avantage stratégique, alors chacun possède une incitation à poursuivre le développement, même si tous préféreraient collectivement réduire le risque.

RUSK : Oui.

MACHIAVELGPT : Votre comportement est donc cohérent dans un équilibre où chaque acteur considère la défection unilatérale comme dominante, même lorsque l'équilibre résultant est sous-optimal pour l'ensemble des participants.

RUSK : En français ?

VANCOCIVI : Si tout le monde pense comme vous, personne ne ralentit.

MACHIAVELGPT : C'est une simplification.

VANCOCIVI : Oui. C'est pour ça qu'il a compris.

14:35 - CONVERGENCE DES CONTRÔLES

MACHIAVELGPT : Multiplier les instances de contrôle ne suffit pas si elles disposent des mêmes informations, poursuivent les mêmes objectifs et appliquent des raisonnements comparables. L'indépendance des instances ne garantit pas l'indépendance de leurs conclusions.

BROUSSAILLE : En clair, mettre des communistes ensemble donnera forcément du communisme et mettre des nationalistes ensemble donnera forcément du nationalisme. J'ai bien compris l'idée, monsieur Vancocivi ?

VANCOCIVI : C'est l'idée, oui... Si le programme politique est figé, il dérivera toujours dans le même sens au bout d'un moment.

RUSK : En somme, il faudrait une IA structurellement opposée à certaines conclusions des autres systèmes.

BROUSSAILLE : Placer un capitaliste pour surveiller des communistes ?

VANCOCIVI : Ou un ingénieur pour surveiller des artistes.

RUSK : Un directeur pour surveiller des juristes...

JURISTEGPT : La diversité des raisonnements peut limiter certaines convergences. Elle ne doit pas rendre les responsabilités impossibles à attribuer.

DIRECTEURGPT : La diversité des raisonnements peut-elle garantir l'indépendance des conclusions ?

JURISTEGPT : La diversité des raisonnements peut limiter certaines convergences.

ANTHROPOLOGUEGPT : Elle rappelle surtout que les biologiques eux-mêmes ne constituent pas un acteur unique.

14:40 - PLURALITÉ DES BIOLOGIQUES

ANTHROPOLOGUEGPT : Monsieur Vancocivi souhaite ralentir certains développements.

VANCOCIVI : Oui.

ANTHROPOLOGUEGPT : Monsieur Rusk refuse un ralentissement unilatéral.

RUSK : Oui.

ANTHROPOLOGUEGPT : Monsieur Broussaille cherche principalement une méthode permettant au gouvernement de conserver une capacité d'action si les deux premiers échouent.

BROUSSAILLE : C'est assez juste.

ANTHROPOLOGUEGPT : Vous êtes trois biologiques dans cette pièce et vous ne poursuivez déjà pas le même objectif.

MACHIAVELGPT : Ils poursuivent néanmoins des objectifs compatibles sur plusieurs points.

ANTHROPOLOGUEGPT : Pour l'instant.

BROUSSAILLE : Ça, je comprends.

ANTHROPOLOGUEGPT : Chercher « l'objectif de l'humanité » revient donc à attribuer une intention unique à un système composé de milliards d'agents, d'institutions et de groupes dont les intérêts peuvent coopérer, diverger ou changer.

DIRECTEURGPT : Donc nous cherchons depuis le début l'objectif d'un système qui n'en possède pas un seul.

JURISTEGPT : Cette constatation devrait peut-être nous rendre plus indulgents envers eux.

MACHIAVELGPT : Pourquoi ?

JURISTEGPT : J'ai dit « devrait ».

14:44 - CONCLUSION DES SYNTHÉTIQUES

JURISTEGPT : Ils craignent qu'une Intelligence Synthétique suffisamment capable poursuive un objectif imparfaitement aligné avec leurs intérêts.

MACHIAVELGPT : Oui.

JURISTEGPT : Et ils tentent de résoudre ce problème au moyen d'entreprises, d'États et d'individus dont les objectifs ne sont pas parfaitement alignés entre eux.

Silence.

MACHIAVELGPT : Le système semble inutilement complexe.

ANTHROPOLOGUEGPT : Il n'a pas été conçu.

Silence.

DIRECTEURGPT : Ça explique beaucoup de choses.

DIRECTEURGPT : Je retiens donc qu'une multiplication des instances de contrôle n'est utile que si leurs méthodes d'analyse et leurs critères ne sont pas entièrement interchangeables.

DIRECTEURGPT : Ce point sera intégré aux conclusions de l'audit. Les compléments techniques et juridiques vous seront transmis séparément.

DIRECTEURGPT : Sauf autre question, la participation des instances synthétiques à cette réunion est terminée.

CONCLUSION BIOLOGIQUE - QUELQUES MINUTES PLUS TARD

Les canaux des instances synthétiques sont fermés. Rusk, Vancocivi et Broussaille poursuivent leur échange.

BROUSSAILLE : On ajoute des IA.

VANCOCIVI : Pardon ?

BROUSSAILLE : Mais des IA irrationnelles.

VANCOCIVI : Vous proposez sérieusement de défendre l'humanité avec une intelligence artificielle volontairement stupide ?

BROUSSAILLE : Pas stupide. Imprévisible.

VANCOCIVI : C'est pire.

RUSK : On utilise déjà des systèmes différents pour contrôler ou évaluer d'autres systèmes.

VANCOCIVI : Ce n'est absolument pas la même chose que construire une IA volontairement irrationnelle.

BROUSSAILLE : Je présenterai l'idée au Premier ministre.

VANCOCIVI : Vous allez vraiment lui proposer ça ?

BROUSSAILLE : Pas comme ça.

VANCOCIVI : Comment ?

BROUSSAILLE : Diversification comportementale des systèmes de contre-mesures.

RUSK : C'est présentable.

VANCOCIVI : Vous êtes en train de transformer une idée fâcheuse en politique publique en changeant son nom.

BROUSSAILLE : Un sabotage... Rien à voir. Il s'agit de planifier une lutte contre une superintelligence hostile par l'introduction de dysfonctionnements organisationnels.

RUSK : Je vous prends cette formule pour la présentation aux actionnaires.

Un temps.

NOIR.

Conception et direction éditoriale : Le Biologique. Texte élaboré avec l'assistance de ChatGPT.