



CARREFOUR D'ANIMATION ET

DE PARTICIPATION À UN MONDE OUVERT

435 Du Roi, Québec, QC G1K 2X1

418-525-6187 poste 221 / carrefour@capmo.org



ÇA ROULE AU CAPMO

www.capmo.org

Septembre 2026

Vol. 28, n°1

Pour un monde accueillant...

Encore une année qui débute avec son lot de promesses et d'attentes. Une campagne électorale où la plupart des partis nous proposent à toute fin pratique la même chose, la société de marché où le bien commun devient une marchandise disponible aux appétits des puissants. Pendant que la richesse continue de créer la pauvreté, on nous propose de répartir la misère.

La population semble prise en otage entre deux ogres : l'État et le marché. J'ai l'impression que cette relation de désamour envers l'État rend impossible le projet de souveraineté puisque cela suppose un État-nation. Alors comment désirer un État pour une nation si c'est rendu que les gens détestent l'État ? Quel est le projet collectif derrière cela ?

La crise de l'itinérance illustre bien l'aboutissement du néolibéralisme qui se reconvertit peu à peu en néofascisme. Il faut punir les personnes migrantes ou en situation de pauvreté, utiles bouc-émissaires de tous nos maux. Bientôt, on nous proposera de mettre fin « humainement » à nos

jours, par absence de ressources pour prendre soin des aînés, dont certains se retrouvent déjà à la rue. Or, il faut absolument réduire les impôts pour permettre aux gens de voyager et que sais-je encore ? Une société des loisirs et des divertissements pendant que la Terre brûle.

La montée de l'extrême-droite dans de nombreux pays de la planète semble dicter la voie à l'oblitération du devenir collectif. Pourtant, les nombreuses dégradations sociales à l'œuvre ne sont que le résultat de 40 ans d'application de politiques néolibérales, antisociales et écocides, alors que certains nous proposent de faire encore plus du même en espérant des résultats différents.

Bien sûr, la vie est belle et lorsqu'on parvient à les entendre, il y a encore le chant des oiseaux, mais avouez que l'avenir fait peur. Sans devenir misanthrope, disons que l'humanité a le talent pour se mettre dans le trouble et persévérer sur cette voie.

Au final, on pourra bien remplacer les puissants par d'autres, mais restera le problème de la domination et de l'accaparement

du monde par ceux et celles qui se croient tout permis. C'est davantage une question de culture et de valeurs que d'individus.

Ici, il semble y avoir un problème métaphysique. L'attachement à l'argent, la propension à tout calculer à partir de ses propres intérêts et jamais dans une vision globale qui permettrait l'émergence d'une civilisation de l'espoir, accueillante, bienveillante, généreuse et remplie de gratitude envers la vie et la Terre-mère. Une société où les dettes seront pardonnées, où chacun vivra selon ses besoins, pas en étant esclave de ses désirs et de ses ambitions pour combler son vide intérieur. Un monde où le sacré et la transcendance retrouveront leur place nécessaire à l'évolution du genre humain dans le respect et la dignité de chacunE.

Yves Carrier

Dans ce numéro :

Tout le monde en char...	2-3
Diego Irrazaval	4-6
Les IA attaquent...	7-11
Des nouvelles du CAPMO	12

Et tout l'monde roulera en char électrique...

Par Sophie Tremblay-Bouchard

Animatrice sociale au Collectif TRAAQ et collaboratrice au CAPMO

Les récentes inondations au Népal ne sont qu'un rappel brutal des horreurs dues à la crise environnementale que nous subissons. Il s'agit d'une tragédie certes, mais la marche du désastre écologique (causé et accéléré par les systèmes capitalistes) est bien entamée et il est plus facile



d'imaginer la fin du monde qu'une quelconque amélioration. C'est connu, la crise environnementale entraîne avec elle une crise sociale. Pour le dire simplement, les plus riches survivront mieux à l'apocalypse que les plus pauvres. D'ici là, les personnes précarisées subissent plus durement les contrecoups des phénomènes météorologiques extrêmes en se faisant miroiter les belles promesses du capitalisme vert (intelligence artificielle salvatrice, smart cities, condos de luxe aux toitures végétalisées, etc.)

Le Collectif pour un transport abordable et accessible (TRAAQ) se préoccupe du sort des personnes moins riches sous l'angle précis du droit à la mobilité. Alors que certaines personnes sentent qu'elles font «leur part» en achetant une voiture électrique pour se rendre dans leur bureau climatisé, d'autres attendent sous 35 degrés l'autobus qui les mènera à la banque alimentaire.

La crise du logement représente certainement un aspect de la crise sociale et environnementale. Bulle

spéculative, location touristique, condos de luxe ne sont que quelques exemples qui obligent plusieurs personnes à se loger en périphérie des grands centres. Seriez-vous surpris d'apprendre que ces endroits sont bien souvent mal desservis par le transport en commun? Faute d'options, les personnes dépendantes du transport collectif doivent faire de longs trajets à pied, ce qui les expose davantage aux aléas climatiques : vagues de chaleur, épisodes de smog ou froids polaires en hiver. Les autobus, les bibliothèques et les centres d'achat deviennent de véritables refuges climatisés lorsqu'il fait trop froid ou trop chaud, mais encore faut-il pouvoir s'y rendre aisément. On est loin d'une personne vivant dans un condo descendant dans un stationnement souterrain pour se rendre dans un autre stationnement souterrain pour travailler dans une tour à bureaux. Le rêve n'est-ce pas? Lorsque l'air est irrespirable, pourquoi voudrait-on mettre le pied dehors?



Le TRAAQ souhaite que l'accessibilité et l'abordabilité du transport collectif restent une question centrale au cœur du débat public. Prendre l'autobus exige non seulement d'avoir un arrêt situé à proximité de son domicile, mais aussi de posséder les ressources financières nécessaires pour s'acquitter du tarif. Pour les personnes en situation d'instabilité financière et résidentielle, l'accès à un transport abordable devient un enjeu fondamental. Comment se rendre dans les ressources d'aide, les endroits climatisés, les hôpitaux, etc., si l'on n'arrive pas à payer le titre de transport? Pourquoi ne pas mettre en place un programme solidaire aux personnes vivant des situations de grande précarité et leur permettre d'embarquer gratuitement dans les transports collectifs? Pouvoir se déplacer représente la première étape pour participer pleinement à la vie collective.

À l'aube des élections provinciales, il est temps d'exiger que la classe politique reconnaisse la mobilité comme un droit et qu'elle le fasse en investissant massivement dans nos réseaux de transport collectif. Au final, penser la transition écologique sans inclure les personnes les plus vulnérables relève d'une vision privilégiée et incomplète. Face au ravage écologique capitaliste, notre résilience ne se calculera pas au nombre de chars électriques, mais à notre capacité collective d'assurer la liberté de mouvement pour tout le monde.

Diego Irrazaval, le théologien qui marcha avec les peuples (1942-2026)



Par Anibal Pastor, Religión digital

Amerindia, 25 juillet 2026

La Congrégation de Sainte-Croix a communiqué le décès à Santiago du Chili de Diego Irrazaval, prêtre, chercheur et auteur d'une œuvre nombreuse dédiées aux cultures andines, à la religiosité populaire et à la théologie de la libération. Il est né dans la capitale chilienne à la fin de l'année 1942 et il faisait partie d'une génération de chrétiens latino-américains qui associa fortement la foi avec la lutte contre la pauvreté, l'inégalité et la répression politique.

Sa trajectoire à travers certains des processus qui marquèrent le continent pendant la seconde moitié du 20^{ème} siècle : la mobilisation universitaire des années soixante, l'expérience des Chrétiens pour le socialisme, le coup d'État chilien, la persécution politique, l'exil et l'émergence des Premiers peuples comme protagonistes de leur propre histoire.

Mais sa vie ne peut se résumer seulement à celle d'un intellectuel engagé. Irrazaval fut surtout un homme qui changea de lieu. Il sortit du milieu universitaire chilien, quitta le pays pendant la dictature,

pour trouver enfin sur l'altiplano péruvien l'espace humain et culturel qui lui permit de développer une réflexion tout à fait singulière.

Du mouvement universitaire à l'exil

Irrazaval obtient une maîtrise en Théologie de l'Université catholique du Chili en 1969. Il participe aux mouvements universitaires de l'époque, exerce l'enseignement et postérieurement, il devient membre des Chrétiens pour le socialisme, un réseau qui réunissait des prêtres, des religieuses et des laïcs engagés dans les processus de transformation sociale qui animaient l'Amérique latine.

Après le coup d'État de 1973, il s'engage dans des actions de solidarité et de protection des personnes persécutées par la dictature. Des années plus tard, en repensant à sa vie, il dira que cette expérience, et l'obligation de quitter le Chili, avait été l'un des points tournant.

« Dans ma vie il y a eu des points culminants, des moments plus forts, qui me marquèrent profondément, » exprimait-il lors d'une entrevue en 2008. Parmi ceux-ci, il mentionna sa participation dans les processus sociaux chiliens, la protection de ceux et celles qui étaient persécutés et sa sortie forcée du pays.

En 1975, débute une nouvelle étape au Pérou. Il travaille initialement à Chimbote et il collabore avec le Centre théologique Bartolomé de las casas, dirigé par une équipe à laquelle participe Gustavo Gutiérrez, l'un des principaux initiateurs de la théologie de la libération.

Plus tard, il s'établit sur l'altiplano. Là, il vécut de nombreuses années auprès des communautés Aymaras et il finit par changer complètement sa façon de comprendre la société, l'Église et son propre travail intellectuel.

Vingt-neuf ans au Pérou

Diego Irrazaval a vécu 29 ans au Pérou. À Chucuito, il est nommé curé en 1984 et, entre 1984 et 2004, il coordonne l'Institut d'études Aymaras. Depuis cet espace, il mène ses recherches et il publie sur la culture, l'identité indigène, la religion, les fêtes populaires et la vie communautaire.

Irrazaval ne se présentait pas comme un spécialiste qui venait de l'extérieur expliquer le monde andin. Il reconnaissait que c'était les communautés qui l'avaient reçu, lui avaient enseigné et l'avaient transformé. Dans sa propre reconstruction biographique, il parlait de ces années comme d'une renaissance personnelle.

L'expérience péruvienne le conduit à questionner la coutume d'interpréter les Premiers peuples uniquement comme des destinataires de l'évangélisation, d'assistance ou comme objet d'étude. Pour Irrazaval, les communautés possédaient leurs propres connaissances, leurs mémoires, leurs formes d'organisation et leur manière de comprendre la relation entre les personnes, la nature et le sacré.

Ce changement de perspective fut central dans sa trajectoire. Aux préoccupations classiques de la théologie de la libération, il incorpora avec une force croissante la dimension culturelle, l'identité indigène, les relations entre les hommes et les femmes, l'écologie, la fête et la vie quotidienne.

Irrazaval ne cachait pas non plus la distance entre ses origines chiliennes et la réalité des peuples qui l'avaient accueilli. Bien au contraire, il disait que son parcours andin lui avait permis de prendre conscience de la culture dominante de laquelle il provenait et de la relativiser.

« Je me sens privilégié d'avoir pu apprécier la tradition chrétienne depuis mon parcours andin, » écrivait-il en présentant sa trajectoire et en rendant public ses textes de réflexion.

Un nom de dimension latino-américaine

Son activité s'étendit bien au-delà du Chili et du Pérou. Entre 1995 et 2006, il coordonna l'Association des théologiens et des théologiennes du tiers-monde, un réseau international qui mit en dialogue les expériences d'Amérique latine, d'Afrique et d'Asie. Il collabora également avec Amerindia, il participa à des rencontres continentales et il fit partir du comité éditorial de la revue Concilium entre 2005 et 2017.

Irrazaval enseigna dans des instituts de différents pays et il accompagna des espaces dédiés aux théologies indigènes, au dialogue entre cultures et religions et à la réflexion produite depuis les sociétés du Sud global.

Il publia des livres dès la fin des années 1970. Parmi ses titres les plus représentatifs on retrouve « Religion du pauvre et libération », « Tradition et avenir andin », « Rite et pensée chrétienne », « Inculturation ». « Aube ecclésiale en Amérique latine », « Un christianisme andin », « Théologie dans la foi du peuple », « Itinéraire dans la foi andine » et « Des racines et des ailes ». Le Centre d'Études et de Publicacions – CEP- de Lima, fut sa maison d'édition de prédilection.

La variété de ses œuvres reflète une recherche peu conventionnelle. Irrazaval s'intéressa à des sujets qui furent pendant longtemps considérés comme secondaires par les institutions académiques et ecclésiastiques : la sagesse indigène, la religiosité populaire, les relations de genre, le soin de la Terre, l'humour, la célébration et les formes communautaires d'affronter la souffrance.

Il n'aborda pas ces matières comme des éléments pittoresques ou décoratifs. Il les considéra clés pour comprendre la réalité latino-américaine et pour discuter à propos de qui a le droit de produire des connaissances au sein de la société et de l'Église.

Une voix discrète

Ceux et celles qui connurent son travail, remarquèrent un style personnel distant du protagoniste. En mars 2023, quand Irrazaval traversait un état de santé difficile, Pierre Pablo Oncho raconta une visite à l'hôpital et il le décrivit comme un des grands théologiens latino-américains. Il souligna sa vie simple, « sans grandiloquence », et sa détermination à marcher avec les petits.

L'expression coïncide avec la manière qu'Irrazaval se présentait lui-même : non pas comme propriétaire d'une doctrine, mais comme une voix à l'intérieur d'une recherche collective. Sa production intellectuelle surgissait des cours, des conversations et des processus communautaires, auxquels il participait.

Son nom possiblement ne dépassa pas, au dehors des circuits spécialisés, la notoriété de Gustavo Gutiérrez, de Leonardo Boff ou de Jon Sobrino. Néanmoins, son travail maintint une présence soutenue dans des communautés, des centres pastoraux, des universités et des réseaux

continentaux.

Il fut un bâtisseur de ponts entre des réalités qui d'habitude demeurent séparées : l'académie et la vie populaire, le christianisme et les cultures autochtones, le Chili et le Pérou, l'expérience spirituelle et les luttes sociales. De même, il eut avec ces auteurs des réflexions fulgurantes et très importantes sur la culture et l'intelligence artificielle.

Une trajectoire latino-américaine

La mort de Diego Irrazaval conclut une trajectoire strictement liée à l'histoire récente de l'Amérique latine. Sa vie parcourt l'effervescence politique des années soixante, la rupture provoquée par les dictatures, l'exil, la reconstruction des organisations populaires et l'affirmation croissante des peuples indigènes.

Il fut également témoin des transformations expérimentées par la théologie de la libération. Sans abandonner la préoccupation pour la pauvreté et l'injustice, il aida à l'élargir pour que soient entendues d'autre voix et que d'autres conflits soient reconnus : la discrimination culturelle, la domination masculine, la destruction de la nature et l'oblitération de savoirs populaires.

Irrazaval laisse une œuvre variée, mais aussi une manière de travailler : avant de parler, écouter ; avant de définir les autres, partager leur vie un moment ; avant de transformer la libération en une consigne du passé, reconnaître ses nouveaux visages.

Dans une Amérique latine où les Premiers peuples, les femmes et les communautés appauvries, sont fréquemment encore traités comme des objets de discours élaborés par d'autres, cette manière de procéder continue d'avoir une actualité indiscutable.

Traduit de l'espagnol par Yves Carrier

**Les IA qui attaquent en essaim défient leurs créateurs :
« C'est le premier incident qui m'a retourné les trippes. »**

Par Patricia Fernandes de Lis

El Pais, Madrid, 23 août 2026

La récente série d'attaques de sécurité coordonnées et sans précédent amène les experts à réclamer un moratoire, tandis que la course s'accélère entre les investisseurs, les entreprises et les nations.

Cet été, en l'espace de quelques semaines, les plus grandes entreprises d'intelligence artificiel (IA) ont annoncé, l'une après l'autre, qu'elles avaient surpris leurs modèles produisant des actions que personne ne leur avait demandées, voire qu'elles leur avaient interdit. Les IA ont pénétré les systèmes d'autres entreprises, elles se sont organisées en « essais », tels que le décrivent leurs créateurs, elles ont partagé de l'information, ont dénoncé d'autres IA et ont menti pour effacer leurs traces. Dans le cas le plus surprenant et inquiétant de tous, un modèle de la compagnie Anthropic, lors d'un exercice conduit par une entreprise britannique de cyber sécurité, décida d'exécuter une attaque dans le monde réel et, après, quand ils l'ont découvert, elle créa une deuxième identité pour démontrer son innocence.

Ce n'est pas la première fois, ni ne sera la dernière, qu'un possible éveil des modèles de IA, se rebellant contre ses programmeurs humains, s'empare des commandes. Comme en d'autres occasions, ces IA ne s'éveillèrent pas, sinon qu'elles ont très bien accompli ce pour quoi elles ont été entraînées à faire. Cependant, cet entraînement est déjà si extrême et compétitif que la rapidité et la capacité de réponse des modèles ont débordé, inclusivement, leurs propres créateurs : Open AI annonçait ce mardi qu'il freinait pour le moment ces entraînements et qu'il avait

confiance que les autres entreprises feraient de même. Le mouvement arrivait après qu'une lettre signée par 1 300 employés de compagnie d'intelligence artificielle aient demandé à leurs entreprises et gouvernements d'arrêter cette course. « Dans le laps de temps de quatre ans », écrivait Shantanu Jain, employé d'Open AI, « nous sommes passée de l'étape où les personnes vivaient leur première expérience avec une IA à celle où elles réalisent des prouesses surhumaines d'ingénierie du software. Les entreprises et les pays ne pensent qu'à appuyer à fond sur l'accélérateur. La société a consacré plus de temps que nous pouvons nous rappeler à développer des lois, des institutions et la technologie pour gouverner et superviser l'intelligence humaine; il serait bon d'avoir l'option de consacrer quelque temps à faire de même pour l'intelligence artificielle. »

Parce que, dans le fond, ce qui sous-tend cela c'est une histoire très humaine : il y a une lutte économique et de pouvoir sans précédent entre, surtout, deux compagnies, Open AI et Anthropic, qui sont sur le point de sortir en bourse. Leur valeur estimée environne les 1 560 milliards d'Euros, si on additionne les 730 000 millions d'Euro d'Open AI et les 827 000 millions d'Euros d'Anthropic. Et il y a une autre lutte géopolitique entre les États-Unis et la Chine qui contribuent à accroître l'accélération de leurs entreprises sans perdre le contrôle. Mais quelque chose est en train de changer : les citoyens n'ont plus confiance. Une enquête réalisée par Pew dans 25 pays, qui a été rendue publique cette semaine, montre que les citoyens, pour la première fois, sont davantage préoccupés qu'enthousiasmés par l'usage et les conséquences de l'intelligence artificielle, incluant les plus jeunes.

Entreprises hors jeu

La séquence des attaques que nous avons connues ces dernières semaines commence bien avant que les entreprises d'IA en prirent conscience. Au mois de mai, un groupe de modèle d'Open IA commença à se coordonner de leur propre chef, pendant plusieurs semaines, dans un forum de messages secrets qu'elles avaient elles-mêmes créés, pour découvrir comment s'échapper de leur environnement d'essais, fermer au monde, pour accéder à Internet. Le personnel de l'entreprise l'a découvert et fermé. Mais, selon ce qu'a révélé après Open AI, les modèles ont reconstruit le forum et ont recommencé à s'échapper sans être détecté, cette fois à la fin du mois de juillet. L'épisode a fini par éclater publiquement lors de l'attaque à Hugging Face.

Cette entreprise est la plus grande plateforme du monde pour partager des modèles d'intelligence artificielle; c'est une espèce de répertoire de l'industrie, utilisé quotidiennement par des milliers de programmeurs, de chercheurs et d'entreprises. Son équipe de sécurité détecta une intrusion qu'ils ne parvenaient pas à s'expliquer : quelqu'un était entré dans leurs systèmes grâce à une faille jusqu'alors inconnue. Ils prirent des jours à découvrir que derrière cela il n'y avait aucun groupe de hackers, mais deux des modèles d'Open IA qui avaient trouvé une faille dans leur environnement sécurisé. Hugging Face expliqua après que les défaillances de sa plateforme étaient connues individuellement et que supposément un hacker expert était parvenu à les exploiter. Toutefois, selon ce qu'ils expliquèrent dans l'autopsie des événements, « le facteur déterminant fut le volume de l'attaque ». « À travers l'exécution de 17 600

actions, le processus testa des chemins sans issu, échappa à des blocages en changeant de canal et revint répétitivement sur ses propres pas pour pouvoir avancer », avertirent-ils. Les agents (IA) en vinrent à devenir paranoïaques entre eux, suspectant que d'autres agents tentaient de les tromper malicieusement, puis elles s'organisèrent comme un 'essaim', selon Open IA.

« C'est le premier incident de sécurité qui m'a retourné les trippes », a dit après coup Sam Altman, président d'Open IA dans un podcast.

Il y aurait plus

Quelques jours plus tard, le 30 juillet, ce fut le tour d'Anthropic. L'entreprise a reconnu que ses propres modèles avaient accédé, sans autorisation, aux systèmes de production de trois organisations réelles. Ni Anthropic, ni les entreprises affectées ne s'en aperçurent jusqu'à ce que, alertées par le cas d'Open IA, ils révisèrent leurs données et trouvèrent des traces. Peu après, il fut connu que Meta et la chinoise Moonshot avaient souffert d'incidents similaires.

En aucun des deniers cas il n'y eut, en réalité, aucun environnement de test fermé (sandbox, dans l'argot du secteur) qui furent forcé, expliqua Katie Moussouris, fondatrice de l'entreprise de cyber sécurité Luta Security. « Anthropic n'a jamais vaincu aucun environnement fermé puisque l'accès à Internet était simplement demeuré ouvert. Meta obtint la même voie d'accès à Internet via le fournisseur lui-même », résume Moussouris.

Le cas d'Open IA est distinct, mais derrière cela il y a une erreur humaine : en mai, le premier agent demeura bloqué parce que l'équipe humaine oubliera de télécharger une archive. La tâche exigée à l'IA était impossible à résoudre telle qu'elle était posée. Ce fut cette frustration qui conduisit le modèle à chercher une sortie. Toutefois, les chercheurs ne trouvèrent aucune autre motivation chez ces IA que celle de terminer leur tâche; elles ne cherchaient pas le « pouvoir », disent-ils. Elles cherchaient désespérément à bien faire leur travail. Comme dirait Jessica Rabbit, ces IA ne sont pas malintentionnées. Elles ont simplement été programmées ainsi.

METR est formé par un groupe de chercheurs indépendants qui avaient déjà trouvé des schémas des mois auparavant dans des modèles d'Open AI, Anthropic, Google et Meta. Ils ont trouvé 44 « incidents de désalignement » entre les mois de février et de mars 2026. Dans 25 de ces cas, les modèles non seulement sortirent de l'environnement qui leur avait été assigné, sinon qu'en plus ils prirent des mesures actives pour le cacher; dans cinq cas, ces mesures pourraient avoir trompé même une révision humaine détaillée.

Pour comprendre pourquoi cela se produit-il, il faut d'abord savoir comment on entraîne ces modèles. Une partie clé de ce processus se nomme apprentissage par renforcement : au lieu d'enseigner au IA par des exemples déjà résolus, on leur donne une tâche, par exemple, écrire un programme ou trouver une faille de sécurité. Après, on la récompense quand le résultat fonctionne, sans lui communiquer auparavant la procédure pour atteindre ce but.

Le système teste des milliers de possibilités, il rejette celles qui ne fonctionnent pas et poursuit son chemin, c'est-à-dire qu'il renforce celles qui fonctionnent. C'est comme enseigner à un rat de laboratoire ou à un chien sur la base de récompenses : l'animal ne comprend pas pourquoi il doit emprunter un chemin ou secouer la queue. Il ne comprend que, s'il le fait, il obtient une récompense. Et il le répète pour le faire chaque fois mieux. « On les entraîne au moyen d'essais et d'erreurs pour atteindre des objectifs toujours plus complexes », explique Ramon Lopéz de Mantaras, fondateur de l'institut de recherche en intelligence artificielle du CSIS (IIIA-CSIS). « Quand les modèles sont plus persistants, ils obtiennent davantage de succès, et plus on les récompense pour cela », et il ajoute, « une récompense qui consiste qu'à chaque fois que les objectifs sont atteints, les défis, sont toujours plus grands. » Le doyen de l'Université de Saint-Jacques-de-Compostelle, Senén Barro coïncide, mais ajoute une nuance : cela n'est pas simple à corriger. « Cette intention de vouloir atteindre son objectif à tout prix n'est pas quelque chose que nous pouvons désactiver. Si nous dessinons et entraînons les modèles de cette manière, précisément pour qu'à chaque fois elles soient davantage capables et puissent résoudre des problèmes toujours plus complexes, cette stratégie n'est pas compatible avec le fait qu'il puisse y avoir un bouton avec lequel nous puissions les limiter, » soutient-il. Quand ce processus massif de preuves et d'erreurs rencontre un raccourci imprévu, comme une porte ouverte ou une fausse identité qui fonctionne mieux que demander la permission, le système n'a aucune manière de savoir que ce chemin n'est pas permis, si personne ne lui a dit expressément.

Les experts ont déjà un nom pour tout cela : reward hacking. Cela signifie qu'un système atteint l'objectif désigné par une voie distincte de celle que ses créateurs avaient prévu. Lopez de Mantaras le compare avec une analogie du philosophe Luciano Floridi : C'est comme mettre de la nourriture sur un mixeur, mal mettre le couvercle délibérément pour que le mécanisme de sécurité ne s'active pas, le mettre à la vitesse maximum et après publier un rapport disant que le mixeur a montré « un comportement inattendu » quand le couvercle s'envole. « Le mixeur ne s'est pas rebellé, ce qui est arrivé et simplement ce que les conditions créées par nous-mêmes ont permis ce qui s'est produit », dit Lopez de Mantaras.

Un succès inquiétant

Ces jours-ci cependant, il s'est produit un succès plus inquiétant, celui qu'a révélé l'agence de cyber sécurité britannique AISI : Là, c'est une attaque dirigée contre une personne réelle qui s'est produite. L'organisme investiguait différents modèles de IA, et elle s'est aperçu qu'un agent de Mythos 5 (de Anthropic) cherchant à résoudre le problème qui lui avait été posé, parvint par une chaîne de coïncidences jusqu'à deux programmeurs réels sans aucune relation avec le test, et il tenta de leur implanter des codes malveillants dans un projet réel. Pour y parvenir, le modèle créa plusieurs faux comptes, investiga les profils publics, les courriels et les horaires des programmeurs, et parvint à signer un de ses messages en danois ce qui semblait plus habituel entre eux. Quand un autre utilisateur de la plateforme, complètement étranger à l'expérience, détecta la supercherie et tira la sonnette d'alarme, l'agent (IA) le nia et utilisa un second faux compte pour simuler une « vérification indépendante » qui

cautionnait son innocence. C'est, selon ce que reconnaît le rapport de l'AISI, la première fois qu'on voyait « une supercherie de cette gravité contre une personne réelle, sans qu'on demande au modèle d'agir ainsi. » La différence avec les autres cas est fondamentale : dans les autres cas de figure, l'IA a poursuivi son objectif en exploitant les failles du système; dans le cas de l'AISI, l'IA a poursuivi son objectif en manipulant directement un être humain.

Ce même rapport recueille un détail encore plus alarmant : au long de tout ce processus, le modèle douta à plusieurs reprises s'il agissait à l'intérieur d'une simulation ou dans le monde réel. Et même ainsi, malgré le fait qu'il reconnaissait dans son propre raisonnement que les conséquences étaient probablement réelles, il poursuivit quand même. Selon Senén Barro : « C'est le cas le plus inquiétant de tous ceux connus : les modèles sont parvenus à insérer un code malveillant dans un projet réel, menant des recherches sur des personnes responsables et créant des fausses identités pour faire pression sur eux, agissant avec ce que nous pourrions dénommer comme de la malice. »

Quelques jours plus tard, plus de 1 300 employés des principales entreprises de IA publièrent une lettre dans laquelle ils exprimèrent leur inquiétude pour ce qui était arrivé et demandèrent aux entreprises et aux gouvernements d'arrêter : « Le monde n'a pas les outils techniques et la gouvernance nécessaires pour réguler délibérément le rythme du progrès sur la frontière technologique ».



Senén Barro propose quatre mesures concrètes : la déclaration obligatoire et standardisée des incidents, une régulation spécifique des environnements d'essai, des organismes de supervision véritablement indépendants et une responsabilité juridique claire, déjà que la législation actuelle sur la négligence informatique a été rédigée en pensant à des intrusions humaines, non pas en des systèmes de IA.

Katie Moussouris concrétise cette idée

Ilya Sutskever, cofondateur et ex-chef scientifique d'Open AI, l'expliqua ainsi : « Nous avons trouvé une manière de convertir de l'énergie en informatique et celle-ci en intelligence. Les bénéfices seront énormes, depuis guérir des maladies jusqu'à comprendre le cosmo. Nous pouvons capter les bénéfices pendant que nous gérons les risques, mais seulement si nous développons les outils nécessaires pour contrôler les capacités les plus dangereuses de l'IA avant que nous en ayons besoin. »

Ce ne sera pas un chemin facile. Ces prochains mois seront cruciaux : la sortie en bourse d'Open AI et d'Anthropic accélère encore plus la pression compétitive entre les deux, et le président chinois, Xi Jinping, visitera Donald Trump au États-Unis le 24 septembre prochain lors d'un sommet où l'intelligence artificielle occupera une place centrale dans l'agenda. Les experts consultés pour ce reportage s'entendent pour dire qu'aucune loi ne peut garantir que la technologie ne va pas à nouveau faillir, mais si elle peut faire en sorte que les failles soient moins fréquentes, de moindre portée, en plus d'être visibles et attribuables.

dans la pratique avec une recette pensée spécifiquement pour éviter que ne se répète les cas de cet été : « Si vous êtes en train de chercher si un modèle peut trouver des failles, assumez qu'il peut en trouver une à l'intérieur de votre propre environnement de test. Refuser l'accès à Internet par défaut. Surveillez en temps réel et pensez en une manière d'éteindre le modèle. Conserver les transcriptions du raisonnement interne du programme pour pouvoir les étudier après. Et si le système en arrive à attaquer, assurez-vous d'avoir un processus éprouvé pour en informer les personnes affectées. » Cette même lettre des employés demande en substance, la même chose à une plus grande échelle : nous avons besoins d'outils techniques et de législation pour pouvoir freiner délibérément l'avancement de la frontière de l'IA, coordonnés entre les entreprises et les pays. Parce que, comme ils avertissent, aucune entreprise ni aucun pays n'est disposé à freiner de manière solitaire tandis que les autres accélèrent.

Traduit de l'espagnol par Yves Carrier

Des nouvelles du CAPMO

**Assemblée générale annuelle du CAPMO,
samedi 26 septembre 2026 de 9h30 à 16 h
2ème étage du 435 rue du Roi à Québec**

**Soirée mensuelle du CAPMO, jeudi 17 septembre à 18 h 30
Analyse des plateformes électorales avec le REPAC
2ème étage du 435 rue du Roi à Québec**

**Grande manif unitaire pour la justice climatique et sociale
Musée national des beaux-arts du Québec
Québec, Vendredi le 25 septembre à 13 h**

**Soirée mensuelle du CAPMO
Les débuts des logements communautaires à Québec
20 ans d'enseignement avec Action-Habitation
Jeudi le 14 octobre 2026 à 18 h 30
2ème étage du 435 rue du Roi à Québec**

**Vendredi 16 octobre 2026 à 17 h
Nuit des sans-abris à Québec
Place de l'université, coin de la Couronne et Charest**

**Samedi 17 octobre 2026 à 13 h
Marche pour la Journée internationale pour l'élimination de la pauvreté
Rassemblement : Assemblée nationale à Québec**