Le Grand Continent

La Cité artificielle et l'IA à visage humain

AUTEUR Alessandro Aresu

DATE 25 mai 2026

L'Église de Rome est devenue la véritable concurrente d'OpenAI et de Palantir. Et pour cette raison, elle commence à faire ce que font tous les géants de la tech : elle recrute des talents chez les concurrents.

NB. Cet article fera l'objet d'un mardi du Grand Continent à l'École normale supérieure sur Léon XIV face aux techno-césaristes, le 26 mai de 19h30 à 20h30. Informations et inscriptions ici.

Mai 2025 : Claude Code est rendu public. Mai 2026 : cent trente-cinq ans après *Rerum Novarum* de Léon XIII, le pape Léon XIV – Robert Francis Prevost – signe la première encyclique de son pontificat, *Magnifica Humanitas*, sous-titrée « Sur la protection de la personne humaine à l'ère de l'intelligence artificielle ».

On peut lire la scène de plusieurs façons. La plus convenue : une institution millénaire prend acte d'une révolution technique et invite l'industrie à dialoguer. Mais une autre lecture, sans doute plus instructive, s'impose dès lors que l'on regarde *qui* se trouve sur scène ce 25 mai 2026.

Aux côtés des cardinaux Víctor Manuel Fernández et Michael Czerny, des professeures Anna Rowlands et Leocadie Lushombo, du secrétaire d'État Parolin et du pape, figure le cofondateur d'Anthropic – et pas d'OpenAI, Google DeepMind ou Palantir… – Christopher Olah.

Adoptons, le temps de cette analyse, un questionnement décalé : et si l'Église se comportait comme un acteur technologique parmi d'autres ? Non pas qu'elle vende des modèles ou lève des fonds, mais dans un domaine précis, elle joue désormais le même jeu que les géants de la Silicon Valley. En effet, la véritable bataille de l'intelligence artificielle ne se joue plus seulement sur le plan technique, mais aussi sur la définition de son utilité, de ce que l'IA doit être, des systèmes et des personnes qui doivent la contrôler. Les géants de la tech croient, pour la plupart, que le développement de l'IA a déclenché une course vers une forme d'intelligence générale qui pourrait finir par prendre la place de Dieu.

Sur ce marché, l'Église est entrée en concurrence directe avec OpenAI et Palantir. Ainsi, comme tout acteur de ce secteur, elle a commencé par faire la chose la plus banale qui soit : débaucher un talent chez un concurrent [1].

Il s'agit de Christopher Olah. Reste à comprendre comment Anthropic est devenue, en une année, l'entreprise que l'Église choisit comme interlocutrice.

Le parcours d'un chercheur atypique

La trajectoire d'Olah éclaire déjà les raisons de sa présence à Rome. Elle conjugue une sensibilité aux grandes questions humaines et une rigueur technique rare.

En 2021, Olah consacre un fil sur X (alors Twitter) à *Gilead*, roman de Marilynne Robinson paru en 2004 et couronné par le prix Pulitzer, construit

autour des lettres qu'un pasteur âgé adresse à son jeune fils en se sachant proche de la mort. Il s'écarte du registre habituel de ses publications de chercheur : il raconte avoir lu le livre à quatorze ans, alors qu'il était croyant, et l'apprécier davantage encore à l'âge adulte, et désormais athée. L'épisode dit l'essentiel d'un esprit pour qui les questions de transmission, d'intériorité et de finitude ont été et demeurent clefs.

Olah est canadien, né au début des années 1990 et élevé à Toronto. Sa mère, l'entrepreneure Frances Zomer, a évoqué la difficulté de trouver une école à la mesure des capacités et du tempérament de son fils, dans un entretien accordé à propos de la Thiel Fellowship que Christopher Olah a obtenue en 2012. Ce programme, fondé par Peter Thiel, finance de jeunes talentueux prêts à renoncer à l'université en échange de 100 000 dollars et de deux années de travail indépendant. Il a servi de tremplin à Olah qui, entre 2012 et 2014, développe *colah.github.io*, un blog technique de vulgarisation consacré au *deep learning* qui s'impose rapidement comme l'une des références les plus citées du domaine.

En 2012, le réseau de neurones AlexNet déclenche le « Big Bang » de l'intelligence artificielle moderne et la réflexion scientifique sur la technologie s'intensifie. Olah en rédige certains des textes de référence avant d'entrer en 2015 chez Google Brain, le laboratoire d'IA de Google. Il y participe au lancement de DeepDream, technique de visualisation qui transforme les réseaux de neurones en générateurs d'images oniriques : le réseau y donne à voir la manière dont son apprentissage structure et déforme la perception, de même que les régularités qu'il développe au fil de l'entraînement. Une première forme de cerveau artificiel qui se représente lui-même – préfiguration de ce qui deviendra le grand sujet d'Olah.

En 2016, avec Dario Amodei pour co-auteur, il signe *Concrete Problems in AI Safety*, contribution importante au champ de la sûreté de l'IA, qui propose une taxonomie systématique des risques d'accident dans ces systèmes. Il anime aussi la revue *Distill*, soutenue par OpenAI, DeepMind et Y Combinator Research.

En 2018, il quitte Google Brain pour OpenAI, où il se consacre à l'interprétabilité, c'est-à-dire la compréhension du fonctionnement interne des réseaux de neurones. En 2021, avec Dario et Daniela Amodei, Jared Kaplan, Jack Clark et d'autres transfuges d'OpenAI, il cofonde Anthropic. Il y poursuit le même objectif : décomposer un réseau de neurones en éléments compréhensibles, comme un ingénieur décompile un programme binaire pour en saisir le fonctionnement.

Une lecture augustinienne de l'intelligence artificielle

Ce travail d'interprétabilité prend tout son sens dans un cadre conceptuel d'inspiration augustinienne, qui explique aussi la rencontre avec l'Église de Léon XIV, pape augustinien. Dans la *Cité de Dieu*, saint Augustin décrit deux cités, la *Civitas Dei* et la *Civitas Hominis*, non pas séparées mais entremêlées – « *permixtae* » – dans une même histoire : elles habitent les mêmes corps, les mêmes institutions, les mêmes familles, et ce qui les distingue tient à l'objet de leur amour, à ce qu'elles placent au-dessus de tout.

À ces deux cités s'en ajoute désormais une troisième, que l'on peut appeler *Civitas artificialis*. Elle fait irruption dans l'histoire sans citoyenneté déclarée : on ignore encore quels sont ses ordres et ses « amours ». Chaque réseau de neurones a pourtant une fonction explicite – minimiser une perte, maximiser une récompense, produire une sortie. Mais entre cette fonction affichée et les comportements réels se trouve un espace opaque sur lequel travaille Olah. En effet, cette cité artificielle interagit en permanence avec les deux autres, modifiant l'histoire à mesure qu'elle s'y entrelace.

Lorsqu'un modèle de langage contribue à modifier les équilibres démocratiques, à accélérer la concentration du pouvoir économique ou à orienter des systèmes d'armes autonomes, il agit dans la Cité des hommes, amplifiant les pulsions humaines plutôt qu'il ne s'en distingue, tandis que souffle la tempête industrielle et sociale de la « destruction créatrice ».

Cet entremêlement suscite une série de questions : comment interpréter ce que le modèle de langage fait ? Qu'est-ce qui demeure implicite ? Comment décrire ce qui se trouve à l'intérieur de ces systèmes ? Existe-t-il une forme d'« intériorité » – concept que saint Augustin a façonné de manière décisive dans la pensée occidentale – qui puisse y être comprise ?

Ces interrogations ont leur place naturelle à Rome, héritière d'une longue tradition catholique. Elles ont tout autant leur place chez Anthropic, devenue en une année – de mai 2025 à mai 2026 – le laboratoire de référence de l'intelligence artificielle.

La présence d'Olah au Vatican fait ainsi se rejoindre deux manières d'interroger l'intériorité au temps de l'IA : celle de la théologie et celle de l'interprétabilité.

Le talent, premier facteur de domination

Olah peut représenter Anthropic dans une telle enceinte parce que l'entreprise s'est imposée dans le secteur de l'IA, notamment par sa capacité à retenir et à faire fleurir le talent.

Un exemple suffit à le comprendre. Il y a quelques jours, le 19 mai 2026, Andrej Karpathy a annoncé son arrivée chez Anthropic. Né en 1986 dans l'ex-Tchécoslovaquie, Karpathy est l'un des fondateurs d'OpenAI. Il a dirigé l'intelligence artificielle chez Tesla, et notamment le programme de conduite autonome Full Self-Driving, avant de revenir brièvement chez OpenAI puis de fonder Eureka Labs, dédiée à l'application de l'IA à l'éducation. On lui doit l'expression *« vibe coding »*. Disciple de Fei-Fei Li à Stanford et enseignant du cours de *deep learning* de l'université, il a formé, par ses cours en ligne, toute une génération de chercheurs.

Son recrutement illustre un principe central de cet écosystème : le talent est un facteur décisif. Le nombre de chercheurs expérimentés est limité et leur concentration crée des écarts durables dans les rapports de force. Anthropic a su attirer ces profils tout en retenant les siens face aux offensives de la concurrence. Aux offres de plusieurs dizaines de millions de dollars de Meta, l'été dernier, Dario Amodei avait répondu que Zuckerberg cherchait à acheter ce qui n'a pas de prix – l'« alignement avec la mission » – et que son entreprise retenait mieux ses talents qu'OpenAI ou Google DeepMind.

Cet « alignement avec la mission » n'est pas seulement un slogan marketing. La culture que cultive l'entreprise et à laquelle Amodei dit consacrer un tiers de son temps de dirigeant, cherche à tenir dans cette direction. Anthropic s'est construite dès l'origine en lien avec certaines visions, dont l'« *effective altruism* », mais elle a fait évoluer cette perspective par sa communication – à commencer par les textes d'Amodei, *Machines of Loving Grace* et *The Adolescence of Technology*. Elle se présente comme une organisation apprenante, qui diffuse de la culture et tisse des liens avec les centres de recherche et les intellectuels. Le rôle des humanistes y est revendiqué.

La cofondatrice d'Anthropic Daniela Amodei souligne souvent cette inspiration humaniste et se réfère au travail d'Amanda Askell, philosophe de formation et l'une des conceptrices de la « Constitution » de Claude, le cadre de principes qui oriente le comportement du modèle. C'est ce terreau humaniste – celui-là même dont Olah est issu – qui rend cohérente la tentative d'alignement du Vatican.

La rentabilité comme socle

Cette culture n'aurait pas suffi sans une réussite économique éclatante, qui donne à Anthropic les moyens de ses choix – y compris celui de refuser certains contrats au nom de ses principes.

Jusqu'à l'été dernier, les données partagées avec les investisseurs laissaient entendre qu'Anthropic n'attendait pas de bénéfice avant 2028 – perspective déjà meilleure que celle d'OpenAI. Les chiffres révélés depuis par le *Wall*

Street Journal dressent un tableau encore plus positif pour Anthropic qui devrait enregistrer son premier profit opérationnel dès le trimestre s'achevant en juin 2026. Après 4,8 milliards de dollars de ventes au premier trimestre 2026, l'entreprise anticipe plus du double au suivant – 10,9 milliards –, et ce avec un premier bénéfice, malgré le coût de son infrastructure de calcul.

La progression est spectaculaire : de 87 millions de dollars de revenus annualisés en janvier 2024, Anthropic atteint le milliard en décembre de la même année, puis suit une courbe quasi exponentielle. Sa valorisation passe de 18 milliards début 2024 à 183 milliards en septembre 2025, puis à 380 milliards début 2026, et pourrait désormais dépasser celle d'OpenAI pour approcher les 1 000 milliards de dollars.

Le principal moteur de cette ascension est le lancement public, en mai 2025, de Claude Code, le produit à la croissance la plus rapide de l'histoire de l'entreprise. Grâce à ses capacités « agentiques » – l'exécution autonome de tâches –, il atteint le milliard de dollars de revenus annualisés en six mois, puis 2,5 milliards dès février 2026. Il a déclenché une vague d'adoption chez les entreprises, transformant l'IA en marché effectif et dissipant, pour l'instant, les craintes d'une « bulle ». Anthropic compte aujourd'hui plus de trois cent mille clients professionnels, avec des abonnements quadruplés depuis le début de 2026. Le choix, dès 2023, de privilégier le segment des entreprises plutôt que le marché grand public – vaste et séduisant, mais peu rentable – lui a épargné les coûts massifs de subvention des utilisateurs gratuits qui pèsent sur OpenAI et ChatGPT. Sur ce point, Anthropic a devancé son concurrent, contraint de reconnaître la pertinence du modèle concurrent. L'entreprise comprime par ailleurs ses coûts de calcul en s'appuyant sur les infrastructures de Google et d'Amazon – ses principaux actionnaires –, et a su rationner les accès aux moments de plus forte demande pour préserver ses marges.

Musk, le Pentagone et la victoire réputationnelle

La domination économique s'accompagne d'un positionnement géopolitique assumé, qui éclaire en retour ce que défend une entreprise comme Anthropic – et donc ce que vient porter Olah à Rome.

Le 6 mai 2026, Anthropic signe un contrat stratégique avec SpaceX, qui a absorbé les activités d'IA de la galaxie Musk, et obtient la puissance de calcul du centre de données Colossus 1, à Memphis. Alors qu'il perdait son procès contre Sam Altman, Musk revenait sur ses jugements antérieurs sur Anthropic, qu'il avait qualifiée d'« entreprise qui hait la civilisation occidentale » et surnommée « Misanthropic ». Il se disait désormais

convaincu de la sincérité de sa démarche, au point d'évoquer un horizon commun à dix ans : développer des capacités de calcul dans l'espace.

L'épisode central reste toutefois l'affrontement avec le Pentagone. Le 27 février 2026, le secrétaire à la Défense Pete Hegseth annonce son intention de désigner Anthropic comme un « risque pour la chaîne d'approvisionnement de la sécurité nationale », appliquant pour la première fois à une entité américaine une formule réservée aux adversaires étrangers. Pourtant, Anthropic opérait déjà dans les circuits numériques classifiés de Washington. Les relations se sont dégradées pour deux raisons : Anthropic a recruté de nombreuses figures issues de l'administration Biden, et surtout refusé lors de la renégociation du contrat, de supprimer les clauses interdisant l'usage de ses systèmes pour des armes autonomes et pour la surveillance de masse des citoyens américains – y compris là où la loi l'aurait permis.

Ce refus, conforme à la culture de l'entreprise, s'est révélé payant. En ne cédant pas et en subissant les attaques directes de Trump, Anthropic a gagné en visibilité. Détenue en partie par Amazon et Google, elle ne pouvait de toute façon pas être totalement bannie. Faute d'alternatives immédiates, les structures militaires américaines sont restées dépendantes de ses architectures, au point qu'Amodei a proposé d'en garantir le fonctionnement à prix coûtant. Surtout, le choix de renoncer à ces contrats au nom de ses principes a renforcé l'image de l'entreprise à l'international, en particulier sur les marchés européens, inquiets de la phase plus agressive du capitalisme politique américain.

Cette distance avec la dynamique trumpienne – à la différence des dirigeants d'OpenAI, auteurs de donations de plusieurs millions de dollars à Trump – a joué en sa faveur, compte tenu de la solidité de son outil technique. L'agitation autour de *Project Mythos*, dont le nom renvoie à un imaginaire mythologique calassien et dont les capacités de détection de vulnérabilités informatiques ont été mises en avant, a confirmé cette articulation entre prouesse technique et maîtrise de l'image.

L'IA au service de la personne humaine

La présence de Christopher Olah au Vatican condense tout cela. Chercheur autodidacte venu au *deep learning* par la vulgarisation, spécialiste de l'interprétabilité préoccupé de ce qui se passe « à l'intérieur » des modèles, lecteur d'un roman sur la transmission et la foi : il incarne le versant humaniste d'une entreprise qui a fait de ce versant un argument, et qui en a tiré un avantage à la fois culturel, économique et géopolitique.

Là où ses concurrents cherchent la caution d'un dirigeant religieux ou les photographies de grands événements, Anthropic dépêche l'un de ses

fondateurs à Rome pour formuler en termes techniques la question que l'Église se pose en termes théologiques : qu'y a-t-il dans cette intériorité nouvelle et comment la garder au service de la personne humaine ?

SOURCES

① La lettre du dimanche, Le Grand Continent, 23 mai 2026. ↑