ASSEMBLÉE NATIONALE

17e LÉGISLATURE

Compte rendu

Commission d'enquête sur les dépendances structurelles et les vulnérabilités systémiques dans le secteur du numérique et les risques pour l'indépendance de la France

Mardi
12 mai 2026
Séance de 16 heures

Compte rendu n° 42

SESSION ORDINAIRE DE 2025-2026

Présidence de *Mme Isabelle Rauch, Vice-présidente de la commission*

ASSEMBLÉE NATIONALE

La séance est ouverte à seize heures.

Mme Isabelle Rauch, présidente. Dans le monde de l'intelligence artificielle (IA) générative, Mistral AI fait figure d'exception. Non seulement votre modèle est open source, mais de plus tous vos principaux concurrents sont extra-européens. Beaucoup d'espoirs reposent sur vous.

On présente souvent l'intelligence artificielle comme un secteur encore suffisamment ouvert pour que des acteurs européens puissent rivaliser avec les géants du numérique. Dans la course à l'IA, quels sont les atouts de la France et de l'Europe ? Que pouvons-nous faire pour assurer la souveraineté numérique de l'Europe, et de la France en particulier ?

Avant de commencer, je vous remercie de nous déclarer tout intérêt public ou privé de nature à influencer vos déclarations.

L'article 6 de l'ordonnance du 17 novembre 1958 relative au fonctionnement des assemblées parlementaires impose aux personnes auditionnées par une commission d'enquête de prêter le serment de dire la vérité, toute la vérité, rien que la vérité.

(M. Arthur Mensch et Mme Audrey Herblin-Stoop prêtent successivement serment.)

M. Arthur Mensch, cofondateur et directeur général de Mistral AI. Merci de nous avoir invités pour parler de l'autonomie de l'Europe en matière d'intelligence artificielle et plus largement de services numériques ; c'est un sujet crucial. Selon moi, il ne faut pas séparer ces deux domaines : les services numériques sont essentiellement faits d'intelligence artificielle, et ce sera encore davantage le cas demain. Certes, il y a un peu de stockage et d'exécution d'agents mais, fondamentalement, les services numériques, le cloud, c'est l'intelligence artificielle. Il n'y a pas de distinction à faire.

L'intelligence artificielle redéfinit les équilibres économiques et géopolitiques. La question n'est pas tant de savoir si l'Europe peut rivaliser – y renoncer, ce serait abandonner toute participation au concert des nations – mais de déterminer comment. Le peut-elle sans dépendre d'acteurs extra-européens ? La réponse est évidemment non. En revanche, je pense qu'elle pourra y parvenir en renforçant sa position vis-à-vis de ses fournisseurs et en tant qu'exportateur de technologies, et en faisant en sorte de pallier les déséquilibres économiques, qui vont encore s'aggraver.

Nous avons fondé Mistral AI le 28 avril 2023, avec l'intuition que l'intelligence artificielle générative rebattrait largement les cartes dans le domaine des services numériques et changerait profondément le fonctionnement de l'économie. En effet, l'IA déplace l'intelligence du travail vers le capital et les machines.

Parce que notre expérience chez les gros acteurs américains nous avait convaincus, mes cofondateurs – Guillaume Lample et Timothée Lacroix – et moi, qu'un oligopole se dessinait, nous avons créé Mistral AI, pour le contrer. Nous étions accompagnés d'une quinzaine de personnes, en France et en Angleterre, essentiellement des chercheurs. Nous avons d'abord montré que nous savions entraîner des modèles de langue, puisque c'est ce qu'il faut pour faire de l'intelligence artificielle générative. Nous avons créé un modèle ouvert, c'est-à-dire que nous avons donné ces modèles au monde ; chacun pouvait s'en emparer, les modifier,

les déployer où il voulait. C'est une technique de décentralisation et de déplacement des acteurs établis. Puis nous avons continué à investir dans les modèles – cette année, nous consacrons 1 milliard à la R&D (recherche et développement). Nous sommes un des leaders des modèles ouverts.

Notre modèle économique est double. D'une part, il fonctionne par-dessus : nous proposons une plateforme de déploiement d'agents, de délégation de tâches à de l'intelligence artificielle, et nous créons des applications métier au contact de nos clients, qui sont souvent de grosses entreprises.

D'autre part, il fonctionne en amont : nous hébergeons également les modèles, nous les mettons sur des machines, et nous leur faisons générer ce qu'on appelle des tokens – des jetons. Le token est l'unité économique de l'intelligence artificielle. On entraîne des modèles qui génèrent des séquences ; à partir de ces générations de séquences, on décide d'exécuter des actions, on réfléchit. Les séquences sont faites de tokens, à savoir de chiffres, compris entre 0 et 65 000. On génère ces tokens pour nos clients, en les hébergeant. On transforme des électrons en tokens.

L'entreprise compte désormais 1 000 collaborateurs ; elle est valorisée 12 milliards d'euros et nous visons 1 milliard de revenus d'ici à la fin de l'année. Nos clients sont des entreprises et des institutions publiques ; en France, ce sont par exemple la Dinum (direction interministérielle du numérique), la Caisse des dépôts, France Travail, CMA-CGM, Stellantis, TotalEnergies, BNP Paribas, etc. Nous faisons à peu près 30 % de notre chiffre d'affaires en France, 75 % en Europe. Donc nous sommes un exportateur de technologies, vers les États-Unis et vers l'Asie.

Ce qu'il faut comprendre sur l'intelligence artificielle, générative notamment, c'est qu'elle consiste à transformer de l'énergie en intelligence. Il s'agit de générer des tokens ensuite utilisés pour déléguer des tâches, faire de la recherche mathématique, opérer des chaînes logistiques, améliorer des outils, déployer des systèmes embarqués dans des robots ou des drones, etc. Puisqu'on transforme de l'électricité en tokens, il faut réfléchir à l'intelligence comme on réfléchit à l'énergie – à une ressource naturelle. Notre objectif est de fournir une intelligence peu chère – *affordable*, disons, en anglais ; dont l'approvisionnement est sécurisé – qu'on n'a pas besoin d'aller chercher chez les Américains, par exemple ; et durable – elle transforme une énergie moins carbonée qu'ailleurs, en particulier parce qu'elle est française. On n'y pense pas assez en ces termes : l'intelligence, c'est finalement comme l'électricité.

On a par ailleurs tendance, à tort selon moi, à séparer les services numériques de l'intelligence artificielle. Or le cloud, et sa croissance, c'est maintenant de l'intelligence artificielle ; les services à haute valeur ajoutée, c'est-à-dire ceux qui offrent une forte marge donc permettent de financer de la R&D, c'est l'intelligence artificielle. À partir du moment où on développe l'intelligence artificielle, on peut construire tout le reste des services du cloud, qui sont surtout des commodités. Ça consiste en grande partie à savoir exploiter du logiciel open source, puisque c'est devenu l'infrastructure mondiale.

Notre chaîne de valeur va de l'opération des clusters et des serveurs à la création des applications métier ; nous sommes une unité atomique de la chaîne de valeur plus large, qui va de la construction de semi-conducteurs – avec par exemple ASML, une entreprise qui fait de la lithographie et pour qui nous travaillons – au déploiement dans les entreprises.

Le cloud et l'IA, c'est donc vraiment la même chose.

Il faut arrêter de penser à la souveraineté comme à un isolationnisme et y penser comme ceux qui font des affaires, en se posant la question des leviers. Dans un monde où on importerait la totalité des services numériques des États-Unis, on n'aurait pas de levier pour agir sur eux. Il en va différemment si on crée une partie de nos services et qu'éventuellement on les exporte. On le voit dans les négociations de l'Union européenne : exporter des biens nous donne des arguments. Si nous ne pouvons faire autrement que d'importer la totalité de nos services numériques des États-Unis, nous avons un vrai problème.

Pour comprendre l'importance des enjeux, il faut raisonner à l'échelle de la masse salariale. Chez Mistral, notre consommation d'intelligence artificielle représente 10 % de notre masse salariale. Comme nous fabriquons cette technologie, nous l'adoptons un peu plus vite que d'autres. Mais en extrapolant, on comprend que, d'ici trois à quatre ans, toute l'Europe utilisera l'intelligence artificielle à hauteur de 10 % de sa masse salariale, soit environ 1 trilliard. Nous parlons ici de services numériques, puisqu'il faut bien générer les tokens. Si nous importons pour cela une technologie non européenne, alors notre déficit commercial augmentera de 1 trilliard ; cette somme partira et sera investie dans la R&D ailleurs qu'en France et en Europe.

Il faut donc réfléchir à la souveraineté du point de vue macroéconomique. C'est un sujet essentiel, dont on ne réalise pas l'importance.

Un autre problème est la sécurité économique. Quand vous déployez de l'intelligence artificielle dans des services essentiels, il faut vous assurer que l'accès ne puisse pas être coupé. On l'a vu à différents endroits, la question n'est pas théorique. Évidemment, ce point est essentiel dans tous les domaines régaliens, notamment pour la défense. L'intelligence artificielle est désormais au cœur des centres opérationnels des armées : elle est l'outil qui permet de prendre les décisions sur les champs de bataille et de faire du renseignement, par exemple. Si nous n'avons pas ces technologies, nous devrons les importer, donc nous risquons de nous les voir refuser.

Dernier problème, qui ne concerne pas directement notre entreprise parce que notre activité grand public n'est pas significative : l'intelligence artificielle, parce qu'elle génère du contenu, façonne les représentations culturelles, la langue, l'éthique. En effet, les modèles que nous créons ont une certaine politique ; ils actionnent certains leviers, ils écrivent un certain type de code, etc. Tout cela devient une médiation de l'information et une médiation de l'action. Faute de solution alternative européenne, nous serons contraints d'importer des modèles, donc nous dépendrons de choix faits par d'autres, les États-Unis et la Chine en particulier, et des biais qui en découlent.

Pour finir, je formulerai quelques recommandations. L'intelligence artificielle offre l'espoir de rebattre les cartes. Vous m'avez demandé quels étaient les atouts de l'Europe, en particulier de la France. Le premier, c'est son vivier de talents. C'est une entrée dans le système. Si nous n'avons pas de grandes entreprises, innovantes, capables de capter ces talents et de les convaincre de travailler pour nous, ils partiront porter leur valeur ailleurs. On l'a vu beaucoup ; les choses commencent un peu à se renverser, il faut qu'on en bénéficie.

En dehors des talents, l'Europe n'a pas beaucoup d'atouts. Sa réglementation est plus lourde. Son marché est fragmenté, ce qui veut dire plus de difficultés pour vendre – mais aussi des partenariats de plus long terme, ce qui est un atout. Sa grande diversité culturelle, linguistique par exemple, est aussi un atout pour l'intelligence artificielle.

Nous n'existons que s'il y a un marché. Le problème, c'est que ce marché se matérialise un peu trop lentement : les entreprises n'ont pas encore pleinement conscience de la quantité d'intelligence artificielle qu'elles utiliseront dans les années à venir. La conséquence, c'est que les Américains vont être plus à même d'investir avant la demande.

En effet, en France, nous avons du surplus énergétique – près de 9 gigawatts en moyenne. Il sera en grande partie capté par les entreprises les plus agressives du point de vue capitalistique et capables de mettre de l'argent sur la table pour transformer ces électrons en tokens avant les autres. Nous arrivons à être assez agressifs mais nous aimerions l'être encore plus. Pour cela, nous avons besoin d'une visibilité sur le marché. Celui-ci est constitué d'entreprises privées, qui choisissent de travailler avec nous pendant cinq ou sept ans – il y en a plein –, mais aussi de la demande publique, qui forme près de 50 % du PIB de l'Europe.

Il faut donc faire en sorte que la demande publique ruisselle sur toute la chaîne de valeur, en partant de l'endroit où nous sommes forts, où nous avons de la valeur ajoutée, c'est-à-dire l'intelligence artificielle. Et il faut le faire de manière concentrée, et non distribuer des subventions pour alimenter la demande en bout de chaîne, à savoir dans les services numériques. C'est un atout essentiel, que nos partenaires américains et chinois utilisent massivement, mais que nous avons toujours eu peur d'utiliser en Europe ; il faut absolument changer cela.

De manière générale, nous essayons de structurer les marchés et d'avoir une demande unifiée. Les marchés unifiés, c'est utile, les marchés de capitaux unifiés, c'est utile – on n'en fait jamais assez.

S'agissant des seules infrastructures, la France a la chance d'avoir de l'électricité en surplus. Il faut continuer à en produire : l'intelligence artificielle, c'est environ 1 kilowatt par personne à l'horizon de cinq ans. Ça veut dire 40 gigawatts à fabriquer en France, 400 gigawatts en Europe, soit quelque 20 trilliards à investir. Il faut en avoir conscience ; il faut produire l'électricité ; surtout, il faut faire en sorte que les électrons soient transformés en tokens par des entreprises européennes.

Pour construire 400 gigawatts en Europe, il faut accélérer les projets : il faut délivrer les permis plus rapidement et prévoir une réservation des capacités énergétiques.

Pour répondre à votre seconde question, en matière de services numériques, au-delà de l'IA, on peut faire du *database management*, de l'hébergement, des machines virtuelles, etc. Mais ce sont des services qui ont une marge très faible. Or vous ne pouvez pas construire un business à l'échelle en partant d'un endroit où la marge est faible. Il faut partir de l'endroit où la marge est importante ; ensuite, une fois que vous êtes suffisamment gros, vous utilisez une partie de votre R&D – même faible – pour créer tous les services que vous utilisez.

Si vous partez de l'endroit où la marge est faible, et que de surcroît vous n'avez pas de volume, vous n'avez aucune chance d'arriver à l'échelle. Avec une faible marge mais un gros volume, vous pouvez remonter la chaîne de valeur – c'était la stratégie chinoise. Si vous êtes dans un segment où vous avez une haute marge et peu de volume au début, votre seule possibilité, c'est de descendre la chaîne de valeur. C'est très important d'avoir ça à l'esprit. Donc il faut partir des services haut niveau et descendre vers les services bas niveau, qui ont moins de marge. Sur le plan stratégique, c'est comme ça que nous appréhendons le problème.

Mme Cyrielle Chatelain, rapporteure. Merci pour cette introduction, centrée sur le modèle économique – c'est un élément essentiel.

Commençons par les questions simples : quel est le coût de revient d'un token ?

M. Arthur Mensch. Le token est l'unité économique de l'intelligence artificielle. On facture en millions de tokens ; par exemple, sur un de nos modèles, 1 million de tokens coûtera 1 euro. Si vous m'envoyez 1 million de tokens afin d'accomplir une tâche qui va nécessiter que je génère 1 million de tokens, vous me paierez 1 euro pour l'entrée et 3 euros pour la sortie.

Un token, c'est quelques lettres dans un mot, un chiffre, quelques pixels dans une image, un très court son. Et on sait gérer de la donnée multimodale. C'est l'unité.

Notre modèle économique est partiellement hébergé, ou à la consommation : vous hébergez de la technologie chez nous et vous payez au token ou à l'unité de temps de CPU (unité centrale de traitement) quand vous faites autre chose que de l'intelligence artificielle – vous payez au stockage. Nous avons également un modèle de plateforme : nous apportons notre logiciel chez le client et nous le déployons sur son infrastructure : il paie celle-ci – ses GPU (unités de traitement graphique), ses CPU, son stockage et, par ailleurs, nous lui facturons un modèle de type licence auquel s'ajoute une couche de services pour transformer l'entreprise et construire de la valeur.

Mme Cyrielle Chatelain, rapporteure. Les chiffres que vous avez cités, 1 euro et 3 euros, correspondent-ils à ce que paie le consommateur ?

M. Arthur Mensch. On parle plutôt du développeur, qui va créer des applications avec les données.

Mme Cyrielle Chatelain, rapporteure. Les développeurs sont vos clients ?

M. Arthur Mensch. Oui.

Mme Cyrielle Chatelain, rapporteure. Donc ces chiffres désignent ce que vous facturez à vos clients mais, pour produire ces tokens, vous-mêmes avez dû investir. Nous avons besoin d'évaluer les capex (dépenses d'investissement) nécessaires pour développer le modèle, qu'il s'agisse de l'infrastructure ou de l'énergie. Combien cela vous coûte-t-il de générer 1 token ?

M. Arthur Mensch. Le token est l'unité client du service numérique – quand vous achetez un service numérique, vous achetez des tokens. Pour financer nos projets, nous raisonnons plutôt en mégawatts – l'électricité qu'on va transformer en intelligence.

Quand vous investissez dans la création d'un data center et que vous mettez des cartes graphiques, qui vous permettront de générer des tokens dans le cadre d'un modèle hébergé, il faut investir 50 milliards de dollars sur cinq ans pour créer 1 gigawatt. Donc ça vous coûte à peu près 10 milliards par an. Pour avoir une idée des marges, sachez qu'avec ça, si vous faites les choses correctement, si vous êtes à peu près à l'échelle – 1 gigawatt, c'est à peu près à l'échelle –, vous pourrez vendre environ 20 milliards de tokens par an, donc faire à peu près 20 milliards de revenus. Donc sur la fourniture du service numérique, la marge brute est d'environ 50 %. Significative, elle permet de faire de la R&D, en particulier de financer l'entraînement des modèles, qui coûte extrêmement cher. En effet, nous utilisons aussi nous-mêmes des GPU : nous avons besoin de dizaines de mégawatts pour entraîner des modèles.

J’ai dit que 10 % de la masse salariale européenne pourrait devenir de l’intelligence artificielle. On parle d’environ 400 gigawatts. En effet, on va dire qu’on parle de 10 000 euros par an, c’est-à-dire environ 1 kilowatt de GPU – 1 GPU, c’est 2 kilowatts, donc avec cette somme, vous louez 0,5 GPU. Pour construire ça, il faut donc investir 20 trillions, qui vont rapporter environ 8 trillions par an – autant de revenus qui pourraient partir aux États-Unis. Les ordres de grandeur sont significatifs, vous le voyez.

J’ai dit que la marge brute du fournisseur de services était de 50 % environ. En fait, dans la chaîne qui va de l’électron au token, l’électron représente à peu près 10 % du coût total. Autrement dit, si l’Europe se contente de fournir l’énergie, 90 % de la valeur partira ailleurs, pour être investie ailleurs, recyclée en R&D, et utilisée comme vecteur de puissance.

Mme Isabelle Rauch, présidente. Monsieur le président, puisque vous êtes parmi nous, avez-vous des questions ?

M. le président Philippe Latombe. Quelles limites éthiques vous êtes-vous fixées ? Ma question concerne le domaine de la défense, en écho à la polémique qui a opposé Anthropic au ministère de la guerre des États-Unis – puisque c’est comme cela qu’ils l’appellent maintenant –, mais aussi le remplacement des êtres humains, puisque vous avez parlé du remplacement de la masse salariale.

Avez-vous pu toucher du doigt Mythos et voir ce que ça vaut ? Nous avons posé la question à Vincent Strubel, qui nous a dit qu’il ne l’avait pas lui-même manipulé. Est-il aussi puissant qu’on le dit ?

Avec l’IA, on assiste à une accélération. Des modèles toujours plus puissants sortent régulièrement. Quelle étape ne faut-il absolument pas rater si nous voulons ne pas décrocher ? À quel terme faut-il investir ? À quelle chronologie la puissance publique doit-elle réfléchir ?

Les Chinois pratiquent beaucoup la distillation, notamment avec DeepSeek. Est-ce un bon moyen de rattraper le temps, ou est-ce que cela permet de compenser un peu mais sans réelle utilité, parce que le modèle produit est moins puissant ou efficace que ceux utilisés pour l’entraînement ?

M. Arthur Mensch. Votre première question concerne l’éthique. Nous faisons des systèmes qui aident les humains à être plus efficaces. Ils automatisent des procédés en laissant les humains dans la boucle pour prendre les décisions importantes. Nous avons toute une équipe de designers, composée d’une dizaine de personnes, qui réfléchissent aux interactions entre l’homme et la machine. Quand l’intelligence artificielle résume des informations et aide à décider, ils font en sorte qu’elle fournisse suffisamment d’éléments pour que l’humain reste informé.

Ce n’est pas facile, parce que l’humain a tendance à être un peu paresseux : plus le truc fait à sa place, moins il réfléchit. Donc, pour le laisser dans la boucle, ce qui est important, il faut le stimuler quand il est face à la machine. Évidemment, ce sont des problèmes auxquels nous réfléchissons.

Une des questions éthiques porte plus précisément sur l’engagement dans la défense. Nous travaillons avec le ministère des armées ainsi qu’avec des alliés de la France. Comme il s’agit d’une technologie à usage dual, la réglementation relative au contrôle des exportations s’applique. En tant que société, nous ne pouvons prétendre avoir la légitimité pour définir

nous-mêmes le cadre éthique. En revanche, nous nous inscrivons dans un cadre qui a une légitimité démocratique, ce qui nous permet de travailler correctement et de ne pas faire ce que nous n'avons pas le droit de faire.

Au contraire d'autres acteurs, nous ne prétendons pas avoir un droit de regard sur l'utilisation finale que le ministère des armées fera de nos produits. Nous n'avons pas la légitimité pour expliquer aux armées ce qu'elles peuvent ou ne peuvent pas faire de la technologie. Nous pouvons, dès le stade du projet, leur expliquer la fiabilité des outils, par exemple. Mais, ensuite, c'est l'armée qui décide souverainement de leur utilisation. Elle a pour ce faire une légitimité démocratique que nous n'avons pas. Nous, nous avons un devoir de conseil : nous expliquons comment ça fonctionne. Je peux vous dire qu'aujourd'hui, l'intelligence artificielle est utilisée de manière raisonnable, dans les limites de ce qu'elle peut faire.

Il faut avoir conscience que l'intelligence artificielle peut servir la défense. Elle permet d'avoir des armées beaucoup plus opérationnelles. Mais elle est aussi un facteur de dissuasion décisif. Certaines armées, comme l'armée russe, l'utilisent massivement dans les drones : si, en face, on n'est pas capables d'avoir des contres, eux-mêmes activés par de l'intelligence artificielle, notre dissuasion n'est pas suffisante. Donc le déploiement de l'intelligence artificielle, dans des systèmes autonomes en particulier, il ne faut absolument pas y couper. C'est indispensable pour avoir une dissuasion conventionnelle. Ce qui veut dire qu'il faut maîtriser l'intelligence artificielle de bout en bout.

J'en viens à la cyberdéfense et à la cyberattaque. L'un de nos concurrents américains sait très bien faire du marketing de la peur. On parle de modèles qui seraient d'excellents programmeurs, capables d'orchestrer des attaques, de découvrir des vulnérabilités et de proposer des exploitations. Mais ça fait six mois que ça monte ; de plus, ça monte de manière linéaire et prédictible, et ça monte chez tout le monde en même temps. Un autre de nos concurrents américains a un modèle qui fait exactement la même chose. Certains acteurs chinois le font également. Et nos modèles à nous sont capables de découvrir toutes les vulnérabilités qui ont été citées comme mises à jour par Mythos notamment.

Nous travaillons avec nos clients pour les aider dans ce domaine. Là encore, c'est un sujet régalien : cela fait un argument supplémentaire pour contrôler cette technologie. On ne peut pas envisager que les bases de code de l'armée française soient scannées par Mythos. La dépendance dans ce domaine est tellement irrémédiable qu'il faut trouver des solutions.

Vous avez évoqué la question du temps. Eh bien justement, nous en manquons, notamment parce que, pour transformer de l'électricité en tokens, nous utilisons des ressources naturelles qui ne sont pas infinies. D'ailleurs, les États-Unis saturent leur électricité afin de la transformer en intelligence artificielle. Ils dépensent à cet effet 1 trilliard de dollars, ce qui veut dire qu'ils attendent 2 trilliards en retour. Car ils voient bien que, face à la création d'un immense marché, le gagnant est celui qui dispose des *chips*, qui possède les processeurs et les électrons, qui accède de façon massive à l'énergie.

La France dispose de 90 térawattheures d'électricité par an. On les utilise, entre autres, pour l'intelligence artificielle. Certains acteurs ont largement les moyens financiers nécessaires pour transformer des électrons en IA, quelle que soit la demande. Ils vont le faire dans les deux prochaines années. Ensuite, il n'y aura plus d'électricité et il faudra créer de nouvelles centrales. On assiste à un phénomène, qui s'accélère à une vitesse folle, de monopolisation de la ressource énergétique européenne. On n'a pas conscience de ce processus pourtant irrémédiable.

Nous sommes confrontés à un problème de demande : les Américains, qui nous ont devancés sur le cloud, fournissent toute l'industrie européenne et bien sûr l'État français. Non seulement leur poids est de plus en plus important, mais leur part de marché dans le secteur du cloud et de l'intelligence artificielle augmente, tandis que celle des acteurs non américains a plutôt tendance à se compresser. S'agissant de l'intelligence artificielle, la tendance pourrait s'inverser, comme nous commençons d'ailleurs à l'observer – si l'on en juge par nos clients, par nos revenus et par notre croissance.

Cependant, même si on règle le problème de la demande, celui de l'offre demeure. Une fois que celle-ci est monopolisée par des acteurs américains, il ne nous reste plus rien : nous ne pouvons plus transformer des électrons en tokens. Il faut donc investir, dès aujourd'hui – dans deux ans, il sera trop tard –, des centaines de milliards de dollars, ce qui suppose que les acteurs et les États européens nous permettent d'avoir une forte visibilité sur le marché et la demande, ce qui n'est pas le cas actuellement. Une prise de conscience de cette urgence est nécessaire. Malheureusement, je ne vois pas comment elle pourrait intervenir. L'objectif de Mistral est d'atteindre une puissance de calcul d'1 gigawatt d'ici à 2029, ce qui est encore insuffisant.

J'en viens enfin à la distillation, une opération qui consiste à concevoir, à partir d'un grand modèle, un modèle plus petit et plus efficace. Nous y avons recours de façon massive pour proposer à nos clients des modèles moins coûteux. Nous construisons et entraînons également de grands modèles car c'est indispensable pour connaître le contenu des modèles et pour les contrôler. Cela nous coûte très cher car cela suppose de disposer, en propre, de fortes capacités de calcul, ce qui nécessite des investissements importants – c'est un pur enjeu de R&D. C'est donc une technologie qui permet de réduire les coûts en interne ; elle ne permet pas de rattraper le temps.

M. Arnaud Saint-Martin (LFI-NFP). Le projet Campus IA, auquel vous êtes associé, est en cours d'installation à Fouju, au nord de ma circonscription. J'aimerais avoir des détails sur son architecture générale. Une concertation, à laquelle j'ai participé, a été menée en amont et l'enquête publique est réalisée en ce moment. Pour cet énorme projet, présenté comme notre vaisseau amiral lors du sommet pour l'action sur l'IA de 2025, le fonds souverain d'Abou Dhabi MGX a abondé 35 milliards et un partenariat a été conclu avec l'Américain Nvidia.

Quelle est votre appréciation des impacts environnementaux de ce mégaprojet ? Il consommera 100 hectares de terres arables – une ressource à protéger –, ce qui n'est pas rien, et aussi une grande quantité d'électricité, 1,4 à 1,6 gigawatt – le niveau de la centrale de Flamanville, souvent citée comme référence en la matière.

De même, comment envisagez-vous la question de la souveraineté numérique, dans la mesure où la construction de ce projet inclut un acteur moyen-oriental, venu faire des affaires dans notre pays. J'ai bien compris que la volonté d'attirer de telles entreprises était le fruit d'une politique gouvernementale mais j'aimerais savoir comment vous vous positionnez à l'intérieur de cette architecture sur le plan à la fois économique et technique.

Par ailleurs, à quoi serviront ces data centers modulaires destinés à former une macro-infrastructure ? Quels usages sont envisagés ? Qui sont les potentiels clients que vous démarchez – vous en avez cité quelques-uns ? Quel est le calendrier ?

Il est important que vous apportiez des réponses à la représentation nationale. En effet, ce projet est colossal parce qu'il représente 35 milliards – voire beaucoup plus – et parce qu'il

transformera ce territoire de Seine-et-Marne de façon importante. Il a déjà fait disparaître les champs de betteraves et de blé qu'on y trouvait auparavant – l'archéologie préventive effectue actuellement des fouilles.

Je m'interroge aussi sur le nom « Campus IA ». Des partenariats avec Polytechnique ou d'autres structures de l'enseignement supérieur et de la recherche avaient été évoqués mais ne s'agit-il pas uniquement d'une mesure d'affichage pour vendre le projet ? D'après ce que j'ai compris, on fera peu de recherche sur ce « campus ».

M. Arthur Mensch. Le rôle de Mistral dans le projet Campus IA est très mineur. Nous avons pris une toute petite participation. Au sein de la chaîne de valeur, nous savons construire les clusters : nous achetons les serveurs puis nous les branchons dans les bâtiments – un peu comme des racks physiques. En revanche, nous ne nous occupons ni des bâtiments eux-mêmes, ni des transformateurs, ni de leur refroidissement ; pour cela, nous devons aller chercher des partenaires, nous avons besoin de développeurs de data centers. Une partie de Campus IA jouera pour nous un rôle de fournisseur, une autre partie alimentera les *hyperscalers*.

Il faut construire des infrastructures car 80 % des services numériques européens sont importés des *hyperscalers*. La France et l'Europe doivent avoir pour objectif de réduire cette part.

Vous m'interrogez également sur le nom ; ce n'est pas moi qui l'ai choisi.

Par ailleurs, je n'ai pas d'information concernant la place de la recherche mais il est certain qu'une partie des revenus sera reversée à la recherche publique.

J'en viens à la question du capital. Si l'on dispose de fonds importants, la location d'un data center représente un business assez intéressant, qui séduit beaucoup d'entreprises, car elle assure un bon retour sur investissement – d'autant plus que tout le monde veut des data centers. Il faut préciser que des sites de cette dimension nécessitent des acteurs aux reins solides, capables de dépenser des dizaines de milliards d'euros.

Faute de fonds de pension, en France par exemple, nous devons nous tourner vers des investisseurs étrangers : sur les marchés de capitaux en Europe, personne n'est en mesure d'apporter ce type d'investissements. Il faut opter pour cette solution dès lors qu'un contrôle est garanti et que la gouvernance est bonne – à cet égard, BPIFrance, qui dispose d'un siège au conseil d'administration, joue son rôle – car ces investissements sont bénéfiques et nous permettent d'accélérer, de construire des projets. Bien sûr, on perd de l'argent lorsque des capitaux étrangers entrent en jeu puisque les investisseurs voudront ensuite récupérer leurs billes. Reste que, sans eux, nous ne pourrions rien construire.

Nous ne sommes pas capables, pour l'instant en tout cas, de proposer un modèle totalement intégré : nous n'opérons pas depuis l'installation sur le terrain jusqu'au fonctionnement des services numériques. En revanche, il y a un partage de la valeur. Nous sommes en mesure de créer de la valeur en gagnant de la marge – sur les étapes qui vont des opérations des serveurs jusqu'aux services numériques – et en réinvestissant cet argent dans la R&D.

Il faut donc envisager la souveraineté comme un chemin. Le point de départ, c'est un certain état de l'économie, marqué par différents équilibres sur l'ensemble de la chaîne de

valeur. Si on ne le juge pas satisfaisant, il faut changer les choses, et donc affecter les ressources là où elles sont nécessaires. Cela demande du temps mais, peu à peu, cela produit des effets.

J'en arrive aux questions environnementales. Bien sûr, ce type de projet a un coût, et il y a des externalités. Par exemple, la production de semi-conducteurs nécessite beaucoup d'intrants et des terres rares. Le problème du terrain ne me semble pas essentiel, en particulier en Europe, en raison de la densité qui caractérise ce type d'installation : par exemple, une centaine d'hectares suffit pour un centre d'1 gigawatt – bien sûr, à l'échelle d'une ville, c'est important, mais ça l'est moins à l'échelle du continent.

Les électrons, en revanche, représentent un réel enjeu. Puisque 70 % de l'électricité produite en France est d'origine nucléaire, l'empreinte carbone de l'électron est très faible. Dès lors – et puisque nous avons besoin de tokens –, il est nettement préférable de construire en France qu'ailleurs, par exemple au Texas, car on contribue ainsi à réduire le réchauffement de l'atmosphère.

S'agissant de l'eau – un autre intrant –, des progrès technologiques importants ont été réalisés pour récupérer la chaleur fatale, permettre un refroidissement plus efficace et faire en sorte que les électrons soient majoritairement utilisés pour les calculs.

Par ailleurs, pour avoir consulté les documents, il me semble que tout a été mis en œuvre pour préserver la biodiversité.

Les enjeux environnementaux sont importants pour nous. Certes, d'un point de vue écologique, une forêt est toujours préférable à un data center mais un data center est important d'un point de vue économique. Nous sommes conscients qu'il a des effets sur l'environnement. Il faut donc aussi penser l'intelligence artificielle d'un point de vue écologique. Si l'on considère les tokens comme, d'une certaine manière, une ressource naturelle, il faut envisager les data centers comme des mines.

Nous devons construire des data centers sur notre territoire, sinon d'autres problèmes se poseront. Plus nous les construirons nous-mêmes, plus nous aurons voix au chapitre en ce qui concerne les externalités. Si nous nous contentons d'importer ce que d'autres nous proposent, nous serons contraints d'accepter aussi leurs critères – qui sont bien différents des nôtres – en matière de biodiversité et de réchauffement climatique. Il est absolument essentiel d'internaliser la production, ce qui suppose de faire des compromis – la France a accepté, de la même manière, d'ouvrir des mines de lithium.

Ce qui serait vraiment dommage, ce serait de construire des data centers dont la valeur économique échapperait pour l'essentiel à la France. Quitte à subir des externalités, bâtissons des projets qui nous permettront d'investir dans la R&D et qui auront des effets bénéfiques pour notre pays.

Mme Audrey Herblin-Stoop, directrice de la communication et des affaires publiques de Mistral AI. J'ajoute que Mistral AI est la première entreprise d'intelligence artificielle qui ait procédé à une analyse du cycle de vie de ses modèles. Celle-ci s'est déroulée l'an dernier, avec l'Ademe, l'Agence de la transition écologique, et Carbone 4, et nous nous apprêtons à recommencer pour analyser nos nouveaux modèles. L'objectif est d'offrir davantage de transparence et de proposer une méthodologie. Une telle initiative peut inspirer d'autres entreprises et permet en tout cas à nos clients et aux pouvoirs publics de s'appuyer sur

des éléments concrets et fiables pour opérer des choix. Elle pourra ainsi encourager une forme de standardisation.

M. Arthur Mensch. Nous entraînons la nouvelle génération de modèles en France. Nous sommes d'ailleurs fiers de posséder le plus gros cluster de calcul de notre pays car c'est un important facteur de réduction de l'empreinte carbone.

Par ailleurs, contrairement à certains concurrents, nous préférons réfléchir aux cycles de vie, donc avoir une approche qui intègre toutes les dimensions du projet, plutôt que de prévoir des compensations. C'est un peu trop facile, lorsque son bilan carbone est très élevé, d'atteindre un équilibre grâce à des quotas. Il faut offrir une transparence projet par projet plutôt que de vouloir proposer une transparence globale.

Mme Isabelle Rauch, présidente. Vous avez évoqué, dans vos propos liminaires, la lourdeur de la réglementation. Qu'entendez-vous par là et quelles pistes d'amélioration identifiez-vous ? Le recours à l'analyse du cycle de vie, que vous avez mentionné, ne pourrait-il pas, par exemple, devenir un critère pris en compte par les pouvoirs publics au moment de lancer un projet lié à l'intelligence artificielle ?

Vous avez également expliqué que la fragmentation du marché pouvait constituer une chance plutôt qu'un frein, avant de tempérer immédiatement cette affirmation. Qu'en est-il exactement ?

M. Arthur Mensch. Un des problèmes de l'Europe en matière réglementaire, c'est que, dès que l'on entre dans un pays européen, on doit ouvrir une nouvelle entité, comprendre un nouveau régime de stock-options, une nouvelle réglementation du travail… Par conséquent, il faut très vite une équipe sur place. J'ai moi-même signé des centaines de documents et ouvert des dizaines de comptes en banque pour ouvrir des entités dans une dizaine de pays d'Europe. Faute de réglementation unifiée, de droit social unifié, il faut s'adapter à chaque pays.

Il en va de même en matière de fiscalité, ce qui pose problème lorsqu'on veut instaurer des systèmes d'incitation des employés qui soient intéressants. On procède donc à des petits ajustements, on essaie de faire au mieux mais la diversité des systèmes fiscaux d'un pays à l'autre est assez catastrophique. Cette absence de marché unifié constitue la principale lourdeur du système.

On est aussi confronté à un empilement de réglementations, plus ou moins cohérentes entre elles : le RGPD, le règlement général sur la protection des données, les lois sur le copyright et le *text and data mining*, ou encore l'AI Act, le règlement européen sur l'IA, qui entrera en vigueur en août. Tous ces textes traitent plus ou moins du même sujet – les données, notamment personnelles, disponibles sur internet – sans dire exactement la même chose, ce qui nous oblige à nous documenter, et chacun des vingt-sept pays européens dispose de son propre organe, plus ou moins zélé, chargé de les faire appliquer.

À l'arrivée, l'équipe de *compliance* chez Mistral compte pas moins de cinq personnes. Pour nous, c'est jouable, parce que nous sommes suffisamment gros, et que ces dépenses sont incorporées dans les frais généraux – même si c'est pénible. En revanche, un entrepreneur qui veut se lancer aura peur. Face à cette situation et à l'ambiance générale dans notre pays – où l'on s'est réjoui, en 2024, d'avoir réglementé encore davantage, avec l'AI Act –, il aura envie de partir aux États-Unis. D'ailleurs, c'est ce qui se passe : nous perdons de nombreux entrepreneurs. Des investisseurs pensent que l'Europe a perdu parce qu'elle réglemente trop.

Les États-Unis ne se privent pas de s'emparer du récit selon lequel l'Europe serait perdante à cause des ronds-de-cuir qui réglementent à Bruxelles parce qu'ils sont incapables d'innover. Ce récit est d'autant plus destructeur qu'il est intériorisé par les Européens – on peut parler d'une forme de colonialisme. Il faut renverser ce récit, ce qui suppose de prévoir des réglementations plus simples et unifiées afin que nous puissions aller plus vite.

La fragmentation du marché n'est pas un atout. C'est plutôt un inconvénient parce qu'elle nécessite de composer des équipes différentes, qui vendent de façon différente selon les endroits.

Il faut savoir qu'il existe environ soixante opérateurs de télécommunications en Europe alors que les États-Unis en comptent seulement trois – le budget de chacun étant donc vingt fois plus élevé. Dès lors, lorsqu'ils investissent dans l'intelligence artificielle, ils peuvent mettre beaucoup d'argent sur la table. Voilà pourquoi la technologie américaine est partout.

Les start-up vendent, dans un premier temps, aux autres start-up – car la course à la réussite, dans sa version *venture capitalist*, fonctionne de façon circulaire – puis aux autres entreprises américaines, qui sont énormes et peuvent rapidement acheter. Une fois qu'elles sont ainsi passées à l'échelle, elles exportent vers l'Europe, où aucun acteur n'a émergé à temps, et rachètent les quelques boîtes qui ont développé des compétences, ce qui leur permet de se consolider. C'est ainsi qu'elles absorbent tout le marché.

Entre les pays européens, cela ne se passe pas comme ça : la viscosité est beaucoup plus forte car les entreprises sont plus petites. En outre, il faut parler une vingtaine de langues et maîtriser vingt-sept réglementations différentes.

Cette fragmentation n'est donc plutôt pas du tout une chance. Il y a un aspect positif : c'est que les entreprises sont plus ouvertes au partenariat, ont des approches plus constructives, d'autant plus que nous avons la chance de compter de nombreux champions, en particulier dans le secteur de l'industrie. Et comme les Américains ne peuvent pas vraiment répliquer ces approches partenariales, ces revenus sont plus facilement défendables. La viscosité, qui joue au départ en notre défaveur, peut donc tourner à notre avantage, face à un concurrent qui veut s'emparer d'un marché. Cependant, globalement, il serait préférable que le marché européen soit bien moins fragmenté.

J'en viens enfin à l'idée selon laquelle la réglementation pourrait représenter un outil de défense contre les Américains. Certes, l'intention est bonne – favoriser les start-up européennes – mais dans les faits, nous n'avons jamais réussi à atteindre notre but. La raison est très simple : la réglementation entraîne un surcoût qui est surmontable uniquement par des entreprises assez solides. Donc on se retrouve à importer des acteurs américains. En technologie plus qu'ailleurs, l'existence d'une réglementation favorise les gros. Comme on dit, l'enfer est pavé de bonnes intentions.

J'indique une piste d'amélioration : il faudrait que la demande publique soit recyclée pour faire de la R&D sur le territoire européen – cela vaut aussi bien pour la défense que pour les autres secteurs publics. C'est nécessaire, il faut une prise de conscience politique sur ce point. C'est d'ailleurs ce que les Américains font avec succès depuis les années 1940. Faut-il faire évoluer la réglementation dans ce sens ? Peut-être mais nous devons veiller à maintenir une concurrence très élevée en Europe. Si l'on peut parler de planification puisqu'il est question de choisir où vont les dépenses de l'État, il ne faut pas pour autant préciser comment cet argent doit ensuite être utilisé.

En résumé, réglementer pour défendre, cela ne fonctionne pas.

Mme Isabelle Rauch, présidente. Je n'ai pas dit qu'il fallait réglementer pour défendre. Je demandais s'il ne serait pas intéressant que la réglementation prévoie que les opérateurs publics prennent en considération l'analyse du cycle de vie plutôt qu'un autre critère.

M. Arthur Mensch. Sur ce point, je suis d'accord. En revanche, ce n'est pas parce qu'on ajoutera des règles que les Européens s'en sortiront mieux. Car, lorsque de nouvelles règles apparaissent, il faut avoir recours à des lobbys pour faire en sorte que ces règles s'appliquent de la manière la plus favorable. Or les acteurs américains sont puissants et disposent de beaucoup plus de lobbyistes à Bruxelles que nous ; ce sont eux qui gagnent la partie.

Mme Cyrielle Chatelain, rapporteure. Vous avez cité des infrastructures de 1 gigawatt. Est-ce que cela signifie que celles dont la puissance installée est inférieure ne sont plus adaptées ?

M. Arthur Mensch. Il faut avoir à l'esprit l'idée du réseau plutôt que de parler, comme on l'a beaucoup fait à propos de l'intelligence artificielle, de giga-usines. Il est question de services numériques qui ont besoin d'une puissance électrique qui est distribuée. Il n'est donc pas forcément nécessaire de construire une infrastructure imposante. Une forte colocalisation n'est pas indispensable. On peut tout à fait construire plusieurs sites de 100 mégawatts.

En revanche, lorsqu'on implante un site d'1 gigawatt, par exemple Campus IA, on fait des économies d'échelle. Cela permet de déployer des clusters d'entraînement qui, eux, ont besoin d'être colocalisés. Ils mobilisent simultanément un très grand nombre de GPU et doivent donc être très connectés entre eux pour pouvoir communiquer, par conséquent il est préférable de les rapprocher géographiquement. En outre, l'empreinte environnementale est moins élevée : la taille d'une station de transformation reste à peu près la même, que sa puissance soit de 100 mégawatts ou d'1 gigawatt. Toutefois, en France, le réseau électrique est performant ; par conséquent, nous ne sommes pas obligés de construire uniquement des gros sites.

Par exemple, à Mistral AI, nous ne disposons pas uniquement de gros sites : nous implantons souvent des clusters de 40 mégawatts, nous en avons un de 25 mégawatts en Suède, nous avons aussi un projet de centre de 80 mégawatts en France pour l'an prochain. De toute façon, pour déployer un centre de 1 gigawatt, il faut beaucoup de temps. Nous atteindrons cette échelle seulement à horizon 2028 ou 2029.

Mme Cyrielle Chatelain, rapporteure. Vous avez parlé assez justement du risque de verrouillage si les data centers, une fois construits, sont principalement occupés par des acteurs extra-européens et qu'il n'est ensuite plus possible de récupérer la puissance qui a ainsi été captée. La réglementation ne pourrait-elle pas permettre d'empêcher ce phénomène ?

M. Arthur Mensch. Actuellement, la loi européenne ne le permet pas. L'accès aux électrons en Europe repose sur un principe fondamental de neutralité. Il existe bien sûr des volontés politiques mais le problème, c'est qu'il y a assez peu de centralisation en la matière. Ce n'est pas l'État qui décide d'un plan stratégique afin de déterminer à quel endroit il déploie une installation et à quel acteur il la loue. Cette compétence est distribuée entre les intercommunalités, qui ne comprennent pas forcément les enjeux supranationaux.

La location fonctionne de façon largement sous-optimale même si elle repose sur un principe économique : elle est proposée à ceux qui peuvent payer, ce qui suppose qu'ils aient de la demande. Mistral est par exemple en mesure de payer des sommes assez importantes ; et nous aimerions payer plus, mais il nous faudrait davantage de demande. Une fois que la demande est là, si l'on a une visibilité à long terme et que l'on sait qui achètera les tokens, on peut chercher des financements, lever de la dette et construire un projet plus rapidement que les autres. Or ce type de planification n'est pas d'actualité.

Mme Audrey Herblin-Stoop. Plusieurs projets de réglementation sont en cours de discussion à Bruxelles, notamment le Cloud and AI Development Act et le paquet souveraineté. Il y a là deux leviers d'action : premièrement la définition de ce qu'est un cloud souverain, deuxièmement la préférence européenne dans la commande publique, qui nous semble être une manière – certes imparfaite – de créer des opportunités pour les entreprises européennes.

La définition du cloud souverain doit être la plus juste et la plus pertinente possible. Le fait qu'une entité juridique soit vaguement établie dans un pays ne suffit pas : il faut s'assurer que le contrôle de l'entreprise n'est pas exercé par des entités étrangères et que la donnée stockée n'est à aucun moment soumise aux juridictions américaines notamment – je pense au Cloud Act.

M. Arthur Mensch. Il y a en effet un sujet d'extraterritorialité, mais aussi un sujet économique : où la R&D est-elle réinvestie ? Quand des sociétés européennes sont totalement contrôlées par des sociétés étrangères, tous les revenus repartent au siège. C'est cela qu'il faut regarder.

Mme Cyrielle Chatelain, rapporteure. D'où l'importance de ne pas attendre un verrouillage pour agir ! J'entends qu'il est possible d'accroître la capacité d'achat et le potentiel du marché ; mais je pense que les collectivités locales ont peu leur mot à dire dans les projets. Il me semble que les acteurs qui décident des espaces d'implantation sont soit les acteurs économiques et les investisseurs, soit l'État dans son rôle de pilote.

Avez-vous une idée de la part de la commande publique dans votre chiffre d'affaires ?

M. Arthur Mensch. De tête, je dirais que la commande publique globale représente 20 % de notre chiffre d'affaires, sur la partie logiciels, et la commande publique française 10 %. Nous ne cherchons pas à la faire beaucoup augmenter, mais tout croît.

Nous avons des contrats-cadres significatifs avec le Luxembourg par exemple, où nous faisons du déploiement en administration centrale. Nous ne le faisons pas du tout à ce type d'échelle en France mais beaucoup dans les territoires, avec la Caisse des dépôts, et dans le domaine de la défense avec le ministère des armées.

Les administrations centrales devraient voir l'intelligence artificielle générative comme une manière de créer la productivité dont elles ont besoin, étant donné la démographie européenne en particulier. Je pense que l'IA pourrait être plus largement adoptée. Cela serait bénéfique pour la qualité des services publics, pour les finances publiques ainsi que pour l'écosystème : la commande est recyclée en R&D et permet l'émergence d'autres fournisseurs européens, etc.

Mme Cyrielle Chatelain, rapporteure. Avez-vous connaissance d'études macroéconomiques sur les gains de productivité et les retours sur investissement générés par l'IA ? Nous avons posé la question à plusieurs reprises sans obtenir de réponse.

M. Arthur Mensch. Nous avons beaucoup d'études microéconomiques : nos clients nous disent qu'ils enregistrent des retours largement supérieurs à ce qu'ils dépensent en intelligence artificielle chez nous. Dans certains cas d'usage, ils font 50 millions de retour sur investissement parce que les machines fonctionnent mieux. On observe aussi des gains de productivité d'un facteur 5 dans des services clients ; autrement dit, pour effectuer une tâche, il faut 20 % du temps qu'il fallait avant l'IA. C'est donc très significatif.

Je pense qu'il n'y a pas encore beaucoup d'études macroéconomiques parce qu'on manque de recul : la délégation de tâche à un agent pendant toute la journée n'est possible que depuis six mois. Je le disais, avec 10 % de notre masse salariale consacrés à l'équipement de notre personnel, nous avons fait un gain de productivité de facteur 2 par rapport à il y a six mois. Ce n'est pas une étude macroéconomique, c'est vrai, mais on parle d'un gain de productivité pour un coût représentant 10 % à 20 % des opex (dépenses d'exploitation) ou de la masse salariale dans le monde.

Certes, la productivité ne sera pas multipliée par deux partout ; chez nous c'est assez facile car, compte tenu de la faible empreinte physique de notre activité, l'utilisation des agents suscite peu de frictions. Dans l'industrie lourde, c'est un peu plus difficile, parce qu'il faut tester des systèmes physiques. Mais pour 10 % de la masse salariale – sachant qu'en général, un client achète une technologie lorsque celle-ci ne lui prend pas plus de 50 % de la valeur –, il y a 20 % de gains de productivité. Il faut avoir en tête que ce ne sera pas forcément 20 % de croissance en plus : il y aura de la croissance mais aussi des destructions d'emplois – du moins une modification des emplois existants.

Je reviens à une approche macroéconomique : certains métiers disparaissent presque, et les gens s'adaptent assez vite. Il n'est pas exclu qu'il y ait une augmentation du chômage dans certains domaines et un déplacement de la valeur du travail vers le capital – lequel est, pour le moment, largement extra-européen.

La situation est, je pense, assez explosive : à cause de l'intelligence artificielle, il y a des destructions d'emplois – en tout cas, des modifications très rapides de la structure de l'emploi en Europe ; il y a ensuite des conflits d'usage de l'électricité – tout le monde en veut, mais il n'y en a pas assez –, et donc une inflation ; il y a enfin une explosion du déficit commercial qui, dans le domaine des services, sera multiplié par cinq dans les cinq prochaines années. Une situation comme celle-ci est révolutionnaire. Il faut la prendre en compte, en parler davantage et avoir un point de vue de plus long terme sur ce qu'elle implique pour l'économie, pour la société et pour les équilibres géopolitiques. Si on ne le fait pas assez rapidement, on n'aura plus aucun choix – une situation que l'on ne voudrait pas voir advenir.

Mme Cyrielle Chatelain, rapporteure. Quels sont les cas d'usage dans lesquels l'IA est la plus utilisée ?

Il y a par ailleurs une situation que l'on pourrait qualifier de bulle. En tout cas, face à une capacité de levée de fonds très importante des acteurs de l'intelligence artificielle, il y a un modèle économique dont certains économistes disent qu'il n'est pas stabilisé, s'agissant notamment des coûts d'investissement. Vous avez cité des chiffres qui donnent le vertige. Il y a donc de très fortes capacités d'investissement dans les infrastructures comme les data centers,

qui auront besoin d'une maintenance en continu. Dans le même temps, le coût de l'IA est très faible aujourd'hui pour les clients ; ce n'est peut-être pas votre modèle mais c'est en tout cas la promesse faite par les acteurs américains comme OpenAI. Compte tenu des montants colossaux des investissements et du coût de la maintenance des infrastructures, sera-t-il possible de conserver un prix modéré ? N'y a-t-il pas un risque de verrouillage ? Si les opérateurs augmentent fortement leurs prix, des acteurs économiques ayant revu leur organisation du travail pourraient se retrouver dans l'impossibilité de se dégager de ces nouveaux services numériques.

M. Arthur Mensch. Le premier cas d'usage est le développement logiciel. Chez Mistral, les ingénieurs n'écrivent plus de lignes de code. La façon de travailler, dans ce domaine, a profondément changé au cours des six derniers mois. Auparavant, les contributeurs individuels étaient des artisans qui écrivaient du code. Les gens aimaient cet artisanat ; j'en viens, et c'était mon cas. Aujourd'hui, ce ne sont plus des artisans mais des managers, des donneurs d'ordres : ils demandent à des agents d'écrire le code pour eux en leur décrivant des spécifications. C'est un changement assez profond, qui s'accompagne de gains de productivité plus ou moins importants selon la taille de l'équipe : quand vous êtes seul, vous pouvez aller dix ou vingt fois plus vite ; quand vous êtes cinq, les gains sont un peu moindres parce qu'il y a l'enjeu de la communication. Dans une très grosse entreprise, il faut lever les goulots d'étranglement organisationnels pour atteindre les gains de productivité dont on rêve.

Mme Cyrielle Chatelain, rapporteure. À combien de tokens équivaut un salarié écrivant des lignes de code ?

M. Arthur Mensch. Je compte plutôt en mégawatts. L'ordre de grandeur, c'est environ 1 kilowatt par salarié, soit un demi-GPU (processeur graphique) ou encore 10 000 euros par an par personne – ce qui représente au niveau de l'Europe, c'est-à-dire 400 millions de personnes, 8 trilliards.

Avec 10 000 euros par an, soit 30 euros par jour, vous générez 10 millions de tokens. Sachant qu'une ligne de code, c'est environ 100 tokens, cela fait environ 100 000 lignes de code. Mais quand vous demandez à un agent de faire du code, il fait aussi autre chose. Il va beaucoup réfléchir, et générer des tokens pour se demander : dois-je faire ceci ou cela ? Dois-je faire tourner ce test ? Est-ce que ça marche ? Dois-je le réécrire ? Le code qu'il écrit à la fin est court mais il a fait beaucoup d'expériences – celles que le développeur logiciel faisait auparavant.

Quant à la bulle, on a dit pendant longtemps qu'il y en avait une, mais ce n'est pas le cas. Une bulle, c'est quand la demande est surestimée. Or le problème aujourd'hui est plutôt celui de l'offre. Beaucoup de gens viennent nous voir en nous disant qu'ils ont besoin de plus de tokens que ce qu'ils avaient prévu. Il n'y a pas assez de logique, donc de *chips* ; pas assez de mémoire, de cartes mères, de disques durs ; pas assez d'hélium pour les semi-conducteurs ; pas assez d'électrons. Toute la chaîne des semi-conducteurs est mise sous pression par l'intelligence artificielle. Le manque est donc plutôt du côté de l'offre que de celui de la demande.

Pour que le modèle économique fonctionne, il faut qu'une entreprise de technologie qui investit 50 milliards pour bâtir 1 gigawatt puisse repartir avec 100 milliards ; ce qui signifie que vous devez créer 200 milliards de valeur pour votre client. Est-on aujourd'hui à ces équilibres ? Pas complètement. On met évidemment de l'argent sur la table au début parce qu'on sait qu'une fois qu'on a capturé la demande et l'offre, on peut monter les prix et les

marges. Le risque est donc celui que j'évoquais tout à l'heure : si vous ne déployez pas suffisamment vite, vous n'allez pas capturer suffisamment de parts de marché. Des oligopoles vont se former ; eux vont retrouver leurs billes. Toute l'économie américaine ne met pas 1 trilliard sur la table chaque année si elle ne s'imagine pas repartir avec 2 trilliards. Je ne m'inquiète pas pour eux ; je m'inquiète plutôt pour nous.

M. Arnaud Saint-Martin (LFI-NFP). J'aimerais d'abord vous interroger sur l'économie des travailleurs du clic, auxquels des enquêtes journalistiques ont été consacrées récemment. D'où viennent les annotateurs ? Comment entraînent-ils les modèles ? Combien sont-ils payés et au travers de quelles entreprises ?

Vous avez évoqué des modèles qui ont une certaine politique, qui actionnent certaines choses. On connaît les polémiques autour de X, notamment. Pourriez-vous nous apporter des précisions à ce sujet ? Sur quelle politique vous appuyez-vous pour générer votre business ? Quels sont les verrous éthiques qui cadrent votre activité ?

Je m'interroge enfin sur la soutenabilité financière de votre entreprise car je pense quant à moi qu'il y a une bulle ; une réflexion est en cours à ce sujet en économie politique du numérique. Il y a des effets d'entraînement massifs, avec des entreprises qui surcapitalisent et des valorisations financières indécentes. Comment vous protégez-vous d'une éventuelle implosion ? Quand bien même vous faites état d'une demande pléthorique, comment vous prémunissez-vous contre le possible effondrement d'une économie qui, pour l'instant, fonctionne largement sur la promesse de retours sur investissement qui restent assez théoriques ?

M. Arthur Mensch. Durant les premières années de l'IA générative, les modèles étaient largement entraînés avec de l'annotation humaine – des gens qui donnent leurs préférences, etc. Cela nous a permis d'arriver collectivement à un niveau de modèle qui parvient à engager des conversations.

Au fur et à mesure que la technologie s'est développée, on a eu besoin d'annotateurs de plus en plus qualifiés. Aujourd'hui, les annotateurs chez Mistral sont des gens qui ont des thèses ; ce sont des gens à qui on demande de résoudre des problèmes de sécurité dans un répertoire de code, ou des problèmes de physique des particules. Souvent ce ne sont pas des gens, d'ailleurs, mais des environnements, car ce n'est pas de signaux humains qu'on a besoin, mais de signaux d'environnement : vous demandez à l'agent de faire des modifications sur le code, puis vous mettez celui-ci dans un environnement qui va l'exécuter et vous dire s'il fonctionne ou pas. Vous pouvez faire ça aussi avec de la 3D ou avec des fichiers Excel.

Le sujet n'est donc plus tant celui des annotateurs humains que celui des développeurs de logiciels qui créent les environnements pouvant être utilisés largement à l'échelle, et donner un signal au modèle. Pour nous, il s'agit maintenant de distiller des environnements, de comprendre comment un environnement de code fonctionne, de prendre le signal et de le ramener vers le modèle. Imaginons un outil de simulation numérique : vous le faites tourner de nombreuses fois, vous prenez le signal à la sortie et vous l'utilisez pour que le modèle comprenne un peu mieux la physique que vous êtes en train de simuler.

Le sujet, c'est de ramener le signal de l'environnement vers le modèle, ce qui implique un travail beaucoup moins intense. Le changement s'est fait en 2023-2024.

Pour l'annotation, nous avions toujours travaillé avec des fournisseurs qui nous apportaient des garanties s'agissant des salaires, notamment.

M. Arnaud Saint-Martin (LFI-NFP). Où étaient-ils implantés ?

M. Arthur Mensch. Aux États-Unis, en Europe… Comme nous avons besoin de travailleurs qualifiés et parlant un certain nombre de langues européennes, nous sommes obligés de nous tourner plutôt vers l'Europe. Pour nos activités de robotique, nous travaillons avec Madagascar et nous avons mis en place un cadre de travail pour faire en sorte que les gens soient bien rémunérés, entre autres.

J'en viens à votre question sur la politique. Le modèle répond comme il veut à la question que vous lui posez. Il y a donc un enjeu en quelque sorte éditorial : déterminer ce qu'il va vous répondre quand vous lui posez une question orientée. Notre stratégie en la matière consiste à suivre les instructions et à avoir le moins de biais possible. C'est un sujet scientifique et technique, que nous ne prétendons pas résoudre complètement. La question éditoriale se pose vraiment quand vous faites des chatbots pour le grand public, ce qui n'est pas vraiment notre sujet.

En revanche, les systèmes qui écrivent et exécutent du code peuvent choisir différentes bibliothèques de code et faire des choix plus ou moins sécurisés. Nous faisons donc en permanence un travail de sécurisation du code généré : nos clients nous le demandent. Il y a un vrai enjeu de sécurité, à partir du moment où les modèles sont exécutés pour faire des actions irréversibles, pour exécuter du code et éventuellement pour actionner certaines chaînes logistiques, par exemple. Nous avons une équipe dédiée à la sécurité et à la sûreté : c'est un sujet très important pour nous. Nous le traitons aussi en donnant des outils à nos clients afin qu'ils soient capables d'observer et de vérifier que les applications font bien ce qu'ils ont envie qu'elles fassent.

Mme Cyrielle Chatelain, rapporteure. J'ai une dernière question assez simple : demain, les Américains pourraient-ils racheter Mistral AI ?

M. Arthur Mensch. Nous avons avec nous quelques sociétés de VC (*venture capital*), qui représentent moins de 30 % de notre capital. Nous aurions été ravis de prendre des Européens, mais il n'y en avait pas ! On y perd sur le plan économique : la partie *general partner* du retour sur investissement dans Mistral repart aux États-Unis, par le biais des fonds américains. Quant à la partie *limited partner* (LP), elle revient partiellement en Europe, car les Américains lèvent dans le monde entier. Donc on perd du retour, parce qu'on n'a pas tout l'écosystème. C'est dommage mais c'est comme ça.

Notre objectif est de rester indépendants. Il arrive évidemment qu'on nous demande si l'on peut nous racheter : nous répondons non, car notre mission est de rester indépendants, d'aller vers une cotation, à terme, et de faire en sorte de fournir une alternative.

Si vous réussissez, vous ne vous faites pas racheter. Si vous vous faites racheter c'est que, d'une certaine manière, vous avez raté. En tout cas, c'est comme ça qu'il faudrait réfléchir. De nombreux entrepreneurs, en Europe, réfléchissent à des stratégies d'exit qui consistent plutôt en des rachats par des sociétés américaines. Ces stratégies sont perdantes – enfin, vraiment dommageables sur le plan macroéconomique. Ce qu'il faut, c'est créer des champions européens indépendants. C'est à ça que nous travaillons ! Nous n'avons pas encore gagné, mais nous y travaillons.

Mme Isabelle Rauch, présidente. Souhaitez-vous ajouter quelque chose ?

M. Arthur Mensch. Non, je crois que nous avons couvert beaucoup de sujets !

Je redis qu'il ne faut surtout pas distinguer la question du cloud du reste de l'IA. Nous ne devons pas nous dire qu'on a perdu la bataille du cloud. Au contraire, il faut remonter ; et pour cela, il faut passer par la valeur ajoutée importante, celle qui fait la croissance : l'intelligence artificielle. À partir de là, tout le reste du cloud peut suivre. Il faut concentrer les efforts là où l'on est fort. Il se trouve que nous avons la chance d'être relativement forts en intelligence artificielle et d'avoir de la capacité électrique. Si l'on combine les deux, on peut retrouver une part de marché soutenable. Il faut absolument le faire ; sinon, nous allons devenir un État vassal.

Mme Isabelle Rauch, présidente. Merci d'être venu devant nous.

M. Arthur Mensch. Merci beaucoup de m'avoir écouté.

Mme Isabelle Rauch, présidente. Nous avons bien compris le message final, et ce que l'Assemblée nationale va devoir faire.

La séance s'achève à dix-sept heures trente.

———

Membres présents ou excusés

Présents. – M. Nicolas Bonnet, Mme Cyrielle Chatelain, M. Philippe Latombe, Mme Isabelle Rauch, M. Arnaud Saint-Martin