Roger Bautier
LabSIC – Université Sorbonne Paris Nord

Méconnaissance et autorité de l'IA générative

Fortement médiatisée, l'apparition de ChatGPT en novembre 2022 a attiré l'attention sur les développements de l'intelligence artificielle générative en mettant en lumière les performances des transformeurs pré-entraînés, qui offrent des potentialités à la fois diverses et massives pour le traitement de l'information, l'invention intellectuelle ou la production artistique. Cette médiatisation a provoqué des enthousiasmes et des condamnations qui ne doivent pas cacher le bouleversement culturel suscité, dont témoignent, par exemple, la leçon inaugurale en 2023 du cours de Benoît Sagot au Collège de France (Apprendre les langues aux machines) ou bien l'attribution en 2024 du prix Nobel de physique à John Hopfield et Jeoffrey Hinton, qui avaient ouvert la voie aux réseaux de neurones artificiels dans les années 80, des réseaux qui constituent une des composantes fondamentales des systèmes d'intelligence artificielle générative.

Les systèmes récents, ceux que l'on appelle transformeurs, font appel, en effet, à de nouveaux réseaux de neurones en couches, particulièrement performants. Au-delà de l'accomplissement de tâches linguistiques, les applications sont diverses et impressionnantes et les méthodes utilisées présentent des avantages incontestables, notamment en matière de quantité de données traitées et produites, que celles-ci soient unimodales (comme du texte) ou multimodales (textes, images, sons). Ces systèmes ont montré, en particulier, leur aptitude à générer des suites de mots acceptables, sans que soit concernée leur relation au monde : ces suites peuvent être conformes aux contraintes de la langue et, plus largement aux aspects rhétoriques du discours. L'objectif est de leur fournir une connaissance par l'intermédiaire d'un entraînement sur une première tâche, puis de leur faire réaliser une nouvelle tâche en utilisant la connaissance acquise auparavant. Lorsque le matériau est langagier, le modèle de langue, tenant compte de dépendances à longue distance, est entraîné sur un corpus à retrouver un élément qui a été masqué ou bien à prédire un élément qui suit un ensemble d'éléments connus. Le modèle va pouvoir effectuer un apprentissage auto-supervisé lui permettant d'effectuer une actualisation progressive du poids des connexions entre éléments.

Ces systèmes génératifs peuvent, grâce à ce type d'apprentissage, « répondre » aux requêtes (*prompts*) des humains en donnant des résultats spectaculaires, souvent très adéquats, mais, bien évidemment, ils n'ont aucune expérience du monde : ils manipulent des objets sémiotiques dont ils ne peuvent pas considérer la dimension référentielle. Il faut souligner que le fonctionnement des transformeurs n'est pas totalement élucidé et que les performances auxquelles ils parviennent paraissent donc à la fois spectaculaires et assez mystérieuses. Le niveau de performance semble fortement lié à l'augmentation du nombre de paramètres dans le processus : on peut, en conséquence, envisager que la progression de quelques milliards de paramètres à des centaines de milliards corresponde à ce que la physique statistique appelle une transition de phase. Si tel était le cas, il resterait à connaître les modalités précises de cette transition. Cela ne peut que renforcer la pertinence des interrogations concernant le processus par lequel les résultats sont engendrés, étant donné, notamment, l'écart entre la simplicité, toute relative, de ses principes et la complexité des tâches effectivement réalisées, qui paraît impliquer une élaboration dépassant de beaucoup la seule recopie d'une partie de l'information (gigantesque) disponible sur le web.

Un des dangers les plus évidents réside dans le fait que la machine, en produisant des objets sémiotiques qui ressemblent – quelquefois étonnamment – à ceux produits par les humains, peut obtenir, aux yeux de certains d'entre eux, un statut d'autorité indiscutable, même si celle-ci est totalement infondée. De nombreux commentateurs ont dit que ChatGPT ressemble à un rhéteur de l'Antiquité, ce qui dénote sans doute une appréciation erronée, à la fois, de cet agent conversationnel et de la rhétorique antique. Cependant, il faut bien considérer, d'une part, que les productions des intelligences artificielles génératives, par le fait qu'elles sont d'origine probabiliste, ne peuvent être considérées que comme vraisemblables (on sait qu'elles peuvent même « halluciner ») et, d'autre part, que les humains, dans certaines circonstances, peuvent être amenés à manifester les mêmes qualités et les mêmes défauts. Toujours est-il que l'usage des systèmes génératifs ne peut que susciter la réflexion sur ses conséquences en matière d'accès aux informations et aux connaissances.

Ainsi, on ne peut que remarquer que les données sur lesquelles ils s'entraînent sont fournies à la fois par les humains et par les machines, d'où le risque d'un fonctionnement circulaire (une sorte d'autophagie) qui marginaliserait aussi bien les experts de différents domaines que les informaticiens, en faisant appel, bien davantage, aux productions de multiples individus et aux performances des réseaux de neurones artificiels. De plus, par leur générativité, ces dispositifs d'intelligence artificielle sont en position de concourir avec les objets sémiotiques d'origine proprement humaine, dont la diffusion dans l'espace public est déjà massive. Cette concurrence pouvant se réaliser de manière invisible si les différences entre les deux types ne sont pas perceptibles, il est tout à fait envisageable qu'elle puisse se traduire par une suprématie quantitative des productions machiniques. On comprend dès lors que le développement de l'IA générative impose un examen attentif de ses implications dans toutes sortes de domaines : dans le domaine réglementaire en général, comme dans celui de la protection des données personnelles ou dans celui de son impact écologique.

Dans cette perspective, ne doit-on pas plus particulièrement promouvoir une culture de la transparence ? En effet, il peut exister une opacité intentionnelle (de la part d'une entreprise ou de l'État), éducationnelle (par manque de formation de l'usager) ou opératoire (issue de la complexité des algorithmes). Tout autant, une culture de la responsabilité est requise, puisque la machine n'est pas un agent moral et que la responsabilité est partagée entre l'usager, le concepteur, l'entraîneur et le fabricant. De même, c'est une culture de l'utilisation des données qui doit être défendue : par exemple, ce n'est pas parce que des données sont publiques qu'elles ne sont pas personnelles et il n'est pas interdit d'entraîner des IA sur des données personnelles, mais sous contrôle. Plus largement, d'autres impératifs encore mériteraient d'être mis en lumière. Il s'agirait d'abord d'insister sur l'intérêt d'une culture du débat : la prospective est sans doute nécessaire, avec une réflexion généralisée sur l'impact des IA génératives (notamment le risque de boucle autophagique). Il s'agirait aussi de favoriser la diffusion d'une culture scientifique car cette réflexion ne peut avoir lieu que si les participants possèdent le minimum de culture scientifique leur permettant de comprendre les évolutions techniques. Enfin, c'est une culture politique qui, souvent, doit être renforcée, étant donné que ces évolutions ne sont pas sans lien avec des idéologies qui apparaissent, implicitement ou explicitement, dans les travaux des concepteurs