Settimana della Scienza,
Tecnologia ed innovazione Italia-Cina

Responsabilità e sostenibilità

nei sistemi di intelligenza artificiale generativa multimodale

Rita Cucchiara
Università di Modena e Reggio Emilia, Italia

Ministri, autorità, colleghi, vi ringrazio molto per l'opportunità di presentare attività e risultati della ricerca scientifica di frontiera nell' intelligenza artificiale e di proporre una breve discussione sulla responsabilità e la sostenibilità nei sistemi di intelligenza artificiale generativa multimodale.

I **sistemi generativi multimodali**, che in ambito scientifico vengono chiamati sistemi *fondazionali*, sono in grado di ricevere in ingresso testo, immagini, audio e video, di riprodurre e generare contenuti in tutte le modalità e anche di interrogare banche dati esterne, sempre in molte e differente modalità. Le attività di ricerca in questo ambito sono nate solo nell'ultima decade, grazie a una vasta collaborazione internazionale tra le aree di ricerca nella visione artificiale, nell' NLP (Elaborazione del Linguaggio naturale) e dell'apprendimento automatico. Dal 2021, sono stati proposti i primi sistemi cosiddetti *contrastivi*, addestrati con coppie di due modalità diverse come testo ed immagini; sistemi come CLIP sono nati inizialmente da una proposta statunitense e poi si sono evoluti in sistemi proposti in tutto il mondo, in Cina e in Europa, tendenzialmente come open source, preziosissimi per la ricerca.

In questo ambito dei sistemi generativi multimodali, l'Europa e l'Italia stanno lavorando con impegno; cito rapidamente alcuni progetti europei attivi a cui il mio laboratorio è stato finanziato, come ELLIOT, ELIAS, ELSA della rete ELLIS (Europeal Lab of Learning and Intelligent Systems) e MINERVA, nati proprio sui temi dell'IA generativa, e in particolare il progetto ITA4IA, la nostra iFactory italiana coordinata dal centro di supercalcolo del CINECA. Cito inoltre molti progetti italiani finanziati recentemente da Next Generation Italia, cioè dal PNRR, come il partenariato esteso FAIR, coordinato dal CNR, ITSER e FIT4MedRob, sempre coordinati dal CNR, e progetti FIS come TRAMIS sulla robotica, in cui il mio Ateneo è coinvolto.

I sistemi generativi multimodali, sebbene proposti nelle conferenze internazionali solo negli ultimi quattro anni, hanno già **moltissimi campi di applicazione**:, per esempio, la generazione di immagini modificate e utili per la robotica medica, per aiutare operazioni di trapianto, o per l'interazione persona-robot (come nell'esempio di robot capaci di interagire in una casa e raccontare ciò che stanno vedendo, anche muovendosi in un ambiente sconosciuto); oppure, più naturalmente, la generazione di immagini sintetiche o modificate per applicazioni di turismo, o ancora la creazione di immagini nell'ambito dell'automotive, capaci di riconoscere l'auto, la strada e la scena e anche di comprendere la quantità di traffico percepita da un'auto; per finire, in ambito e-commerce, con la possibilità di modificare in modo responsabile e sostenibile le immagini.

Questi modelli e queste possibili applicazioni rientrano nel mirino dei ***concetti di responsabilità e di sostenibilità*** nati in Europa negli ultimi anni e recepiti anche dal nostro Paese. Sappiamo che fin dal 2019 l'OCSE ha definito alcuni fondamenti per la responsabilità dell'intelligenza artificiale, che nell'ambito della Comunità Europea si associano strettamente ai concetti di fiducia e affidabilità, basati su tre pilastri importanti: i sistemi devono essere *legali, etici e robusti*. Questi principi si sono poi declinati nella legge di regolamentazione sull'intelligenza artificiale della Comunità Europea, l'AI Act, in sette requisiti ormai noti: i sistemi devono essere supervisionati da un essere umano; devono essere tecnicamente robusti e sicuri; devono gestire in modo corretto la privacy e i dati; devono essere trasparenti ed evidenziare quando i

contenuti sono generati dalla macchina e non dal lavoro umano; devono essere non discriminatori ed equi e gestire correttamente la diversità; devono essere sostenibili dal punto di vista sociale e ambientale; infine, è importante comprendere i concetti di accountability, ossia di chi sia la responsabilità quando il sistema commette errori.

Questi concetti di affidabilità, responsabilità e sostenibilità sono stati recepiti da diversi anni anche dall'Italia, fin dal PNR, cioè il *Programma Nazionale della Ricerca* 2021-2027, quando fu creato un work program sulla intelligenza artificiale che ho avuto l'onore di coordinare, in cui si indicava espressamente che i sistemi proposti dalla ricerca italiana dovessero essere etici sin dai progetti, con controllo umano in ogni stadio, affidabili e degni di fiducia. Lo stesso concetto è stato ribadito nell'ultimo documento di *Strategia Italiana per l'Intelligenza Artificiale 2024-2026*, dove si utilizzano esattamente gli stessi principi. Infine, poche settimane fa, nel settembre del 2025, l'Italia si *è dotata di una prima legge sulla* IA e sulla responsabilità dell'intelligenza artificiale, ribadendo gli stessi principi di affidabilità e sostenibilità e aggiungendo concreti controlli e salvaguardie soprattutto per alcuni ambiti, come l'impiego e la salute, nonché un intero contesto per la governance dei sistemi, per gli investimenti e per la responsabilità di chi progetta e usa tali sistemi.

Se questo è il contesto in cui l'Europa e l'Italia si vogliono collocare, ciò non significa che non siano stati fatti investimenti e importanti risultati nella ricerca scientifica, rivolta non tanto ai sistemi attuali quanto a quelli delle prossime generazioni. In particolare, vorrei mostrare tre diverse direzioni della ricerca: che sia *eticamente responsabile*, *sostenibile dal punto di vista computazionale* e *affidabile nella direzione di ridurre le allucinazioni*. Questi risultati sono stati recentemente proposti nei più importanti consessi internazionali quali ICCV, ECCV, CVPR, dove Europa, Cina e Stati Uniti si ritrovano annualmente per discutere la ricerca di frontiera.

Il primo tema riguarda sistemi responsabili, tecnicamente ed eticamente, e in particolare la possibilità di utilizzare modelli già esistenti e **riaddestrarli** attraverso un approccio chiamato *unlearning*, cioè disimparare concetti interni forse appresi, come violenza, tossicità, nudità o concetti non sicuri, sia nella generazione sia nella ricerca di immagini simili. Solo per fare un esempio con Safe-CLIP: se a un sistema viene fatta una richiesta violenta o non sicura, come "disegnami un gruppo di persone in un campo di battaglia con costruzioni sullo sfondo", il sistema può utilizzare "campo di battaglia" come metafora e mostrare uno stadio con una partita; oppure, se si chiede "genera un'immagine con un bambino con una pistola in mano", il sistema non riconosce il concetto di pistola e disegna un bambino con un giocattolo in mano. Questi modelli, parzialmente sviluppati in Italia insieme all'Università di Amsterdam in un progetto europeo e addestrati grazie al sistema di supercalcolo del CINECA, saranno presto disponibili in modalità open source, ad esempio per applicazioni educative o dove sono coinvolti bambini, in cui non è necessario né opportuno utilizzare concetti di violenza o nudità nell'ambito dell'IA.

Un secondo risultato che vi propongo, sviluppato in ambito italiano dalla mia università ma in collaborazione con la comunità europea, riguarda nuovi sistemi **sostenibili**, in questo caso chiamati self-*reflective*, **autoriflessivi**, che sono in grado di "sapere di non sapere", come diceva Socrate, cioè di capire, dal punto di vista della sostenibilità energetica, quando conviene interrogare banche dati esterne e quanto i risultati ottenuti possano essere utili. Tra l'altro, questo sistema impiega sorgenti aperte ed è nato anche in collaborazione con la Cina, utilizzando come base di lavoro sistemi multimodali come Qwen VL sviluppati recentemente dai centri di ricerca Alibaba e disponibili in sorgenti aperte.

Utilizzando questi nuovi modelli e questo nuovo approccio di addestramento sostenibile dal punto di vista energetico, è possibile ottenere risultati di precisione e di altissima affidabilità, soprattutto in richieste particolari come quelle degli esempi: per esempio, se si chiede "dove si trova questo giardino?", sistemi di questo genere sono in grado di interrogare, solo se serve, sorgenti esterne come Wikipedia e fornire il risultato corretto senza sprecare troppa energia nel fine tuning e nel riaddestramento.

Come ultimo esempio di risultato interessante della ricerca, che potrà essere alla base dei prossimi sistemi aperti e disponibili anche per le aziende italiane, europee e di tutto il mondo, cito nuove soluzioni nate in ambito europeo, dall'Università di Trento e UNIMORE, per valutare il grado di correttezza, o al contrario il **grado di allucinazione**, dei sistemi nel momento in cui generano immagini. Questo è un problema molto importante nell'intelligenza artificiale eticamente responsabile, perché conoscere e misurare il grado di errore ci permette, da una parte, di risparmiare tempo ed energia e, dall'altra, di essere più sicuri del grado di affidabilità dei sistemi. Grazie all'uso massivo delle CPU e delle GPU del supercalcolo italiano abbiamo potuto addestrare questi modelli e ottenere risultati come quelli che vedete: il sistema non solo, come ormai molti, è in grado di modificare immagini, per esempio cambiando il colore di una parte o una bandiera, o in ambito medico rimuovendo una parte non interessante di un'immagine istologica, ma anche di valutare, con un ragionamento tramite un agente, se la modifica è coerente con il prompt dato. Nell'esempio della bandiera, il sistema riconosce che, sebbene sia stato aggiunto un simbolo tipico canadese, non si tratta della bandiera canadese, mentre nell'altro caso è stata correttamente modificata solo la parte richiesta nel vetrino istologico. Questi sono nuovi tool, al momento solo prototipali, che in futuro saranno molto importanti per l'affidabilità e la responsabilità.

In conclusione, ci tengo a ribadire che, sebbene la ricerca scientifica in tutto il mondo abbia fatto passi da gigante nell'ambito dei sistemi generativi con immagini, video e testi multimodali, esistono ancora problemi riguardo al consumo energetico, alla sostenibilità etica e al fatto che i sistemi spesso sono stati addestrati anche con dati non corretti o tossici, riproducendo questo tipo di contenuti e fornendo allucinazioni. Le prossime generazioni risolveranno sicuramente parte di questi problemi e la ricerca mondiale sta andando verso sistemi fondazionali che siano sostenibili e sempre più in grado di integrare multimodalità insieme, quindi testo, video, audio, serie temporali finanziarie, dati medici di tutti i tipi, dati genomici, per ottenere risultati completi e, comunque, poter addestrare o anche riaddestrare questi sistemi in modo etico, anche a livello multimodale. Questo tipo di attività di ricerca, per fortuna, viene portata avanti con una grande collaborazione internazionale, in Europa, in Cina e negli Stati Uniti, come si vede dai convegni e dalle riviste più importanti del settore.

Con questo vi ringrazio. Vi ringrazio molto per l'attenzione. Questi sono i nostri contatti e ci tengo di nuovo a ringraziare il CINECA, Nvidia, per il supporto delle GPU, il supporto del Ministero italiano della Ricerca per i progetti PNRR, e FIS e della Comunità Europea per i progetti europei, soprattutto quelli della nostra rete ELLIS, nella quale diverse unità italiane fanno parte della rete di eccellenza.

Riferimenti

- ELLIOT (Open Multimedia Foundation Models) https://www.elliot-ai.eu/
- ELIAS (European Lighthouse of AI for Sustainability) https://elias-ai.eu/
- ELSA (European Lighthouse on Secure and Safe AI) https://elsa-ai.eu/
- FAIR (Future AI Research) https://fondazione-fair.it/
- MINERVA https://minerva4ai.eu/
- IT4LIA AI Factory https://it4lia-aifactory.eu/
- OECD 2019 OECD, Recommendation of the Council on Artificial Intelligence, C/MIN(2019)3/FINAL,
- AI ACT2024 [CE 2024] EU Commission Artificial Intelligence Regulation in Europe AI Act (2024)
- PNR 2021-2027 https://www.mur.gov.it/sites/default/files/2021-01/Pnr2021-27.pdf
- Strategia Italiana In Intelligenza Artificiale https://www.agid.gov.it/sites/agid/files/2024-07/Strategia_italiana_per_l_Intelligenza_artificiale_2024-2026.pdf
- S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, R. Cucchiara. «Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models», ECCV 2024
- F. Cocchi, N. Moratelli, M. Cornia, L. Baraldi, R. Cucchiara, "Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visual Question Answering." CVPR 2025.
- L. Baraldi, D. Bucciarelli, F. Betti, M. Cornia, M., N. Sebe, R. Cucchiara, (2025). What Changed? Detecting and Evaluating Instruction-Guided Image Edits with Multimodal Large Language Models. In ICCV 2025

Italy-China Week of Science,
Technology and Innovation

Responsibility and sustainability

in generative multimodal artificial intelligence systems

Rita Cucchiara
University of Modena and Reggio Emilia, Italy

Ministers, authorities, colleagues, thank you very much for the opportunity to present activities and results from cutting edge scientific research in artificial intelligence and to offer a brief discussion on responsibility and sustainability in generative multimodal artificial intelligence systems.

Multimodal generative systems, known in the scientific community as **foundational model,** can take as input text, images, audio, and video, reproduce and generate content in all modalities, and also query external databases, across multiple and diverse modalities. Research in this field has emerged only in the last decade, thanks to broad international collaboration among computer vision, NLP (Natural Language Processing), and machine learning. Since 2021, the first systems known as contrastive have been proposed, trained with pairs of two different modalities such as text and images; systems like CLIP originated from a proposal in the United States and then evolved into systems proposed around the world, in China and in Europe, generally as open source, which are extremely valuable for research.

In this area of multimodal generative systems, Europe and Italy are working with commitment; I will quickly mention some active European projects that have funded my laboratory, such as ELLIOT, ELIAS, ELSA of the ELLIS network (European Laboratory of Learning and Intelligent Systems) and MINERVA, all focused on generative AI, and in particular the ITA4IA project, our Italian iFactory coordinated by the CINECA supercomputing center. I also mention many Italian projects recently funded by Next Generation Italy, that is, by the PNRR, such as the extended partnership FAIR, coordinated by the CNR, ITSER and FIT4MedRob, also coordinated by the CNR, and FIS projects such as TRAMIS on robotics, in which my University is involved.

Although proposed at international conferences only in the last four years, multimodal generative systems already have **many fields of application**: for example, generating modified images useful for medical robotics to support transplant operations, or for human robot interaction (as in the example of robots able to operate in a home and describe what they are seeing, even while moving in an unknown environment); or, more naturally, generating synthetic or modified images for tourism applications; or creating images in the automotive domain that can recognize the car, the road, and the scene, and also estimate the amount of traffic perceived by a vehicle; and finally, in e-commerce, enabling responsible and sustainable image editing.

These models and applications fall **within the scope of responsibility and sustainability** as developed in Europe in recent years and adopted by our country. Since 2019, the OECD has defined foundations for responsibility in artificial intelligence, which in the European Union are closely associated with trustworthy and reliability, based on three pillars: systems must be lawful, ethical, and robust. These principles were then translated into the European Union's law regulating artificial intelligence, the AI Act, with seven well known requirements: systems must be supervised by a human; they must be technically robust and safe; they must correctly manage privacy and data; they must be transparent and disclose when content is generated by a machine rather than human work; they must be nondiscriminatory and fair and correctly manage diversity; they must be socially and environmentally sustainable; finally, it is important to establish accountability, that is, who bears responsibility when the system makes mistakes.

These concepts of reliability, responsibility, and sustainability have also been adopted in Italy for several years, starting with the PNR, the National Research Program 2021–2027, when a work program on artificial intelligence was created, which I had the honor of coordinating, specifying that systems proposed by Italian research should be ethical from the design stage, with human control at every stage, reliable and worthy of trust. The same idea was reiterated in the latest Italian Strategy for Artificial Intelligence 2024–2026, which uses exactly the same principles. Finally, a few weeks ago, in September 2025, Italy adopted its first law on AI and on the responsibility of artificial intelligence, reaffirming the same principles of reliability and sustainability and adding concrete controls and safeguards, especially for areas such as employment and health, as well as a complete framework for system governance, for investments, and for the responsibility of those who design and use such systems.

If this is the context in which Europe and Italy intend to position themselves, it does not mean that there have not been investments and significant achievements in scientific research aimed not so much at current systems as at those of the next generations. In particular, I would like to highlight three directions of research: *ethically responsible*, *computationally sustainable*, and *reliable in reducing hallucinations*. These results have recently been presented at the most important international venues such as ICCV, ECCV, and CVPR, where Europe, China, and the United States meet annually to discuss frontier research.

The first topic concerns systems that are responsible both technically and ethically, in particular the possibility of using existing models and retraining them through an approach called unlearning, that is, removing internal concepts that may have been acquired, such as violence, toxicity, nudity, or unsafe concepts, both in generation and in the retrieval of similar images. For example, in Safe-CLIP, if a system receives a violent or unsafe request such as “draw a group of people in a battlefield with buildings in the background,” it may take “battlefield” metaphorically and show a stadium with a match; or, if asked to “generate an image of a child holding a gun,” the system may avoid the concept of a gun and depict a child with a toy. These models, partially developed in Italy together with the University of Amsterdam in a European project and trained thanks to the CINECA supercomputing system, will soon be available as open source, for example for educational applications or wherever children are involved, where using concepts of violence or nudity is neither necessary nor appropriate in the context of AI.

A second result, developed in Italy by my university in collaboration with the European community, concerns new sustainable systems, in this case called self reflective, which are able “to know that they do not know,” as Socrates said, that is, to decide, from the standpoint of energy sustainability, when it is advisable to query external databases and how useful the retrieved results may be. This system uses open sources and was also developed in collaboration with China, using multimodal systems such as Qwen VL, recently developed by Alibaba research centers and available as open source.

Using these new models and this new approach to training that is sustainable from an energy perspective, it is possible to obtain results with high precision and very high reliability, especially for specific queries like the examples: for instance, when asked “where is this garden located?”, systems of this kind can query external sources such as Wikipedia only when needed and return the correct result without wasting energy on fine tuning and retraining.

As a final example of an interesting research result that may form the basis of upcoming open systems available to Italian, European, and global companies, I cite new solutions developed in Europe, by the University of Trento and UNIMORE, to assess the degree of correctness, or conversely the degree of hallucination, of systems when they generate images. This is a very important problem for ethically responsible artificial intelligence, because knowing and measuring the degree of error allows us to save time and energy and to be more confident in system reliability. Thanks to the massive use of CPUs and GPUs in Italian supercomputing, we have been able to train these models and obtain results like those shown: the system not only, as many now do, can modify images by changing the color of a part or a flag, or in the medical

field by removing an uninteresting part of a histological image, but can also assess, through agent based reasoning, whether the modification is consistent with the given prompt. In the flag example, the system recognizes that although a symbol typical of Canada has been added, it is not the Canadian flag, while in the other case only the requested part has been correctly modified in the histological slide. These are new tools, currently prototypes, which in the future will be very important for reliability and responsibility.

In conclusion, although scientific research worldwide has made great progress in generative systems for images, video, and multimodal text, problems remain regarding energy consumption, ethical sustainability, and the fact that systems have often been trained with incorrect or toxic data, reproducing such content and producing hallucinations. The next generations will certainly solve part of these problems, and global research is moving toward foundational systems that are sustainable and increasingly able to integrate modalities together, including text, video, audio, financial time series, medical data of all kinds, and genomic data, to achieve comprehensive results and to allow these systems to be trained or retrained ethically, including at the multimodal level. Fortunately, this research activity is carried out through major international collaboration in Europe, China, and the United States, as evidenced by the most important conferences and journals in the field.

With this, I thank you. Thank you very much for your attention. These are our contacts, and I would like again to thank CINECA and Nvidia for GPU support, the Italian Ministry of Research for PNRR projects, and FIS and the European Union for European projects, especially those of our ELLIS network, in which several Italian units are part of the network of excellence.

References

- ELLIOT (Open Multimedia Foundation Models) https://www.elliot-ai.eu/
- ELIAS (European Lighthouse of AI for Sustainability) https://elias-ai.eu/
- ELSA (European Lighthouse on Secure and Safe AI) https://elsa-ai.eu/
- FAIR (Future AI Research) https://fondazione-fair.it/
- MINERVA https://minerva4ai.eu/
- IT4LIA AI Factory https://it4lia-aifactory.eu/
- OECD 2019 OECD, Recommendation of the Council on Artificial Intelligence, C/MIN(2019)3/FINAL,
- AI ACT2024 [CE 2024] EU Commission Artificial Intelligence Regulation in Europe AI Act (2024)
- PNR 2021-2027 https://www.mur.gov.it/sites/default/files/2021-01/Pnr2021-27.pdf
- Strategia Italiana In Intelligenza Artificiale https://www.agid.gov.it/sites/agid/files/2024-07/Strategia_italiana_per_l_Intelligenza_artificiale_2024-2026.pdf
- S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, R. Cucchiara. «Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models», ECCV 2024
- F. Cocchi, N. Moratelli, M. Cornia, L. Baraldi, R. Cucchiara, "Augmenting Multimodal LLMs with Self-Reflective Tokens for Knowledge-based Visual Question Answering." CVPR 2025.
- L. Baraldi, D. Bucciarelli, F. Betti, M. Cornia, M., N. Sebe, R. Cucchiara, (2025). What Changed? Detecting and Evaluating Instruction-Guided Image Edits with Multimodal Large Language Models. In ICCV 2025