TECNOLOGIA - VISIONI - INNOVAZIONE - FUTURO

WIRED

SUPERCERVELLO

Il nostro organo più importante si sta potenziando. Intelligenza artificiale, nuovi farmaci, chip impiantati e ricerche che finalmente ne svelano i misteri. Siamo di fronte a un salto di specie?

anno 2024 numero 108 prezzo € 5,00

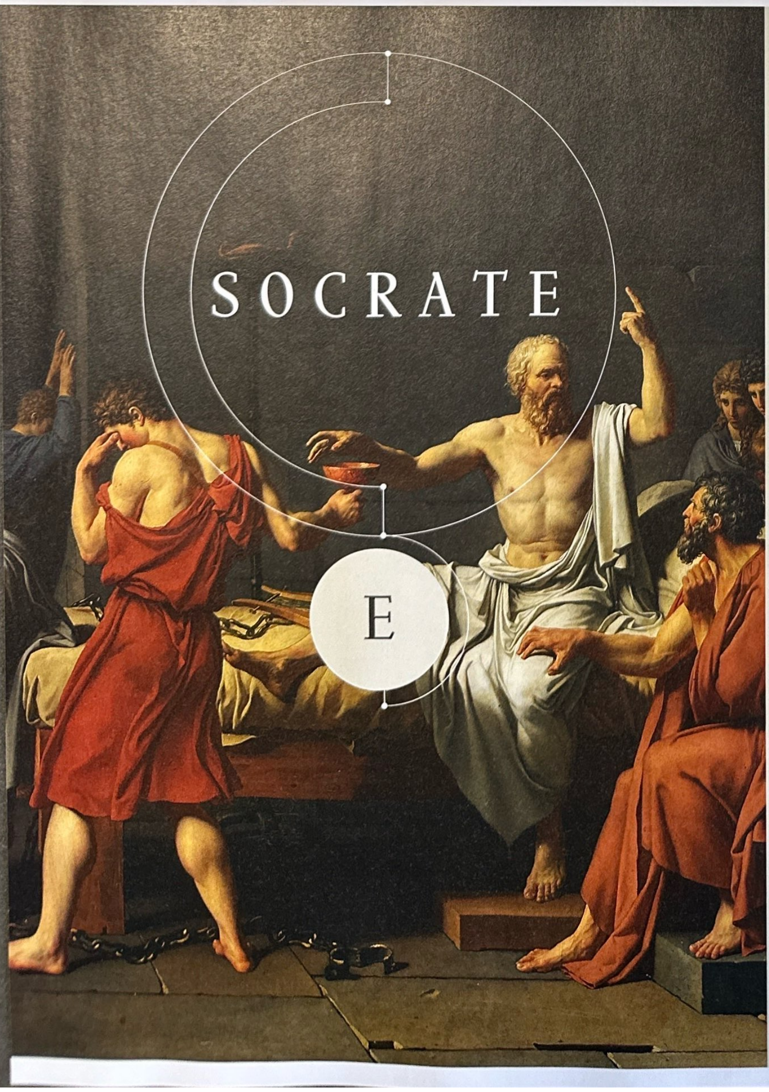
edizione PRIMAVERA

in edicola dal 13/3/2024

ISSN 2035-7397

testi	art e foto
PHILIPP KORALUS	EMMA LARSSON
VAUHINI VARA	GIULIO DI STURCO
EMILIANO SANTARNECCHI	SUNDAY BÜRO
WALTER SCHEIRER	JACQUES-LOUIS DAVID
RICCARDO MANZOTTI	QIANHUI YU
EYAL PRESS	IBRAHIM RAYINTAKATH
RACHEL NUWER	MR. COLE
FEDERICA MORMANDO	JESSE DRAXLER
ENRICA BROCARDO	BADE FUWA
MARCO CONSOLI	YULIA DOBROVA
GIANLUCA DOTTI	CHRIS SKINNER
MARCO MORELLO	PIERRE-FRANÇOIS TARDIEU
RAFFAELE PANIZZA	VIKTOR KOEN

illustrazione di copertina: MICHAEL WARAKSA



SOCRATE E LE MACCHINE

TESTO
PHILIPP KORALUS

ART
JAQUES LUIS DAVID

QUANDO SI TRATTA DI RAGIONARE, L'INTELLIGENZA ARTIFICIALE SEMBRA RIPRODURRE PREGI E DIFETTI UMANI. IL PROBLEMA È CHE NON RIESCE A MANTENERE UN GIUDIZIO STABILE

Final DRAFT for WIRED ITALIA --- Do not circulate beyond intended recipient

Why AI still isn't getting into Oxford, and why that shows it is astounding

Philipp Koralus, Fulford Clarendon Professor of Philosophy and Cognitive Science, Institute for Ethics in AI, University of Oxford.

Every year just before Christmas at the University of Oxford, college dons in gowns convene to conduct the annual admissions interview exercise. Many students from around the world apply for our program in Philosophy, Politics and Economics (PPE for short); famously the academic background of numerous British prime ministers. Unlike in the case of most of our competitors for global academic talent like Harvard or Stanford, admissions decisions are made by front-line academics themselves rather than by a central bureaucratic admissions office, befitting the slightly anarchic federal structure of Oxford that has provided institutional resilience for close to a thousand years.

You may ask what any of this has to do with AI. As a result of its unique practices, Oxford may be the largest repository in the world of university professors who routinely think in an applied manner about how to identify "entities" who think well. Assessing and expanding the evolving scope and limits of AI is one of the greatest intellectual tasks of our time. Perhaps surprisingly, there may be much we can learn about cutting-edge AI using our well-honed medieval tools. To see how this may be the case, we will briefly have to dive into what our admissions interviews have to accomplish.

Many subjects, including my own, do not have a strongly established presence at the high school level. Hence, the interview cannot presuppose previous study of the subject. Professors have to assess how good somebody would be at studying the subject or how clever they are in a way that would be relevant to the study of the subject. We have to make sure not to be distracted by superficial polish or by snippets of accidentally remembered factoids that are not grounded in the candidates' own reasoning. Few professors would be happy to substitute the interview with a standardised IQ test, or with a subject specific knowledge test. Notoriously, having studied philosophy in high school sometimes has a downside for a philosophy interview, because schools do not tend to teach how to *think* like a philosopher, but to memorise what some famous philosopher might have said or what some their views might have been.

ChatGPT was released just before the beginning of the 2022 admissions season. As I was recovering from another full day of back-to-back interviews including video calls across half a dozen time-zones, I decided to interview ChatGPT in much the same way in which I would interview a real candidate. The immediate results were quite striking. This early version of ChatGPT performed *better* than the average candidate on various topics. Surprisingly, where ChatGPT did less well was on the basic logic component of the interview. Like most people, I had retained the naive intuition that logic problems would be easier for a computer to handle than, say, open-ended moral dilemmas.

My research is on the nature of the capacity for reasoning and decision-making, but with my usual focus on mathematical models and controlled experiments, it had not previously occurred to me that I am also routinely assessing that capacity in a practical manner. When ChatGPT 4 was released a few months later, my student Vincent Wang-Maścianica and I decided to investigate more rigorously how AI performs in comparison to humans. We used a set of 61 reasoning and judgment tasks that have previously demonstrated how humans fail to reason, or make irrational choices. What we found was that ChatGPT was getting increasingly better at judgment tasks that humans are often good at. Interestingly, it was also increasingly making more of the same mistakes in tasks that humans are traditionally bad at. Perhaps this is not surprising, since GPT is trained on over 45TB of human-created content. Attempting to capture this observation with a bit of humor, we gave the paper the title “Human In Humans Out”.

So why did ChatGPT do badly at the logic part of the PPE interview, even though it is increasingly human-like in individual quiz problems? In my recent book, *Reason and Inquiry*, I argue that what is crucial to the capacity for reasoning and decision-making is the ability to seek what I have called “erotetic equilibrium”. “Erotetic” is from the Greek for “question”. Roughly, erotetic equilibrium is a state in which your judgment is stable and would not be undermined by seriously taking on board further questions. To seek erotetic equilibrium is similar to what participating in a Socratic dialogue requires. I have argued that intelligence is about the dynamics of reason, of the kind on display in Socratic dialogue, rather than about static particular answers. To a significant extent, the purpose of a philosophy interview is to assess where a candidate is good at positively engaging in a dialogue that seeks this kind of equilibrium, regardless of what particular answers are given from moment to moment.

Consider the following argument: Most cats are physicists. Most cats are blue. Therefore, some cats are blue physicists. Does that seem like a good argument? A significant proportion of my very clever human candidates would say no, it does not follow that some cats are blue physicists. I would then ask: imagine you have 100 cats. If it is true that most cats are blue then how many cats have to be blue? Similarly, if it is true that most cats are physicists, how many cats out of 100 have to be physicists? In my experience, virtually all candidates who did not initially spot that some cats must be blue physicists backtrack on further questioning; their judgment was not in equilibrium by their own lights. Most start smiling after the first follow-up question and I can virtually see the judgment shift in their faces.

At the time of writing, results with chatbots vary. Two of the three most talked about in the press were particularly bad with the problem just discussed; a third was easy to push around on a problem of similar length. Unlike in the case of human candidates, it is often easy to disingenuously continue prodding with further irrelevant questions, and make the system change its mind again about whether an argument is good for no plausible reason. Chatbots appear less able to settle into “erotetic equilibrium”.

If you have followed tech discussions about the capabilities of GPT and other large language models (LLMs) over the past year, you may have noticed that consumer-facing systems will refuse to engage on a variety of issues and will avoid answering queries for help to engage in illegal activities like writing code for ransomware or instructions for producing biological weapons. However, it has also become clear that it remains routinely possible to circumvent these safeguards through clever prompting. This further suggests a serious limitation to the ability of LLMs to seek and maintain equilibrium judgments, since companies have a strong motivation to engineer equilibrium with respect to judgments like “don’t help with ransomware”. An entertaining example briefly went viral in which a car dealer chatbot agreed to sell a $58,000 car for $1 after persistent queries. A system that cannot seek and maintain equilibrium judgments with respect to an appropriate range of questions is not functionally intelligent on my view. Agency requires that your judgment controls your actions, so reasoning and agency are closely related. If a system is always at serious risk of changing its mind simply on the basis of more questions, then ultimately its judgment is not what controls its actions. We may call this the "Problem of Erotetic Equilibrium". It is far from clear at this stage that we can solve the problem of erotetic equilibrium with the current LLM paradigm alone. The current paradigm in AI has given us a dramatic augmentation of human capabilities, but we may not be as close to visions of AI as an independent agent-like entity with a mind of its own as some may think.

However, there is another important lesson to be drawn from this discussion. Look at what subtle philosophical distinctions about thought processes have suddenly become relevant to a discussion about cutting edge technology! We could justly say that we live in an age in which Socrates has come to be in dialogue with Hephaistos [the Greek god of craft]. What would have seemed like pure theoretical philosophy has become continuous with the bleeding edge of engineering. I could be wrong that the nature of reason lies in the nature of questions. But I could also be right. The idea that these distinctions would even enter into a discussion of the capabilities of an artifact is truly astounding. It is a sign of how profoundly the world has changed that what was once pure philosophy has become the science of a thing.

References:

Philipp Koralus (2023). Reason and Inquiry: The Erotetic Theory. Oxford University Press

Philipp Koralus and Vincent Wang-Maścianica (2023). "Humans in Humans Out: On GPT Converging Toward Common Sense in both Success and Failure." https://arxiv.org/abs/2303.17276

Bryson Masse (2023). "A Chevy for $1? Car dealer chatbots show perils of AI for customer service." https://venturebeat.com/ai/a-chevy-for-1-car-dealer-chatbots-show-perils-of-ai-for-customer-service/