

“L’Agi [intelligenza artificiale generale, ndr] si sta avvicinando molto rapidamente – afferma Leopold Aschenbrenner, ricercatore di OpenAI coinvolto nel gruppo di ricerca Superalignment, istituito a luglio -. *Vedremo modelli sovrumani, che avranno ampie capacità e potrebbero essere molto, molto pericolosi, e non abbiamo ancora i metodi per controllarli*”. OpenAI ha dichiarato che dedicherà il 20 per cento della sua potenza di calcolo al progetto Superalignment.<sup>1</sup>.

*Queste sono le preoccupazioni, cercando di prevedere le future evoluzioni dei sistemi di I.A. in particolare i Large Language Modules (LLM).*

“I ricercatori di OpenAI hanno esaminato il processo, ribattezzato “supervisione”, che viene utilizzato per far sì che i sistemi come GPT-4, il grande modello linguistico alla base di ChatGPT, in siano più utili e meno dannosi. Attualmente questo processo prevede che gli esseri umani forniscano dei feedback all’AI, indicando quali risposte sono utili e quali invece dannose. Con il progredire della tecnologia, i ricercatori stanno cercando di capire come automatizzare questo processo: non solo per risparmiare tempo, ma anche perché sono convinti che con l’aumentare della potenza dell’AI fornire dei feedback utile potrebbe diventare impossibile per l’uomo.” I ricercatori hanno utilizzato ChatGPT-2 per controllare in tempo reale le prestazioni ChatGPT-4<sup>2</sup> (...)

I ricercatori descrivono la possibilità addestrare da un modello AI più forte utilizzando un sistema meno capace come *“un elemento chiave per il problema più ampio del superallineamento”*<sup>3</sup>.

La procedura messa in atto in OpenAI è suggestiva<sup>4</sup>, per seguire il procedere di una I.A. si utilizza una seconda I.A. di rango inferiore; possiamo immaginare in futuro una struttura di controllo sempre più stratificata e articolata, una rete di dialoghi tra I.A., le *‘intelligenze naturali’* possono stare in superficie (roughly human-level), filtrando le informazioni che vengono dalla rete delle I.A.; c’è da interrogarsi a fondo sul grado di complessità che queste operazioni di sorveglianza/interpretazione/validazione possono raggiungere avendo ad un capo reti neurali con un numero di nodi/conessioni che cresce esponenzialmente.

L’accento messo dai produttori dell’I.A. sulla necessità di vigilare sull’evoluzione della loro stessa tecnologia -ricordiamo l’incontro tra i principali produttori e la presidenza USA- in un primo momento aveva indubbiamente anche una funzione promozionale, tendente ad impedire la messa in atto di controlli troppo stringenti da parte del governo, ora -alla luce di quanto si sta facendo e investendo- le preoccupazioni appaiono molto più reali. L’uso dell’I.A è sempre più pervasivo, trasversale ad ogni filiera produttiva, logistica o amministrativa; tenerne sotto controllo l’evoluzione richiede di intervenire su una rete di applicazioni sempre più complessa ed estesa, con un ventaglio di conseguenze che si dispiega a dismisura in tutte le dimensioni della formazione sociale.

Se è quello descritto è lo stato dell'arte nei laboratori dove si realizzano e si evolvono -possiamo dire in modo esponenziale- le tecnologie dell'I.A. e le prestazioni dei sistemi, dobbiamo chiederci quale possa essere l'efficacia delle norme e delle procedure che si stanno definendo a livello nazionale e internazionale<sup>5</sup>, la conferenza di Londra<sup>6</sup>, negli Stati Uniti con l'Executive Order del presidente Biden<sup>7</sup>, nell'Unione Europea l'Artificial Intelligence Act emanato dalla Commissione su cui è stato raggiunto un accordo provvisorio tra il Parlamento ed il Consiglio<sup>8</sup>. Se l'EO di Biden si basa sul coinvolgimento e la responsabilizzazione di tutte le istituzioni ai diversi livelli -allo scopo di intervenire su tutti i gangli della società, dei processi di produzione e riproduzione sociale pervasivamente investiti dalle tecnologie di I.A.- la direttiva europea trova il suo fondamento nella definizione di livelli di pericolosità delle applicazioni di I.A. da cui dipende tutta l'architettura normativa e discendono le procedure di controllo e autorizzazione.

*Il dispositivo di controllo che OpenAI sta costruendo corrisponde ai primi due punti elencati nel Fact Sheet relativo all'Executive Order di Biden, tuttavia contemporaneamente ne evidenzia l'ardua e complessa realizzazione.*

Require that developers of the most powerful AI systems share their safety test results and other critical information with the U.S. government. Develop standards, tools, and tests to help ensure that AI systems are safe, secure, and trustworthy.

Come abbiamo evidenziato nel nostro articolo già citato<sup>9</sup> e quindi da rileggere, si evidenzia:

“Nell' E.O. sull'A.I. è illuminante l'elenco dei capitoli: New Standards for AI Safety and Security, Protecting Americans' Privacy, Advancing Equity and Civil Rights, Standing Up for Consumers, Patients, and Students, Supporting Workers, Promoting Innovation and Competition, Advancing American Leadership Abroad, Ensuring Responsible and Effective Government Use of AI.”

Esso rispecchia il tentativo di seguire lo sviluppo verticale nella sua crescente complessità tecnologica e orizzontalmente in tutte le articolazioni della società.

L'Artificial Intelligence ACT elaborato dalle istituzioni europee, presentato dalla Commissione il 21 aprile 2022 dopo un lungo processo di elaborazione<sup>10</sup>, ha richiesto un anno e mezzo per arrivare all'accordo 'provvisorio' tra Parlamento e Consiglio<sup>11</sup> con modifiche rispetto al testo elaborato dalla commissione: definibili sommariamente come:

- norme sui modelli di IA high-impact general-purpose che possono causare un rischio sistemico in futuro, nonché sui sistemi di IA ad alto rischio

- un sistema di governance riveduto con alcuni poteri esecutivi a livello dell'UE
- l'estensione dell'elenco dei divieti, ma con la possibilità di utilizzare l'identificazione biometrica a distanza da parte delle autorità di contrasto negli spazi pubblici, fatte salve le garanzie
- una migliore tutela dei diritti attraverso l'obbligo, per gli operatori di sistemi di IA ad alto rischio, di effettuare una valutazione d'impatto sui diritti fondamentali prima di mettere in uso un sistema di IA.

Il processo di approvazione di questo 'accordo provvisorio è il prodotto di una lunga e apra contrattazione<sup>12</sup>.

Al centro del dibattito c'era la questione se le autorità statali dovessero essere autorizzate a implementare sistemi biometrici basati sull'intelligenza artificiale in grado di identificare e classificare le persone in base a caratteristiche sensibili come genere, razza, etnia, religione e affiliazione politica, nonché sistemi di riconoscimento delle emozioni e politiche predittive.

Alla fine, il Parlamento ha ceduto alle richieste degli stati e ha firmato<sup>13</sup> una serie di "condizioni rigorose" che consentiranno alle autorità di utilizzare la biometria in tempo reale per cercare le vittime di rapimento, tratta e sfruttamento sessuale, prevenire le minacce terroristiche e localizzare le persone sospettate di aver commesso reati gravi, come terrorismo, omicidio, stupro e rapina a mano armata. Al contrario, sarà vietata la categorizzazione biometrica basata su caratteristiche sensibili, il punteggio sociale, la polizia predittiva, lo sfruttamento delle vulnerabilità e il riconoscimento delle emozioni sul posto di lavoro e nelle istituzioni educative.

A seguito dell'accordo provvisorio odierno, nelle settimane seguenti proseguono i lavori a livello tecnico per mettere a punto i dettagli del nuovo regolamento. Una volta conclusi i lavori, la presidenza sottoporrà il testo di compromesso ai rappresentanti degli Stati membri (Coreper) per approvazione. L'A.I. Act entrerà in vigore due anni dopo la sua approvazione, da oggi ad allora il panorama tecnologico sarà presumibilmente radicalmente diverso.

La metodologia che si sta cercando di sviluppare in OpenAI avvolgere le A.I. più potenti in un reticolo di tecnologie, progressivamente meno potenti, denota il grado di complessità ed imprevedibilità con cui si confronta il tentativo di predire e governare gli sviluppi futuri.

Tutte le previsioni, e analisi che abbiamo citato negli articoli precedenti, disegnano il quadro di rapporti sociali soggetti ad una trasformazione radicale, con centinaia di milioni di posti di lavoro messi in gioco, in buona sostanza gli equilibri economici e sociali, le identità stesse di soggetti e gruppi sociali viene messa in discussione. La soglia di rischio non è

certa definita dalla capacità di scoprire una miriade di nuovi materiali possibili, di nuove configurazioni proteiche che in passato non sarebbe stato possibile investigare. In fondo non si tratta dell'uso da parte degli stati costruire arbitrariamente sempre nuovi e potenti strumenti di controllo, su cui si è discusso durante la stesura dell'A.I. Act; le eccezioni per le quali gli stati sono autorizzati alla profilazione sempre più sofisticata per combattere i crimini più odiosi e soprattutto l'onnipresente 'pericolo terrorista'. Le istituzioni e gli apparati della sicurezza e della difesa hanno di fatto campo libero nell'uso delle tecnologie più avanzate nella profilazione, dell'identificazione, dell'analisi comportamentale, ciò accade nel contesto di sempre nuove 'emergenze' dichiarate una dopo l'altra.

L'utilizzo ed il controllo sull'utilizzo di queste tecnologie sono affidati a democrazie che nel nuovo secolo stanno vivendo sempre nuove derive. Ciò che le nuove tecnologie digitali stanno determinando in realtà è uno stato metastabile della formazione sociale, nel contesto -è quasi inutile ripeterlo- del massimo fattore di instabilità vale dire la crisi climatica. Il riscaldamento globale è il prodotto di un accumulo di CO<sub>2</sub> dall'inizio dell'era industriale, contemporaneo ad una successione di rivoluzioni tecnologiche che sono giunte alla soglia della transizione energetica, ecologica, digitale e biotecnologica. Abbiamo uno stato metastabile della formazione sociale priva di una effettiva stabilità strutturale. Queste dinamiche non possono che esaltare le dinamiche conflittuali e competitive, di cui sono protagonisti gli stati ed i poteri oligopolistici.

Tuttavia il carattere 'disruptive' dell'innovazione tecnologica fin qui descritto non è ciò che preoccupa in OpenAI o meglio è un altro livello di questo potenziale sconvolgente, è la sostanziale imprevedibilità dei comportamenti dell'I.A., la totale mancanza di trasparenza, anzi la capacità stessa delle applicazioni di competere con i dispositivi di controllo, in una sorta di inseguimento, in una crescita di complessità.

L'integrazione di più dispositivi di I.A. tra loro, una sorta di architettura dialogante, costituisce un'ormai una strategia che si va espandendo e consolidando. Ne è un esempio l'uso di due dispositivi per la soluzione di problema di matematica di ardua risoluzione. La descrizione più semplice è nel comunicato dell'ANSA<sup>14</sup>.

*Il nuovo sistema di intelligenza artificiale, chiamato FunSearch (che sta per 'functions search', cioè ricerca di funzioni matematiche), è formato da due componenti che lavorano in tandem: un modello linguistico di grandi dimensioni, pre-addestrato a fornire soluzioni creative sotto forma di codice informatico, e un 'valutatore' automatico, che ne esamina le proposte scartando quelle più improbabili e scorrette. Le migliori vengono invece inviate di nuovo al modello linguistico che, attraverso questo procedimento iterativo che si ripete milioni di volte, affina sempre di più le sue soluzioni sviluppando nuove informazioni che vanno al di là della conoscenza umana, dunque delle vere e proprie scoperte.*

L'articolo del MIT<sup>15</sup> entra più nel merito, è disponibile in open access il paper originale<sup>16</sup>.

Già in un precedente articolo *Intelligenza Artificiale, la grande trasformazione: governo e mutazione antropologica*<sup>17</sup> a proposito della *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023, a conclusione della conferenza di Londra promossa dal governo inglese* avevamo dato la sintesi seguente.

*"Il documento rilancia la cooperazione internazionale nel campo dell'I.A. e -come ogni analisi corrente- ne riconosce le straordinarie potenzialità trasformative ed indica nel contempo il coesistere di altrettanto straordinari rischi ed opportunità. Si indica una direttrice per lo sviluppo e l'utilizzo dell'I.A. umano-centrica, affidabile e responsabile. Se l'elenco delle opportunità definisce il quadro del migliore dei mondi possibili, proprio per il carattere pervasivo dell'utilizzo dell'I.A. nelle sue diverse forme, l'elenco dei rischi cita in buona sostanza la possibilità di sovvertire i dispositivi, le procedure con cui si può certificare l'attendibilità di una qualsiasi informazione, la possibilità di penetrare entro qualsiasi relazione e comportamento attraverso le sue tracce digitali e l'utilizzo delle stesse a fini genericamente criminali, in questo elenco la guerra non è citata."* Il nostro giudizio paragonava questa presa di posizione al greenwashing rispetto alla transizione energetica ed ecologica.

Dicevamo: la preoccupazione maggiore in realtà è orientata verso la frontiera degli sviluppi possibili dell'I.A., in questo condividendo l'orientamento prevalente nel confronto pubblico sul tema, l'attenzione è focalizzata su "highly capable general-purpose AI models, including foundation models, that could perform a wide variety of tasks – as well as relevant specific narrow AI that could exhibit capabilities that cause harm – which match or exceed the capabilities present in today's most advanced models"

Centrale è il riferimento ai 'Foundation Models' a cui appartengono tutte le applicazioni di cui si parla negli ultimi mesi l'origine del termine è la seguente<sup>18</sup>

"Il Center for Research on Foundation Models (CRFM) dello Stanford Institute for Human-Centered Artificial Intelligence (HAI) ha coniato il termine "foundation model" nell'agosto 2021, riferendosi provvisoriamente a "qualsiasi modello addestrato su insiemi di dati di straordinaria dimensione (generalmente utilizzando l'auto-supervisione su larga scala) che può essere adattato (ad esempio, messo a punto) a un'ampia gamma di attività a valle". [14] Ciò si basava sulla loro osservazione che i termini esistenti [usati per definirli] non erano adeguati, sostenendo che il termine 'Large Language Model' era troppo ristretto dato che l'attenzione non è solo sul linguaggio; ' modello auto-supervisionato' era troppo specifico per l'obiettivo di formazione; E il "modello preaddestrato" ha suggerito che l'azione degna di nota è avvenuta dopo il "preaddestramento". [15] Dopo aver considerato molti termini, hanno optato per il "modello di fondazione" per enfatizzare la funzione

prevista (cioè, l'adattabilità a un successivo ulteriore sviluppo) piuttosto che la modalità, l'architettura o l'implementazione.)”

Dopo questa citazione del nostro precedente articolo, si comprende come sistemi sempre più potenti e complessi, addestrati su moli di dati sempre più estese messe a disposizione dalla rete producono risultati e prestazioni del tutto imprevedibili, secondo percorsi non trasparenti; un limite che i ricercatori di OpenAI cercano di infrangere. A seconda del campo di ricerca a cui si applica la I.A. le problematiche possono essere molto diverse, nel caso citato -il problema di matematica combinatoria- è stato prodotto un codice di programmazione non una semplice soluzione al problema, cosa che lo rende riutilizzabile e comprensibile, benché i due milioni di 'dialoghi' tra i due dispositivi -che hanno portato ad un tale risultato- non appartengano certo ad un livello genericamente umano di elaborazione.

I campi di applicazione dell'I.A: e le conseguenze possibili a tutti i livelli della società si estendono a ventaglio in tutte le direzioni, gli apprendisti stregoni, abbiamo visto cercano di tenere sotto controllo la propria creatura, ma le tentazioni del potere la storia ce lo insegna solitamente infrangono ogni soglia di potenza. Gli appelli, le procedure di controllo e verifica messe via via in atto nei laboratori - con una efficacia tutta da verificare- non corrispondono alla logica di gestione da parte degli stati.

Rispetto alle varie ondate di innovazione che hanno caratterizzato la storia e la trasformazione del rapporto sociale di produzione capitalistico -intendendo per fasi quelle caratterizzate dalla classe di innovazione che lo hanno trasformato radicalmente- quella che stiamo vivendo mostra un orizzonte di imprevedibilità del tutto vicino al momento presente, lasciando impredicabile anche il futuro prossimo.

In passato la rivoluzione tayloristica ha integrato come metodologia di organizzazione e di comando sui processi di produzione e cooperazione sociale, in termini di progettazione del prodotto e del processo di produzione, di processi a monte e a valle -pensiamo alla evoluzione del ciclo dell'automobile ed alle trasformazioni indotte in tutta la formazione sociale. Essa per quanto nessuno potesse prevedere il futuro, era nei suoi vari passaggi perfettamente comprensibile, anzi si basava proprio sulla calcolabilità delle procedure messe in atto e nelle sue conseguenze sociali ha contribuito alla nascita di nuovi modelli macroeconomici e forme di governo, sua pure passando attraverso ad un gigantesco conflitto mondiale. Lo stesso Lenin a suo tempo attraversò un lungo processo di elaborazione delle sue posizioni relative al taylorismo da 1913 alla presa del potere; una pubblicazione di riferimento in merito è *The Taylorization of Vladimir Ilich Lenin* di James G. Scoville a cui risponde *Lenin as Scientific Manager Under Monopoly Capitalism, State Capitalism, and Socialism: A Response to Scoville* di Victor G. Devinez pubblicato dalla fondazione Friedrich Ebert Stiftung; il dibattito è ovviamente più ampio, può essere significativo cogliere

l'incrocio tra l'elaborazione di una teoria rivoluzionaria di presa del potere prima e di un processo di transizione dopo ed un processo di trasformazione del capitalismo e quindi un giudizio nel merito entro quella elaborazione.

Oggi non possiamo osservare la radicale trasformazione del sistema capitalistico da dentro un processo rivoluzionario -*Lenin nel 1913 aveva vissuto già il 1905*- ma possiamo anzi dobbiamo prendere atto ed analizzare a fondo-facendo poi di volta in volta le nostre considerazioni e scelte conseguenti- il grado instabilità ovvero la condizione di meta stabilità in cui si trova l'intera formazione sociale globale nella sua trasformazione. In questo contesto le strategie esposte dai diversi governi, dai grandi oligopoli, dalle assise internazionali, ai più prestigiosi Think Thank si focalizzano sul temine della transizione energetica e digitale, energia/informazione la coppia che nelle sue relazioni e trasformazione definisce la traiettoria del sistema capitalistico, purché nel termine informazione si comprenda la manipolazione sempre pi profonda e pervasiva de mondo della vita, di ogni forma di vita ed ecosistema.

La COP28 ci ha esposto le strategie -scarsamente efficaci sul piano dell'interdizione del riscaldamento globale- con cui nei diversi paesi e regioni del globo si intende affrontare il cambiamento climatico. La soluzione nucleare -quella della fissione, la fusione per ora è lontana decenni- è una opzione che per molti paesi è pratica corrente per questi e molti altri è un progetto per il futuro prossimo. La filiera del nucleare richiede ovviamente una trattazione a parte nella sua evoluzione tecnologica, nel ciclo dell'energia e nel suo ruolo nei rapporti geostrategici. Oggi si parla molto di una quarta generazione in realtà composta da approcci tecnologici diversi tra loro, quello dei reattori autofertilizzanti che producono plutonio - e quindi si legano alla produzione di ordigni- per ora è poco praticata, l'esperimento francese si è rivelato di fatto un fallimento.

Per ridurre i tempi di realizzazione molto si discute e si punta sui piccoli reattori modulari (small modular reactors) indicati con l'acronimo inglese SMR, vedi ad esempio la presa di posizione del parlamento europeo<sup>19</sup>. La commissaria europea all'Energia, Kadri Simson, ha annunciato il lancio dell'Alleanza industriale europea per i piccoli reattori modulari, il progetto richiesto dalla Francia e da altri 11 Paesi membri per rilanciare la produzione di energia atomica nel continente<sup>20</sup>.

I costi reali ed i tempi di realizzazione degli impianti in realtà non sono mai quelli preannunciati quando si vogliono lanciare questi nuovi programmi, comunque è un dato di fatto che nel contesto dell'intreccio di crisi e trasformazioni radicali che stiamo vivendo la produzione di energia a mezzo del nucleare è uno dei punti di riferimento per questa fase di transizione in attesa dell'arrivo della fusione.

La questione digitale, l'implementazione su vasta scala delle tecnologie dell'I.A. in

particolare, si salda a quella energetica grazie all'enorme bisogno di energia necessaria ad alimentare i nuovi apparati, i centri di calcolo condiviso il cosiddetto CLOUD, la nuvola. Il mercato del CLOUD è dominato dalle società del Big Tech che per prime hanno avuto bisogno di alimentare a una potenza di calcolo in questo secolo è andata crescendo esponenzialmente, ora dominano il mercato e ne ricavano una parte importante dei propri profitti.

Microsoft infatti ha dichiarato di puntare alla nuova generazione di reattori nucleari per fornire energia ai propri apparati di Intelligenza Artificiale<sup>21</sup>, MS in particolare sembra rivolgersi ad una soluzione basata sui SMR ed è quindi alla ricerca di un partner che gli possa offrire la tecnologia e la soluzione industriale. Se questo progetto di avvierà concretamente come pare assai probabile, sicuramente verrà imitato, se già non sono in corso iniziative analoghe, con tutta la forza economica, tecnologica e finanziaria dei giganti del digitale che dominano il mercato del CLOUD, ma non solo.

Questo, articolo, come gli altri che lo hanno preceduto è parte di un processo di analisi, un work in progress, che pone attenzione alle trasformazioni del sistema capitalistico concentrandosi sul ruolo dell'innovazione tecnologica in tutti i settori ed a tutti i livelli, in sintesi nell'ambito -in realtà onnicomprensivo- della transizione energetica e digitale, quindi nel contesto del cambiamento climatico. Le grandi trasformazioni comprendono necessariamente il mutare progressivo, ma non lineare della configurazione geopolitica, la rottura ed il deflagrare di precedenti equilibri geostrategici, rapporti di forze a livello globale e regionale, ma sempre da quel definito punto di vista.

La parola transizione possiede di per sé una tonalità rassicurante, trasmette la promessa di un approdo felice e soprattutto condiviso partendo da una situazione drammatica in un percorso pieno di inciampi. In realtà vediamo come quest'orizzonte felice e condiviso per ora non esista, mentre si incrementano conflitti e diseguaglianze, aumentano le manifestazioni di processo di degradazione climatica ed ecologica. L'insieme del processo di riproduzione della formazione sociale globale, presenta un crescente numero di punti di rottura a livello locali e regionale, le sue dinamiche complessive appaiono meta stabili, sul punto cioè di perdere la regolarità delle proprie traiettorie.

Questa condizione richiede uno sforzo straordinario di condivisione delle conoscenze entro una strenua lotta per l'esistenza che allo stato attuale delle cose produce più divisioni che cooperazione solidarietà nel campo di chi più ne patisce.

Entro questo mare in tempesta tocca navigare e aggiustare man mano la rotta, ma non da soli.

Roberto Rosso

1. <https://www.wired.it/article/intelligenza-artificiale-openai-team-superalignment-ai-superintelligenti/>  
<https://www.wired.com/story/openai-ilya-sutskever-ai-safety/> [↔]
2. In un esperimento in cui GPT-2 – il generatore di testo di OpenAI distribuito per la prima volta nel 2019 – è stato utilizzato per addestrare GPT-4, l'ultimo modello dell'azienda è diventato meno capace e più simile al sistema precedente. Per risolvere il problema, i ricercatori hanno quindi testato due differenti idee. Il primo metodo prevedeva di addestrare di modelli progressivamente più grandi per ridurre la perdita in termini di prestazioni in modo graduale. Nel secondo, il team ha aggiunto una modifica algoritmica al GPT-4 che ha permesso al modello di farsi "guidare" dalla versione più debole senza ridurre le sue prestazioni. Questa strategia si è rivelata più efficace, anche se i ricercatori la descrivono come un punto di partenza per ulteriori ricerche, ammettendo che i metodi come questi non garantiscono che il modello più forte si comporti in modo perfetto.[↔]
3. <https://spectrum.ieee.org/openai-alignment> – <https://spectrum.ieee.org/the-alignment-problem-openai> [↔]
4. <https://openai.com/blog/introducing-superalignment> <https://openai.com/blog/our-approach-to-alignment-research> <https://openai.com/research/critiques>  
<https://openai.com/research/language-models-can-explain-neurons-in-language-models>

“Our goal is to build a roughly human-level automated alignment researcher. We can then use vast amounts of compute to scale our efforts, and iteratively align superintelligence.

To align the first automated alignment researcher, we will need to 1) develop a scalable training method, 2) validate the resulting model, and 3) stress test our entire alignment pipeline:

1. To provide a training signal on tasks that are difficult for humans to evaluate, we can leverage AI systems to assist evaluation of other AI systems (scalable oversight). In addition, we want to understand and control how our models generalize our oversight to tasks we can't supervise (generalization).
2. To validate the alignment of our systems, we automate search for problematic behavior (robustness) and problematic internals (automated interpretability).
3. Finally, we can test our entire pipeline by deliberately training misaligned models, and confirming that our techniques detect the worst kinds of misalignments (adversarial testing).”

[↔]

5. ((<https://transform-italia.it/intelligenza-artificiale-la-grande-trasformazione-governo-e-mutazione-antropologica/> [↔])
6. <https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023?fbclid=IwAR0oaVzzRtfGZSRAsKplqvgKfzWxELtob4Q2SETI1zvbWn-cRgJ7qElxcFU> [↔]
7. <https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/20/executive-order-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government/>  
<https://www.whitehouse.gov/briefing-room/statements-releases/2023/02/16/fact-sheet-president-biden-signs-executive-order-to-strengthen-racial-equity-and-support-for-underserved-communities-across-the-federal-government/> [↔]
8. <https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament/>

- ent-strike-a-deal-on-the-first-worldwide-rules-for-ai/ [↔]
9. <https://transform-italia.it/intelligenza-artificiale-la-grande-trasformazione-governo-e-mutazione-antropologica/> [↔]
  10. <https://artificialintelligenceact.eu/documents/> [↔]
  11. <https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/> [↔]
  12. <https://www.euronews.com/my-europe/2023/12/08/eu-countries-and-meps-strike-deal-on-artificial-intelligence-act-after-drawn-out-intense-t> [↔]
  13. <https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai> [↔]
  14. [https://www.ansa.it/canale\\_tecnologia/notizie/future\\_tech/2023/12/18/lia-di-google-deepmind-scopre-nuove-soluzioni-matematiche\\_5f4961fa-d28c-47c4-a6e9-d64cf5a8ff18.html](https://www.ansa.it/canale_tecnologia/notizie/future_tech/2023/12/18/lia-di-google-deepmind-scopre-nuove-soluzioni-matematiche_5f4961fa-d28c-47c4-a6e9-d64cf5a8ff18.html) [↔]
  15. <https://www.technologyreview.com/2023/12/14/1085318/google-deepmind-large-language-model-solve-unsolvable-math-problem-cap-set/> [↔]
  16. <https://www.nature.com/articles/s41586-023-06924-6> [↔]
  17. [https://transform-italia.it/intelligenza-artificiale-la-grande-trasformazione-governo-e-mutazione-antropologica/#footnote\\_2\\_33181](https://transform-italia.it/intelligenza-artificiale-la-grande-trasformazione-governo-e-mutazione-antropologica/#footnote_2_33181) [↔]
  18. <https://hai.stanford.edu/news/reflections-foundation-models> <https://crfm.stanford.edu/> [↔]
  19. <https://geagency.it/nucleare-pe-chiede-strategia-ue-su-piccoli-reattori-modulari/> I piccoli reattori modulari (small modular reactors) sono reattori nucleari minori sia in termini di potenza sia di dimensioni fisiche, rispetto alle centrali tradizionali su scala gigawatt, con una potenza compresa tra 10 e 300 MegaWatt. Si basano su tecnologie esistenti e sono progettati per essere costruiti in fabbrica in forma modulare standard e il loro vantaggio principale è che possono essere assemblati in fabbrica e poi spediti e installati sul posto, quindi anche in aree remote con capacità di rete limitata o in aree in cui l'uso di grandi centrali nucleari tradizionali non è possibile. Questa tipologia di reattori utilizza reazioni di fissione nucleare per creare calore che può essere utilizzato direttamente o per generare elettricità e sono di recente tornati al centro del dibattito politico in Ue nel pieno della crisi energetica con la Russia e nel tentativo di diversificare le fonti di approvvigionamento[↔]
  20. <https://europa.today.it/economia/europa-alleanza-nucleare-mini-reattori.html> [↔]
  21. <https://www.theverge.com/2023/9/26/23889956/microsoft-next-generation-nuclear-energy-smr-job-hiring> [↔]