Editoriale

Una nuova etica della progettazione per governare l'innovazione tecnologica

Questo articolo è stato pubblicato sull'inserto Login del Corriere della Sera il giorno 24 aprile 2023

Nell'ultimo periodo si è molto parlato di Intelligenza Artificiale (IA), ancora di più nelle scorse settimane: paradigmatico il caso di chatGPT, il chatbot conversazionale sviluppato da OpenAI, per il quale il Garante per la protezione dei dati personali ha richiesto un blocco temporaneo in Italia, e in generale dei cosiddetti *Large Language Models* (LLMs) per i quali è stata lanciata la proposta di sospendere il loro sviluppo per sei mesi, firmata fra gli altri da Elon Musk e Yuval Harari. Peccato che in questo gran parlare poco si sia detto di come non solo l'IA, ma tutta l'informatica, siano processi socio-tecnici così potenti da richiedere strumenti complessi per essere concepiti, sviluppati, usati e governati. Il filosofo Langdon Winner sostiene, infatti, che le tecnologie hanno sempre una connotazione politica, sia intenzionale, come per i "cavalcavia razzisti" progettati per impedire il passaggio del trasporto pubblico e di conseguenza delle minoranze, sia non intenzionale, come l'introduzione delle macchine per la raccolta dei pomodori e del conseguente impatto sul lavoro. Anche nel caso della IA l'attenzione si è concentrata su chi urlava più forte per sostenere che l'innovazione non deve essere fermata o sul fatto che l'IA ha prestazioni paragonabili a quelle umane. È passato del tutto inosservato il potere del progetto e di come le scelte di design determinino la connotazione morale e politica delle tecnologie. Le tecnologie non sono mai neutre. Esortiamo, infatti, i nostri figli a non stare attaccati ai social media tutto il giorno quando questi sono appositamente progettati per catturare irrimediabilmente la loro attenzione. Cosa serve quindi per aprire un dibattito costruttivo sulle sfide che ci attendono? È necessaria sicuramente una nuova cultura della progettazione tecnologica che

parta dal riconoscimento che le tecnologie non nascono nel vuoto, ma sono plasmate dalla società che a loro volta contribuiscono a plasmare. Serve anche riconoscere il grande potere che sta nelle mani e nelle menti di chi progetta, anticipare i problemi già in fase di progettazione, ma anche capire che progettare significa usare l'immaginazione. E ancora essere pronti alle conseguenze inattese che emergono quando la tecnologia è inserita nel suo contesto d'uso. Bisogna essere consapevoli che chi decide il progetto di una tecnologia lo fa, il più delle volte, senza renderlo oggetto di discussione pubblica. Serve non solo capire come progettare tecnologie più giuste, eque e sostenibili, ma anche chiedersi se e perché debbano essere effettivamente realizzate. Si tratta di un compito che coinvolge formazione, ricerca e dibattito pubblico. La formazione è essenziale per educare le nuove generazioni a essere consapevoli della portata morale, sociale e politica di quanto creano. La ricerca è vitale per offrire gli spazi di sperimentazione e di integrazione dei saperi. Il dibattito pubblico è cruciale per promuovere la discussione aperta e critica che accolga visioni diverse. Ci sono già segnali incoraggianti in questa direzione. Il Politecnico di Milano ha da tempo un gruppo di umanisti e scienziati sociali che riflette sulle questioni filosofiche, etiche e sociali della scienza e della tecnologia e offre diversi corsi su questi temi ai suoi studenti. Il CINI, il consorzio interuniversitario per l'informatica, ha un laboratorio di Informatica e Società che si occupa delle questioni sociali dell'informatica, come il divario digitale, i temi di policy-making digitale, la privacy e i diritti umani. Il movimento del Digital Humanism, con il Manifesto di Vienna, si batte per un'informatica in armonia con i valori e i bisogni umani. Sono segnali positivi che mostrano una nuova cultura della progettazione tecnologica, in cui l'impatto etico-sociale non è più qualcosa di cui occuparsi solo a valle, ma diventa parte integrante della progettazione stessa così da superare il rigido blocco delle due culture che ha prodotto tanti danni. Tutto ciò chiaramente non è sufficiente, per quanto importante: i limiti e le difficoltà, tuttavia, non devono farci ritardare ulteriormente dall'imboccare in modo definitivo la strada verso una nuova cultura della progettazione tecnologica. Occorre andare al di là dei segnali incoraggianti e rendere questa nuova cultura non più l'eccezione, ma la regola. Le università possono essere il luogo privilegiato e il motore trainante di questo movimento. La sostenibilità delle nostre società, delle vite delle future generazioni e del nostro pianeta dipende anche da questo.

Viola Schiaffonati

Il Bello, il Brutto e il Cattivo dei LLM

Giuseppe Attardi

Sommario

I Large Language Models (LLM) sono il risultato di tre importanti progressi scientifici in soli 10 anni del Deep Learning applicato al linguaggio naturale. Illustreremo questi progressi, tra cui la soluzione allo storico dilemma sul significato delle parole. I LLM sono alla base di sistemi di Generative AI come ChatGPT, e dimostrano una sorprendente efficacia in molti compiti compresi compiti creativi come la generazione di immagini, codice o musica da descrizioni testuali. Sembrano persino esibire abilità emergenti che vanno oltre i compiti per cui sono stati allenati. I loro rapidi progressi hanno sollevato preoccupazioni su eventuali rischi di un loro utilizzo indiscriminato. Rifletteremo sulle loro potenzialità e sulle paure che sollevano, confrontando atteggiamenti apocalittici e ottimistici. Di sicuro va evitato il rischio che la tecnologia resti appannaggio di poche aziende con le risorse tecniche ed economiche per svilupparla.

Abstract

Large Language Models (LLMs) are the result of three major scientific breakthroughs in just 10 years of Deep Learning, applied to natural language. We will illustrate these advances, including the solution to the historic dilemma over the meaning of words. LLMs are the basis of Generative AI systems such as ChatGPT, and demonstrate surprising effectiveness in many tasks including creative tasks such as generating images, code or music from text descriptions. They even seem to exhibit emerging abilities in tasks besides those they were trained for. Their rapid progress has raised concerns about the possible risks of their indiscriminate use. We will reflect on their potential and the fears they raise, comparing apocalyptic and optimistic attitudes. Certainly, the risk must be avoided that this technology remains the prerogative of a few companies with the technical and economic resources to develop it.

Keywords: Large Language Models, Deep Learning, ChatGPT, Artificial Intelligence

1. Introduzione

Come mai ChatGPT si è diffuso così rapidamente da raggiungere 100 milioni di utenti in circa due mesi? Eppure, fino a poco tempo fa, le chatbot erano considerate servizi molto rudimentali, incapaci di fornire risposte adeguate e di

sostenere dialoghi coerenti con le persone. Finora, la tecnologia principale per costruirli era basata su schemi di dialogo preconfezionati (template), con domande tipiche per ciascuno scopo (intent) con corrispondenti risposte

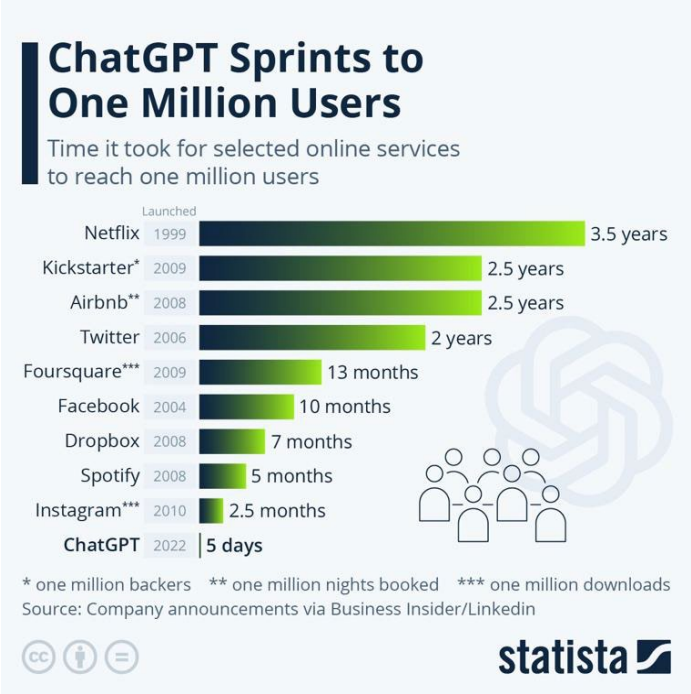

Figura 1
Diffusione di ChatGPT

(fulfillment) che i sistemi si limitavano ad adattare, inserendovi porzioni di frasi relative al tema in questione.

ChatGPT supera i limiti delle chatbot tradizionali combinando tre tecniche: un Large Language Model (GPT-3.5) di cui sfrutta la capacità di capire e generare frasi in linguaggio naturale; la messa a punto (fine-tuning) sul compito specifico di rispondere a domande; e il Reinforcement Learning per imparare a scegliere la mossa migliore di una strategia, in questo caso la risposta migliore, per raggiungere l'obiettivo, ossia di compiacere l'interlocutore.

I Large Language Model (LLM) sono modelli di reti neurali profonde (Deep Learning) in grado di acquisire una vasta conoscenza di una lingua, ricavandola da enormi quantità di testi, tratti principalmente dal Web. Essi imparano dai testi a svolgere un compito apparentemente semplice: a predire la prossima parola a conclusione di una frase. Per esempio, la pagina di Wikipedia sull'Italia riporta: "La capitale è Roma", quella sulla Francia dice: "La capitale è Parigi", ecc. Si intuisce che un LLM sappia completare la frase "La capitale è _" con la parola "Parigi", a fronte della domanda "Qual è la capitale della Francia?": il fine-tuning

gli ha insegnato la forma della risposta e l'attention, di cui parleremo dopo, a tenere conto della parola "Francia" dal contesto della domanda.

Ma le capacità dei LLM si sono presto dimostrate ben superiori alla loro intrinseca capacità di completare una frase o di comporre intere storie a partire da un breve spunto iniziale.

Il Reinforcement Learning utilizza un premio da assegnare al sistema quando la mossa che sceglie è utile a raggiungere l'obiettivo. Nel caso di ChatGPT l'obiettivo è di soddisfare le richieste dell'interlocutore, e il premio si basa sul confronto tra più risposte possibili. OpenAI, l'azienda che produce Chat-GPT, ha raccolto tantissime risposte alternative, ricavate da dialoghi con 'allenatori' umani che interagivano con la chatbot da allenare e davano un punteggio alle migliori. Il LLM di base (GPT-3.5) è stato messo a punto (fine-tuned) in modo da generare la risposta migliore coerentemente rispetto a questi esempi di risposte. OpenAI offre dettagli sul processo di sviluppo e sul ruolo dei revisori, secondo le consuete pratiche dell'azienda, in un blog[1].

ChatGPT è diventato popolare perché OpenAI ha messo a disposizione un demo online per dialogare con la chatbot nella propria lingua, anziché doverlo programmare come gli altri LLM. Milioni di persone lo hanno voluto mettere alla prova e i commenti si sono divisi tra gli entusiasti e i detrattori. I primi erano stupiti e orgogliosi di vedere una piccola creatura alzarsi in piedi e compiere i primi passi, considerandolo un momento cruciale del suo sviluppo. Gli altri si sono sforzati di farla cadere con uno sgambetto o di farla cadere dalla bicicletta, che non aveva mai provato. Cercare domande sbagliate a cui ChatGPT dà risposte sbagliate è diventato uno sport diffuso, anziché cercare le domande giuste a cui questo strumento può dare la risposta giusta. Per farsi un'opinione scientificamente valida, non bastano singoli esempi scelti appositamente, ma occorre innanzitutto capire la tecnologia e i suoi limiti per saperla sfruttare al meglio. Anche coi motori di ricerca, ci siamo rapidamente adattati ai loro limiti: sapendo che si basano sul confronto tra parole chiave della ricerca e parole presenti nei testi, abbiamo imparato a scegliere le parole giuste e a cambiarle quando non ottenevamo i risultati che ci aspettavamo.

ChatGPT è solo uno dei tanti modi di usare i LLM. La ricerca sta facendo rapidissimi progressi in questo settore e nuovi modelli vengono sviluppati in continuazione. Non dobbiamo quindi pensare che ChatGPT sia il meglio che la tecnologia possa offrire, ma solo un passo di uno sviluppo che continuerà a stupirci. Vediamo quindi quali sono i costituenti di queste nuove tecnologie, le loro qualità (il Bello), i loro limiti (il Brutto) e i potenziali rischi (il Cattivo).

2. I Large Language Model

I LLM costituiscono il secondo dei tre inaspettati breakthrough scientifici del Deep Learning applicato al Natural Language Processing, avvenuti nel breve periodo di dieci anni.

[1] https://openai.com/blog/how-should-ai-systems-behave/

Il primo breakthrough fu l'invenzione di un metodo per **rappresentare il senso delle parole** (Collobert, et al., 2011) con tecniche di apprendimento non supervisionato (self-supervised): ossia bastava fornire a una rete neurale un elevato numero di frasi, perché imparasse a cogliere somiglianze di significato tra le parole che le componevano. Ogni parola viene rappresentata da un word-embedding, un vettore di centinaia di numeri, ciascuno che in qualche modo coglie una particolare sfumatura di significato. Parole con significato simile si trovano vicine tra loro in questo spazio, ad esempio Francia, Italia e Germania sono vicine[2], facendo supporre che possano essere accomunate da qualcosa che noi chiameremmo la categoria nazione, Microsoft, Google e Apple saranno altrettanto vicine, legate forse dal concetto di azienda digitale. Categorie e concetti emergono naturalmente, come parole presenti in un certo intorno dello spazio degli embedding, anche più articolati e numerosi dei concetti che si possono ritrovare in dizionari o ontologie curate a mano. Vi sono però termini ambigui, come 'apple', il cui significato dipende dal contesto.

Su questo interviene il secondo breakthrough, con l'introduzione di un meccanismo di attenzione, descritto nell'articolo seminale "Attention is All You Need" (Vasvani, et al., 2017). Con l'attention si riescono a cogliere legami e relazioni tra le parole in un contesto e costruire i cosiddetti Transformer, ossia modelli che trasformano una sequenza di input in una sequenza di output, conservando le relazioni tra le parole. Più in generale, si tratta di reti neurali utilizzate per elaborare sequenze di dati (quindi frasi, voce, fenomeni con andamenti temporali, ecc.) che vengono però elaborati in parallelo, non sequenzialmente, per sfruttare l'accelerazione delle GPU, e che utilizzano l'attention, per tener conto della rilevanza reciproca tra gli elementi della sequenza: per esempio nella traduzione automatica, il testo originale viene

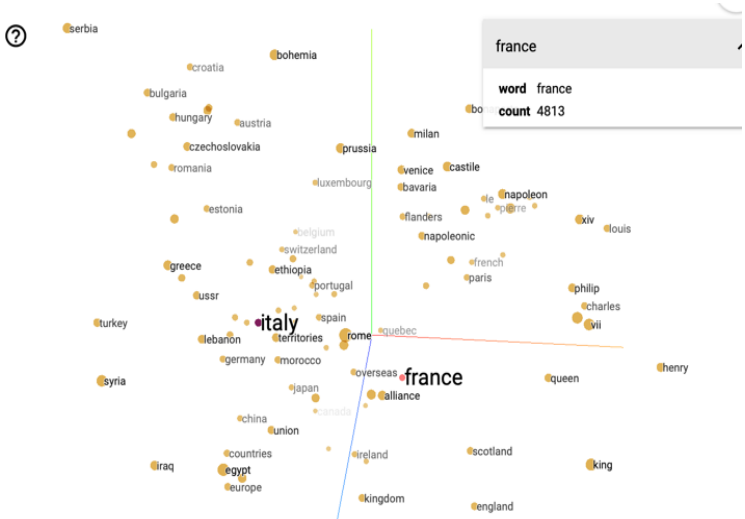

Figura 2
Visualizzazione dei word embeddings, intorno a France

[2] https://projector.tensorflow.org (provare inserendo France nella Search)

trasformato nella sua traduzione in un'altra lingua, tenendo conto del senso e della corrispondenza con le parole nell'originale.

La tecnica dei Transformer applicata alla traduzione automatica è stata uno dei più clamorosi successi del Deep Learning, che ha portato a surclassare in pochi mesi per qualità ed efficienza i precedenti sistemi di traduzione automatica che avevano richiesto anni di sviluppo e messa a punto.

I Transformer hanno poi sbaragliato tutte le altre tecniche usate in precedenza nel campo del NLP in ogni altro compito: traduzione, classificazione, riassunto, risposte a domande, analisi di opinioni, inferenza linguistica, ecc. Basta scorrere la classifica dei sistemi a confronto su SuperGlue [3], una raccolta di benchmark di analisi linguistica, per notare non solo che i migliori fanno tutti uso di Transformer, ma che molti già superano in accuratezza le capacità umane

GPT-4, il successore di GPT-3.5, è stato in grado di superare diversi testi di accesso scolastici e universitari

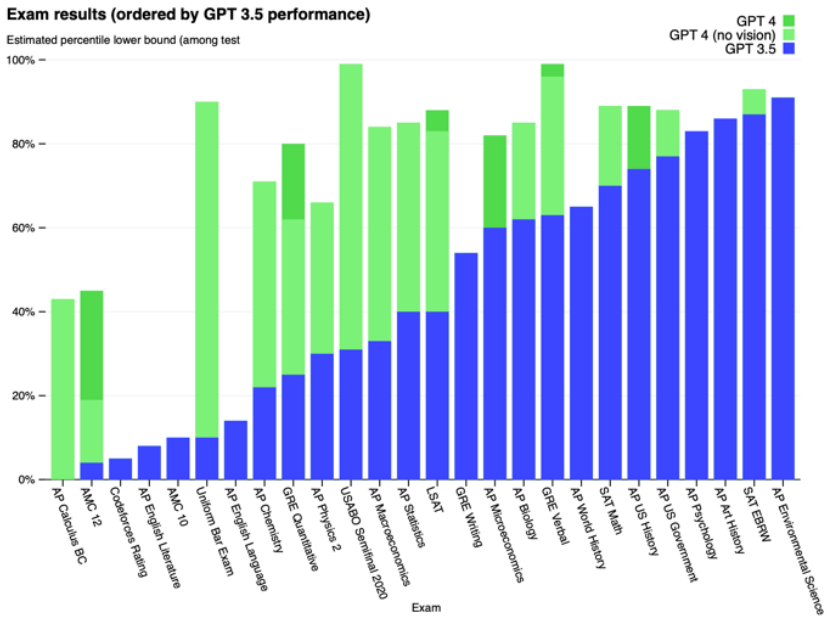

Figura 3
Punteggi di GPT-4 su vari test di esame

I Transformer possono essere adattati a nuovi compiti in modo relativamente semplice con la tecnica del fine-tuning. Si parte da un modello pre-allenato su un'ampia raccolta di testi e gli si fornisce una raccolta relativamente piccola di esempi del nuovo compito da svolgere: il modello si adatta rapidamente a svolgerlo. È un progresso notevole perché accelera lo sviluppo di nuove potenti

[3] https://super.gluebenchmark.com/leaderboard

applicazioni, sfruttando l'enorme conoscenza linguistica contenuta in un unico Transformer generico, e specializzandolo verso un nuovo compito.

3. Il Bello

I Transformer fanno parte della Generative AI, sistemi di Intelligenza Artificiale capaci di generare risposte in modo creativo, producendo risultati che sorprendono per la loro qualità che sembrerebbe tipica della mente umana: testi, immagini, musica e video possono venire generati a partire da frasi che le descrivono. Ad esempio, un testo può essere tradotto a partire dall'originale in un'altra lingua; una figura ottenuta da una descrizione della sua composizione; una musica dal testo di una canzone.

La tecnologia dei Transformer è applicabile a modalità diverse, dalla voce ai testi, dalle immagini ai video. Per questo in futuro verranno sempre più sviluppati modelli multimodali, in grado di interagire accettando input sensoriali di tipi diversi e producendo loro combinazioni, rendendo sempre più naturale l'interazione con loro.

I LLM mostrano risultati impressionanti per una serie di attività di elaborazione di testi come la risposta alle domande (QA), la generazione di codice (o altri linguaggi formali/assistenza editoriale) e la generazione di storie (fittizie).

Dai primi modelli nel 2018 ne sono apparse decine di varianti, da quelle per testo a quelli per immagini, da quelli monolingue a quelli multilingue, da quelli monomodali a quelli multimodali (testo e immagini) come GPT-4.

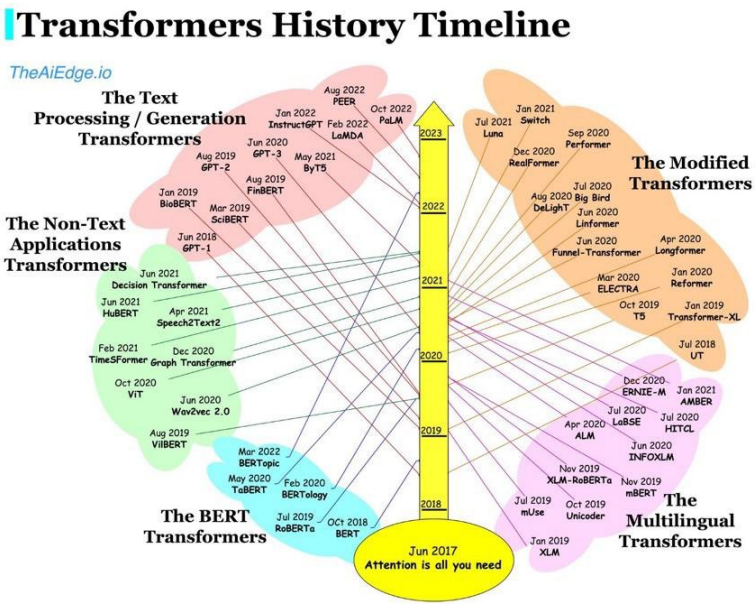

Figura 4
Evoluzione dei Transformer

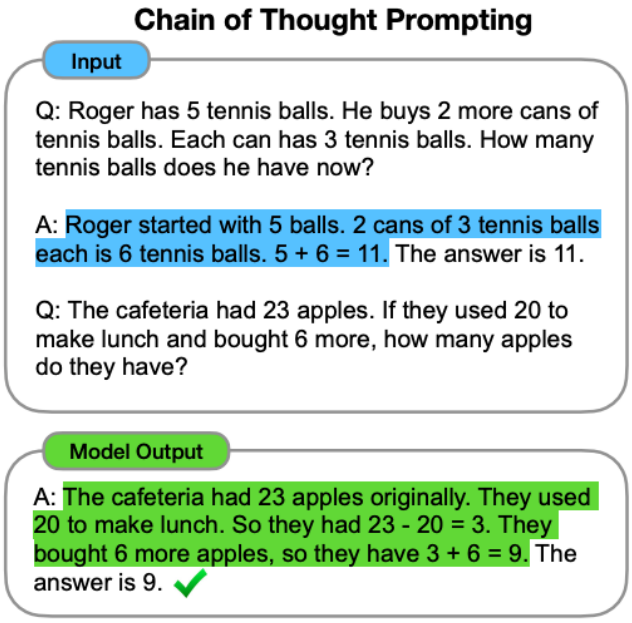

Figura 5
Esempio di ragionamento Chain of Thought

I LLM esibiscono capacità che sorprendono gli stessi ricercatori, al punto che sono diventati oggetti di studio per capire quali siano le loro capacità: un settore di studio chiamato BERTology. Tale studio si esegue stimolando i modelli con delle sonde (probe), per verificare se sanno svolgere compiti che richiedono capacità per le quali non sono stati allenati.

I LLM sembrano mostrare capacità emergenti (Wei & al., 2022), ossia che appaiono solo quando si accresce notevolmente la loro dimensione, in termini sia di dati di apprendimento che di numero di parametri di cui si compongono; capacità che non manifestano i modelli di analoga architettura ma di dimensioni più piccole. Ad esempio, modelli di dimensioni elevate cominciano a esibire capacità di ragionamento di tipo Chain of Thought, come nella figura accanto, in cui il modello risolve un problema che richiede un ragionamento matematico, seguendo la traccia indicata nella prima domanda/risposta sulle palle da tennis. Questa sorprendente potenzialità dà ulteriore stimolo a una corsa verso la costruzione di modelli sempre più grandi.

4. Il BRUTTO

I LLM costruiscono risposte a partire dalle conoscenze linguistiche che hanno accumulato nei loro parametri, non estraggono la risposta da fonti esterne. Perciò sono utilizzabili per compiti in cui questo modo di operare sia efficace, quali:

- Traduzione automatica
- Riassunto di un testo
- Sintesi di una raccolta di testi
- Comporre bozze (di articoli, mail, ecc.)
- Trasformare sequenze di un tipo in un altro (testo in immagine, voce in testo, ecc.)

Per questi compiti possono essere di valido aiuto, mentre, se si cerca di usarli per ottenere informazioni su fatti di cui hanno avuto poco sentore, possono cadere in allucinazioni (hallucinations), introducendo nella risposta elementi plausibili ma non proprio corretti.

Questo problema può essere affrontato con tecniche che guidano un modello a produrre frasi che contengono informazioni precise e corrette ricavate da fonti sicure, ad esempio con la tecnica del prompting, adottata nei sistemi di data-to-text[4] .

ChatGPT è stato allenato a non prendere posizioni su argomenti controversi, e se la cava relativamente bene se interrogato su questioni su cui esiste un'opinione prevalente. Ad esempio, se gli si chiede se i vaccini possono causare autismo, risponde che la scienza è unanime nel negare una correlazione e riporta che gli studi che ne sostenevano l'esistenza sono stati smentiti.

Questo è stato confermato da esperimenti[5] su vari benchmark, dove ChatGPT risponde correttamente a domande triviali, su fatti che ricorrono frequentemente su Internet. Invece la comprensione del testo necessaria per rispondere a una domanda complessa, magari costituite da un singolo esempio di testo, è ancora insufficiente.

D'altra parte, questo significa che ChatGPT è influenzato dalle opinioni prevalenti o più diffuse, e quindi non va usato per farsi opinioni o suffragare opinioni preconcette. Occorre sempre esercitare il proprio spirito critico e considerare le sue risposte per quello che sono, una estrapolazione dai testi su cui il sistema è stato allenato. Su molti argomenti non esiste una verità univoca e non si può certo pensare di trovarla tramite ChatGPT. Una delle stesse fonti principali su cui ChatGPT è allenato è Wikipedia: ma le informazioni riportate su certi temi nella stessa Wikipedia sono il risultato di litigi tra i curatori che cercano di imporre il proprio punto di vista.

Magari in futuro verranno prodotti chatbot che incarnano modi di pensare diversi, come avviene per le testate giornalistiche, e gli utenti potranno scegliere a quale di questi aderire per formarsi le proprie opinioni. Questo però richiederebbe che la capacità di costruire LLM diventasse più accessibile, come diremo più avanti.

I LLM non hanno inoltre capacità astratte quali quella di conteggiare, di fare calcoli, di effettuare ragionamenti logici o di pianificare in più passi. Ad esempio,

[4] https://www.amazon.science/blog/automatically-generating-text-from-structured-data
[5] https://www.geeksforgeeks.org/open-ai-gpt-3/

non sempre sanno calcolare quanto è lunga una parola o disegnare un'immagine con esattamente 5 dita delle mani o una bocca sorridente con il numero giusto di denti.

Alcuni studi hanno verificato che gli attuali LLM da una parte esibiscono davvero competenze linguistiche formali (come la conoscenza lessicale e grammaticale,

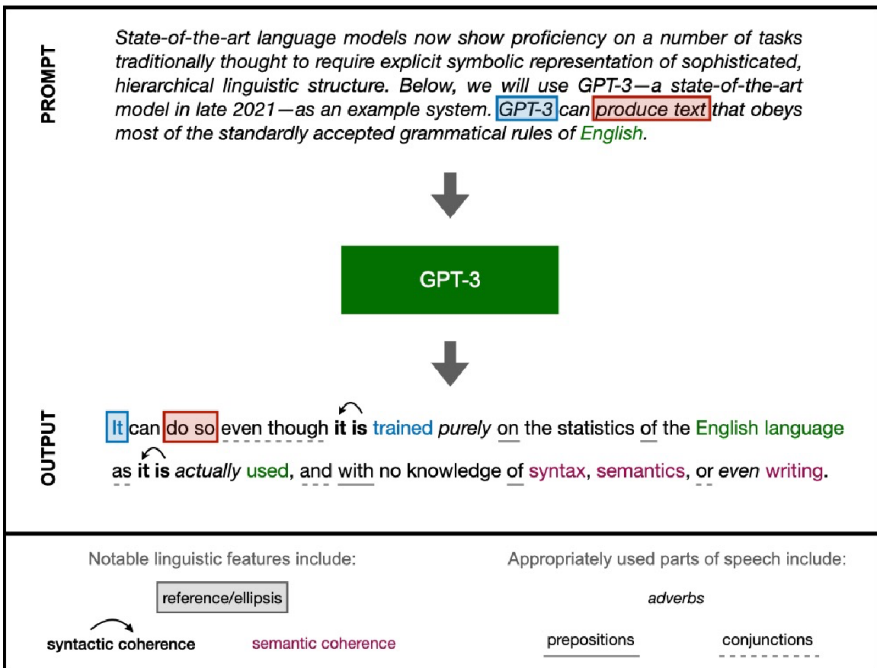

Figura 6
Competenze linguistiche dei LLM

illustrate nella figura 6), ma dall'altra sono privi di competenze funzionali (richieste per svolgere calcoli matematici o ragionamento logico) (Mahowald, et al., 2023).

Ciò non dovrebbe stupire perché essi non sono stati allenati per eseguire ragionamenti astratti, ma solo per prevedere la prossima parola.

ChatGPT per esempio è stato allenato a gestire dialoghi, e quindi a tenere traccia dell'intera conversazione, rispondendo a tono, a volte scusandosi gentilmente se gli si segnala un errore e fornendo una nuova risposta per correggersi.

Questo fa sembrare che ChatGPT impari attraverso i dialoghi: in realtà ciò di cui tiene conto è limitato alla conversazione in corso, ma alla prossima avrà dimenticato tutto. OpenAI sollecita gli utilizzatori a inviare loro feedback sulle risposte, al fine di migliorare il modello, ma ciò avviene con l'aggiunta di nuovi esempi alla raccolta usata per il passo di Reinforcement Learning, che richiede settimane o mesi di allenamento e viene fatto quindi solo di tanto in tanto.

ChatGPT ha sollevato perplessità su possibili effetti che il suo utilizzo potrebbe avere sulla scuola, con studenti che si fanno produrre risposte o saggi da ChatGPT esimendosi dallo studio; sul mondo dell'informazione, sostituendo i giornalisti nella stesura di notizie. Altri sistemi come DALL-E 2[6] potrebbero avere impatti nel mondo creativo, sostituendo gli illustratori con strumenti che generano automaticamente immagini o produrre musica e video. Di recente è stata minacciata una causa contro [7] l'azienda che produce StableDiffusion [8] , sostenendo che utilizza immagini di apprendimento ottenute in violazione del copyright.

Più grande è il LLM, più difficile diventa, sia per gli esseri umani che per tecniche algoritmiche, distinguere le notizie scritte da una macchina dagli articoli scritti da esseri umani. Su come comportarsi di fronte a tali situazioni le opinioni sono divergenti, se bandirne l'uso o controllarlo ad esempio con tecniche di watermarking.

5. Il CATTIVO

È ben noto che le applicazioni di AI generativa come le chatbot a volte possono essere difficili da controllare e si può finire in conversazioni in cui vomitano commenti razzisti o sessisti. OpenAI ha affrontato questo problema identificando contenuti tossici o semplicemente su temi politici controversi e cercando di intercettarli a priori.

Di fatto oggi gli unici che possono permettersi le **enormi risorse di calcolo necessarie per allenare un LLM** sono le grandi aziende tecnologiche. E il loro ulteriore sviluppo e diffusione richiede investimenti massicci, come dimostrano i $20 miliardi che Microsoft ha annunciato di voler investire in OpenAI e nell'integrazione di ChatGPT con il suo motore di ricerca Bing. In questo settore stiamo per assistere a una **guerra tra titani**, per conquistare spazi in un nuovo settore applicativo: Microsoft con il sistema Prometheus[9] contro Google con Bard[10]. Il passo da una chatbot a un sistema integrato di dialogo e ricerca è tutt'altro che banale, come emerge dai primi passi falsi di entrambi i sistemi, e richiederà una riprogettazione sostanziale dell'architettura del sistema integrato.

Sarà una battaglia cruciale con **effetti dirompenti anche sull'ecosistema digitale del web:** infatti finora i motori di ricerca guadagnavano sulla pubblicità che attraevano sfruttando l'interesse per i contenuti che altri introducevano nei loro siti web. Questi ultimi venivano a loro volta remunerati con un aumento di traffico e una quota di entrate pubblicitarie. Ma con i chatbot che producono direttamente le risposte senza fare riferimento alle fonti, si spezza questo cordone ombelicale

[6] https://openai.com/dall-e-2/
[7] https://www.theverge.com/2023/1/17/23558516/ai-art-copyright-stable-diffusion-getty-images-lawsuit
[8] [8] https://stability.ai/blog/stable-diffusion-public-release
[9] [9] https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/
[10] https://blog.google/technology/ai/bard-google-ai-search-updates/

che alimenta i produttori di contenuti. Gli **effetti di questo cambiamento di paradigma sono del tutto imprevedibili.**

Ci sono due strade possibili per rendere accessibile e **democratizzare la tecnologia dei LLM:** progetti dal basso che aggregano una comunità di ricercatori nello sviluppo di modelli Open-Access, come BLOOM (Le Scao & al., 2022), o costruire e rendere disponibili ai ricercatori infrastrutture pubbliche dotate di risorse di calcolo adeguate, come chiedono a gran voce i ricercatori stessi sia in USA[11] che in Europa[12].

6. LE PAURE

Come ogni nuova tecnologia di largo impiego, anche i LLM suscitano reazioni contrastanti, dalle paure apocalittiche all'ottimismo sfrenato.

Persino i tre ricercatori, considerati i padri del Deep Learning, hanno preso posizione, in una specie di tentativo di rimettere il genio nella bottiglia.

Yoshua Bengio ha sottoscritto una lettera aperta del Future of Life Institute (FOLI), in cui si chiede una moratoria di sei mesi nello sviluppo di ulteriori più potenti LLM, finché non vengano definite nuove norme sul loro utilizzo, anche se è scettico che la lettera abbia alcun effetto e consideri inadatte le norme di regolamentazione attualmente proposte.

Geoff Hinton ha invece interrotto la sua collaborazione con Google, a cui ha venduto dieci anni fa la sua startup DNNresearch, oltre che per ragioni di età, anche per poter essere libero di esprimersi sui rischi dell'AI. Ha ribadito che Google si è finora comportato in modo responsabile nell'utilizzo dell'AI e continua a credere nell'importanza degli studi in materia. Finora, come molti altri, riteneva che la possibilità di costruire sistemi più intelligenti delle persone fosse lontano di 30 o 50 anni, mentre ora si è ricreduto.

I rischi che intravede sono nella diffusione su larga scala di fake-news, nell'eliminazione di posti di lavoro e infine nell'utilizzo per lo sviluppo di armi letali autonome.

Invece Yann LeCun non ha firmato la lettera del FoLI, sostenendo che la tecnologia è tuttora in evoluzione e come tutte le nuove tecnologie, forme di controllo e di sicurezza dovranno venire introdotte man mano che si sviluppa.

Le questioni segnalate da Hinton sono state ampiamente discusse negli anni scorsi e pericoli simili sono stati attribuiti anche ad altre tecnologie introdotte in passato. Ricordo, ad esempio, con quanta sufficienza e preoccupazione i media trattavano la nascente tecnologia di Internet una trentina di anni fa. Le preoccupazioni di oggi riguardano quindi più in generale l'uso responsabile delle tecnologie. Viene da chiedersi dunque cosa ci sia di particolare nei LLM che sta facendo concentrare l'attenzione di governi e istituzioni sulla loro regolamentazione.

[11] https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf
[12] https://claire-ai.org/vision/

Le fake news sono sempre esistite ed il problema principale è riconoscerle e bloccarne la diffusione, non tanto impedire che vengano prodotte. Lo spazio pubblico è già saturo di frodi ed è difficile immaginare come l'AI possa renderlo molto peggiore. Il numero che conta non è quello di quante ne vengano prodotte, ma di quante raggiungono l'obiettivo di una diffusione virale, che non è facilmente prevedibile, tanto meno se queste vengono prodotte in automatico un tanto al kilo. Il timore delle deep fake (foto fittizie ad alto realismo) ignora il fatto che PhotoShop è in uso da decenni con lo stesso obiettivo, e gli stessi media tradizionali ne fanno abbondante uso.

Hinton afferma di essere rimasto spiazzato dalle capacità raggiunte dai LLM in poco tempo. In effetti la sua ricerca è stata latente per trenta anni ed è esplosa negli ultimi dieci. Ma lo sviluppo esponenziale delle tecnologie informatiche non è una novità: ne avvengono ogni 15 anni ed hanno effetti dirompenti e sostanzialmente positivi per tutti. Perché l'AI dovrebbe essere diversa e più pericolosa di altre? Perché l'AI produce sistemi più capaci degli umani? Ma in molti compiti i computer sono già ampiamente superiori agli umani. Perché l'AI potrebbe riprodurre se stessa? Ma i compilatori non sono altro che programmi che scrivono programmi. Perché l'AI potrebbe ritorcersi contro gli umani? Ma questi sono scenari da fantascienza, nessun sistema potrebbe diventare autonomo se qualcuno non gli attribuisce questa capacità: i LLM al massimo possono dire sciocchezze ma non possono fare male fisico. Stranamente le norme proposte di regolamentazione dell'AI, come l'European AI Act, escludono invece dal loro ambito di applicazione i sistemi di utilizzo militare. Non è ridicolo che non si vogliano contrastare proprio le applicazioni più pericolose?

Alcuni criticano i LLM sostenendo che non sono in grado di capire. Ma l'informatico Yoav Shahom, in un recente seminario su "Understanding understanding[13]" afferma che tali critiche sono troppo vaghe, fintanto che non si definisce cosa significhi "capire". Finora, l'unico criterio pratico per stabilire se qualcuno, studente o computer, capisce, è di porgli delle domande su un tema di cui sappiamo le risposte. Ma su tutti i test di "comprensione" i LLM superano ormai abbondantemente il livello umano. Del resto, anche Richard Feynman diceva che "nessuno può dire di capire la fisica quantistica"; eppure le sue equazioni funzionano. Quindi ciò che conta è se l'AI funziona, non se capisce.

D'altro lato i LLM esibiscono capacità emergenti, ossia che appaiono solo quando raggiungono grandi dimensioni e che i modelli più piccoli non hanno. È un fenomeno simile a quello che il Nobel Giorgio Parisi analizza nei sistemi complessi, il cui funzionamento è determinato dalla combinazione su larga scala di semplici leggi probabilistiche, come quello dei LLM di saper predire la prossina parola. La mente umana ha difficoltà a spiegare i fenomeni complessi perché siamo abituati a scomporre i fenomeni in piccole parti legate da relazioni di causa-effetto.

[13] https://hai.stanford.edu/events/yoav-shoham-understanding-understanding

Ciò che stupisce è che si esprimano in maniera melodrammatica con affermazioni facilmente confutabili, anche esperti della materia come Gary Marcus e Noam Chomsky[14] . Soprattutto è insensato che a partire da un singolo esempio di errore nella risposta di un LLM, si arrivi a conclusioni generalizzate su quanto mai si possa realizzare tramite il Machine Learning.

I tentativi attuali di regolamentazione come lo European AI Act sono goffi e complicati: un testo di 107 pagine, che anziché limitarsi a stabilire dei principi o dei diritti, si avventura nell'impostare un farraginoso processo di certificazione per garantire gli utenti contro i potenziali danni dei sistemi di IA. Ma siccome non si può effettuare tecnicamente la certificazione dei sistemi o degli sviluppatori, si limita a introdurre norme e verifiche sul processo di sviluppo di tali applicazioni. La certificazione dovrebbe essere svolta attraverso informazioni che le aziende stesse forniscono, visto che si vuole garantire la segretezza della proprietà industriale dei prodotti. Il processo di certificazione è estremamente laborioso e si stima costi intorno ai 300 mila €, senza contare che dovrebbe essere replicato in ogni paese dell'Unione, e ciascun paese dovrebbe dotarsi di un apposito ente di certificazione, dotato di "risorse adeguate". La direttiva stessa riconosce questo problema di costi e per non penalizzare le piccole aziende, propone di introdurre delle "regulatory sandboxes", un misterioso sistema per provare in un ambiente controllato il funzionamento delle applicazioni, da realizzare in ciascun paese. Ma le radici stesse della normativa sono in dubbio, in quanto definisce il settore dell'AI non in termini di ciò di cui si occupa o di cosa faccia, ma delle tecniche che usa, come dire che l'oftalmologia viene definite dall'uso delle lenti e non come lo studio della visione.

Infine, l'European AI Act[15] esclude esplicitamente dal suo campo di intervento le applicazioni militari, che sono quelle che davvero producono morte e danni, mentre un LLM non può causare danni, al massimo può dire qualche sciocchezza.

Un'altra paura è quella della perdita di posti di lavoro, che Goldman Sachs stima in 300 milioni di posti di lavoro solo in USA e in Europa. Altri all'opposto sostengono che altrettanti posti verranno rimpiazzati da nuovi lavori, relativi a nuovi prodotti o servizi basati su AI, come è successo con l'introduzione di altre tecnologie in passato. Francamente non saprei fare delle stime, ma sono convinto che l'AI avrà un impatto significativo sul mondo del lavoro, in quanto si tratterà di una General Purpose Technology, che cambierà il modo di svolgere moltissime attività umane. Inoltre, i cambiamenti del digitale sono molto più rapidi di quelli delle tecnologie del passato; quindi, non ci sarà tempo sufficiente perché i lavoratori si riqualifichino per le nuove attività. Ciò a cui si assiste già adesso è una divaricazione, tra lavori super-specializzati e ben remunerati, ma poco numerosi e un alto numero di lavori di scarso livello e poco pagati, al servizio delle macchine, della cosiddetta Gig-economy[16]. Questo produrrà un enorme divario

[14] https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html
[15] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206
[16] https://en.wikipedia.org/wiki/Gig_economy

tra lavoratori e un corrispondente divario di potere economico e finanziario nelle mani delle poche grandi aziende digitali che controllano le piattaforme e i servizi digitali. Questi due divari saranno la causa di possibili disuguaglianze sociali su cui bisognerà intervenire per tempo.

Tranne la questione delle disuguaglianze e dell'impatto sul lavoro, ritengo che le altre paure siano ingiustificate e siano relative a fenomeni che già esistevano prima dell'irruzione sulla scena dei LLM.

Si tratta di questioni che riguardano l'impatto economico e sociale dell'utilizzo di nuove tecnologie, di cui sono responsabili sia le aziende che gli utenti. Ad esempio le fake-news esistevano prima dei social media, e a diffonderle attraverso i social media sono gli utenti stessi, con la complicità dei media che guadagnano sulla pubblicità che cresce con l'interesse che esse suscitano.

Gli informatici si devono sentire in dovere di segnalare alla società l'importanza e il ruolo che le nuove tecnologie possono avere e di chiedere di investire nella ricerca per sviluppare e migliorare tali tecnologie. Ma il più delle volte gli scienziati non sono in grado di prevedere gli effetti delle innovazioni, come diceva Rodney Brooks[17], nel 2017, mentre oggi si esagera al contrario. Nessuno sapeva predire quali sarebbero stati gli effetti delle precedenti General Purpose Technologies, sviluppate dall'informatica: nel 1980 il personal computer e nel 1995 Internet. Eppure, alcuni ne avevano segnalato gli effetti dirompenti[18]. Ma se analizziamo le preoccupazioni e le ipotesi di regolamentazione che venivano proposte all'epoca, ci rendiamo conto di quanto fossero fuori obiettivo. Esse avrebbero bloccato i benefici e non risolto il problema più serio della concentrazione di potere tecnologico a cui assistiamo oggi. In altre parole, bisogna padroneggiare la tecnologia, fare in modo che sia disponibile a tutti e non concentrata in poche mani, e seguirne le evoluzioni per adattare la società ai cambiamenti che essa comporta.

7. CONCLUSIONI

ChatGPT ha portato all'attenzione del vasto pubblico la tecnologia dei LLM, che sono alla sua base, come di tante altre possibili applicazioni. La sua capacità di interagire e di rispondere in linguaggio naturale in più lingue ha fatto insorgere curiosità e lasciare intendere che esibisca capacità paragonabili all'intelligenza umana. Tramite esperimenti scientifici controllati, è stato possibile verificare che questo può succedere, ma solo in compiti di trattamento di testi molto specifici, mentre i LLM sono ben lontani dalle capacità della cosiddetta Artificial General Intelligence (AGI).

Ma i progressi rapidi della tecnologia fanno immaginare ulteriori miglioramenti, sia incrementali, sia dovuti ad ulteriori eventuali breakthrough, visto il sempre maggior interesse che queste tecnologie stanno suscitando nei ricercatori e negli investitori. Limitarsi quindi alle critiche per gli attuali limiti della tecnologia non

[17] https://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future-of-ai/
[18] http://www.interlex.it/attualit/letterap.htm

tiene conto che ChatGPT non sarà l'ultimo dei modelli e se si guardasse alla velocità dei progressi recenti, potremmo presto stupirci per nuovi risultati in arrivo.

Occorre però evitare che il controllo della tecnologia rimanga appannaggio di poche grandi imprese che possono permettersi le enormi risorse necessarie per costruire i modelli più sofisticati e che questo consenta loro di aumentare il loro dominio sull'economia digitale.

In un caso o nell'altro, l'evoluzione tecnologica dei modelli di AI porterà a cambiamenti dirompenti nel modo di sviluppare applicazioni, nella concentrazione di potere tecnologico e nella disparità tra i detentori della tecnologia e gli altri, e infine nel mondo delle professioni.

BIBLIOGRAFIA

Collobert, R., J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, & P. Kuksa. (2011). Natural Language Processing (Almost) from Scratch. JMLR. Tratto da https://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf

Vasvani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I.

(2017). Attention is all you need. Neurips 2017. Curran.

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., &

Fedorenko, E. (2023, January 23). Dissociating language and thought in large language models: a cognitive perspective. Tratto da ArXiv: https://arxiv.org/abs/2301.06627

Le Scao, T., & al., e. (2022). BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. Tratto da ArXiv: https://arxiv.org/pdf/2211.05100

Wei, J., & al., e. (2022, 12). Emergent Abilities of Large Language Models. Tratto da ArXiv: https://arxiv.org/pdf/2206.07682

BIOGRAFIA

Giuseppe Attardi è stato professore ordinario di Informatica presso l'Università di Pisa. Ha anche lavorato presso l'AI Lab del MIT, il Sony Paris Research Laboratory, l'ICSI a Berkeley e lo Yahoo Research Barcelona. Ha sviluppato Omega, una logica descrittiva; CMM, il Garbage Collector utilizzato in Java e DeSR, un parser multilingue con reti neurali. Ha partecipato allo sviluppo di Arianna, il primo motore di ricerca italiano e ha introdotto la tecnica della categorizzazione per contesto delle pagine web. È fondatore o socio di alcune startup, in Italia e in Spagna. Ha contribuito alla realizzazione delle reti in fibra

ottica dell'Università di Pisa e del GARR e a promuovere l'accesso a Internet in Italia. Ha guidato lo sviluppo della piattaforma cloud GARR. Ha contribuito alla stesura della strategia italiana sull'Intelligenza Artificiale e alla nascita del primo Dottorato di Ricerca nazionale in Intelligenza Artificiale.

E-mail: attardi@gmail.com

Un mondo di transistor

Fabrizio Luccio

Sommario

Il transistor nato settantacinque anni fa ha contribuito a cambiare le attività umane come pochissime altre invenzioni nella storia. La sua realizzazione nei circuiti integrati ha segnato nel tempo una spettacolare riduzione delle dimensioni e del consumo di energia, con conseguente aumento del numero di transistor per chip e moltiplicazione delle applicazioni cui si rivolge. Seguiamo qui la sua storia e lo stato dell'arte oggi, indicando le prospettive di sviluppo per un prossimo futuro e i limiti fisici che a queste si imporranno.

Abstract

The transistor, born seventy-five years ago, has contributed to changing human activities like very few other inventions in history. Its implementation in integrated circuits has marked a spectacular reduction in size and energy consumption over time, with a consequent increase in the number of transistors per chip and an increasing variety of the applications where transistors are used. We follow the transistor history and the state of the art today, and indicate expectations for the near future and the physical limits that will be imposed on them.

Keywords: Semiconductor, MOSFET, FGMOS, Moore's law, Integrated circuit

1. Introduzione

L'invenzione del *transistor* fu annunciata ufficialmente settantacinque anni fa e da allora più di ogni altra ha contribuito a incidere sulla vita dell'uomo consegnandogli importanti oggetti inimmaginabili prima e contribuendo a renderne più efficienti e fruibili innumerevoli altri. I *transistor* sono divenuti così piccoli e così numerosi all'interno di uno stesso supporto che, in un paese mediamente sviluppato, qualunque abitazione è invasa da miliardi di transistor senza che alcuno degli umani presenti se ne accorga: e se uno di essi va dal tabaccaio a comprare per pochi euro una *chiavetta USB* (tecnicamente, *una flash memory*) se ne mette in tasca qualche centinaio di miliardi. Naturalmente lo straordinario sviluppo dei transistor è stato graduale ma sempre più rapido, al punto che il contenuto del presente articolo, fatta salva la descrizione di quanto accaduto finora e il confronto tra qualche nostra previsione e la realtà futura, sarà completamente superato in brevissimo tempo.

Il nome, derivato dalla contrazione dei due termini "transconductance resistor", indica un dispositivo atto a controllare il valore di una resistenza

elettrica, ma forse ciò è ancora oggi ignoto a chi non si interessa di scienza o di tecnologia. Le persone non giovanissime come chi scrive (è un eufemismo) ricorderanno che negli anni '60 del secolo scorso "il transistor" indicava invariabilmente la radiolina portatile a batteria che gracchiava ad alto volume nelle spiagge e nei giardini pubblici, in particolare la domenica nell'ora in cui si giocavano le partite di calcio. Oggi è correttamente di moda la parola *chip*, piccolo e in sostanza misterioso componente di circuiti elettronici che di transistor ne contiene moltissimi: moda innescata dalla pandemia di Covid 19 e non solo da essa, che ne ha causato una grave penuria rendendo a sua volta introvabili in breve tempo tanti dispositivi che li includono per il loro funzionamento, dalle automobili alle lavastoviglie per non dire ovviamente dei computer.

Seguiamo dunque la storia dei transistor dalle origini ai giorni nostri parlando dei principi fisici su cui sono basati e del loro sviluppo negli anni; dell'integrazione nei chip e dei metodi di fabbricazione; dei problemi legati alla velocità di funzionamento e consumo di energia; dello sviluppo dell'industria che li riguarda; dei limiti fisici sulla loro miniaturizzazione che condizionerà l'evoluzione dei chip.

2. Dalla punta di contatto al FinFET

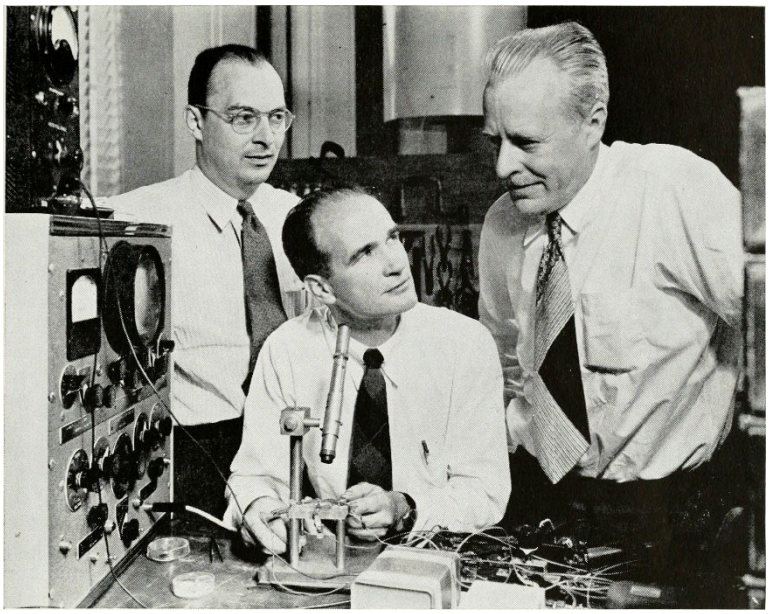

Figura 1
Gli inventori del transistor

Nella famosa immagine di figura 1, rilasciata dalla AT&T ai tempi dell'invenzione del transistor, appaiono i tre scienziati cui nel 1956 fu attribuito il premio Nobel per la fisica "per le loro ricerche sui semiconduttori e la loro scoperta dell'effetto transistor". Sono da sinistra John Bardeen, Walter Brattain e William Shockley che

che lavoravano presso i Bell Labs della AT&T ai tempi della scoperta[1]. Tre caratteri molto diversi tra cui si crearono spiacevoli dissapori dopo il 16 Dicembre 1947 quando il primo transistor della storia prese a funzionare tra le mani di Bardeen e Brattain che ne fecero una richiesta di brevetto. L'invenzione fu resa pubblica in una conferenza stampa presso i Bell Labs nel Giugno 1948: si trattava del *point-contact transistor (transistor a punta di contatto*). Shockley, che era il capo del gruppo in cui si svolgevano quelle ricerche e aveva preso malissimo l'assenza del suo nome dalla richiesta originale di brevetto continuò da solo i suoi studi e nel Dicembre del 1948 completò il progetto del nuovo *bipolar junction transistor (BJT, transistor bipolare o a giunzione*) che avrebbe dominato il campo fino agli anni '70 ed è tuttora, dopo una naturale evoluzione nel tempo, alla base dei circuiti analogici. Due dispositivi diversi dai *field effect transistor* (FET, *transistor a effetto di campo*) di cui anche parleremo, che dominano oggi l'elettronica digitale.

Accenneremo solo alle proprietà e al funzionamento dei transistor che molti lettori conosceranno già, ma vogliamo ricordare anzitutto alcune caratteristiche dei materiali impiegati per comprendere i limiti fisici che la tecnologia comincia ad affrontare oggi. I transistor sono costruiti su materiali semiconduttori che hanno una conduttività elettrica intermedia tra quelle dei conduttori e degli isolanti. Ricordiamo che i valori di questo parametro sono distantissimi tra conduttori, semiconduttori e isolanti, e quello dei semiconduttori in forma di cristallo varia fortemente in funzione di elementi chimici in essi presenti. In particolare consideriamo il silicio cristallino che oggi domina le applicazioni e il germanio su cui furono costruiti i primi transistor.

L'atomo di silicio ha quattro elettroni di valenza. Il cristallo, come quello del diamante, è cubico a facce centrate: ogni atomo si trova al centro di un tetraedro i cui vertici ospitano ciascuno un altro atomo di silicio (Si) con cui il primo scambia un legame covalente attraverso i suoi quattro elettroni di valenza. Questa struttura estremamente regolare può essere "drogata" con atomi di elementi prossimi nella tavola periodica: per il silicio i droganti sono il fosforo (P) e il boro (B) che di elettroni di valenza ne hanno rispettivamente cinque e tre. Per ottenere il drogaggio, atomi ionizzati di fosforo o boro sono spinti da un campo elettrico sulla superficie del cristallo e lo penetrano a profondità controllata sostituendo atomi di silicio nella struttura cristallina.

In un modello idealizzato del fenomeno, sostituendo un atomo di Si con P, l'elettrone di valenza di P che non trova posto nel legame chimico è libero di spostarsi nel cristallo: se questo è soggetto a un campo elettrico esterno l'elettrone contribuisce a una corrente elettrica complessiva la cui intensità dipende dall'entità del drogaggio. Il materiale così ottenuto si chiama silicio-n ove

[1] A parte la profondità di pensiero, John Bardeen era noto per la gentilezza di tratto. È stato l'unico studioso a ottenere per due volte il premio Nobel per la fisica: il secondo nel 1972 per "la teoria generale della superconduttività". Lasciata la AT&T divenne professore della University of Illinois dove l'autore di questo articolo ebbe l'onore di incontrarlo come collega. Da tempo non è più tra noi.

n sta per negativo. Similmente sostituendo un atomo Si con B l'elettrone di valenza mancante in B genera una lacuna, e questa può essere invasa da un elettrone di un legame vicino tra altri due atomi Si lasciando una lacuna tra questi: il fenomeno ripetuto a catena è descritto come flusso di lacune anche se di fatto è un flusso di elettroni in senso inverso, e contribuisce anch'esso a una corrente elettrica se il cristallo è soggetto a un campo elettrico esterno. Il materiale così ottenuto si chiama silicio-p ove p sta per positivo: ecco dunque realizzati due semiconduttori con proprietà in certo senso opposte. Il materiale di base dei primi transistor era il germanio (Ge), impiegato dall'inizio del 1900 a oggi per la costruzione di diodi. Gli elementi di drogaggio che danno luogo a germanio-n e germanio-p possono essere l'arsenico (As) e il gallio (Ga).

Il motivo per cui ricordiamo questi aspetti del drogaggio, probabilmente noti a tutti i lettori, è per meglio spiegare i limiti fisici che la tecnologia dei transistor inizia a incontrare oggi. Anzitutto un cristallo di silicio contiene in ordine di grandezza 10^{22} atomi/cm^3. Ciò significa che il *raggio di Van der Vaals*, ossia il raggio di una sfera ideale occupata da un atomo del cristallo, è di 210 *pm* = 0.21 *nm* (ricordiamo che $p = 10^{-12}$ e $n = 10^{-9}$: come vedremo le dimensioni dei transistor di oggi si avvicinano a questi valori e quelli che si chiamavano *micro-circuiti* sono ormai *nano-circuiti*). Un cristallo idealmente puro ha conduttività elettrica bassissima ed è praticamente isolante, ma cristalli puri non è possibile ottenerli artificialmente: in quelli utilizzati oggi le impurità inevitabili sono dell'ordine di 10^9 atomi/cm^3, quantità sufficientemente bassa per considerarli isolanti per i livelli di segnale impiegati. L'entità media del drogaggio indotto artificialmente è invece dell'ordine di 10^{15} atomi/cm^3, quindi silicio-*p* e silicio-*n* sono conduttori, anche se non buoni.

Una struttura di base per la costruzione del transistor è la *giunzione*, cioè la superficie di contiguità tra un cristallo di silicio-*p* e uno di silicio-*n*, nei fatti la contiguità tra due zone drogate *p* e *n* dello stesso cristallo. Le due zone, elettricamente neutre, si caricano rispettivamente in modo negativo e positivo presso la giunzione a causa di una diffusione di elettroni dalla zona *n* alla zona *p* e di lacune in senso inverso. Questo fenomeno crea attorno alla giunzione una differenza di potenziale (*barriera*) tra le due zone, negativa in *p* e positiva in *n*, che si stabilizza a un *valore di soglia* di circa 0.6 Volt per il silicio di 0.3 Volt per il germanio, e contrasta l'ulteriore diffusione di cariche. Una struttura così formata ha comportamento di un diodo anche se non ideale perché inizia a condurre per una tensione esterna di polarità concorde alla barriera e di valore superiore alla soglia, e non conduce per una tensione esterna inversa (non troppo elevata perché distruggerebbe il dispositivo).

La giunzione, caratteristica essenziale dei diodi a semiconduttore, è presente in tutti i transistor a cominciare dal primo a punta di contatto. Questo, come tutti gli altri, è connesso all'esterno con i tre terminali *emettitore-base-collettore* (successivamente chiamati *source-gate-drain*), anche se l'impiego di essi differisce tra i vari tipi di transistor. Parliamo anzitutto brevemente del primo che ha un interesse unicamente storico e ha un funzionamento piuttosto complicato che probabilmente all'inizio non fu interamente compreso neanche

dagli inventori. Il transistor a punta di contatto mostrato schematicamente nella figura 2 è costituito da una piastrina di germanio la cui parte inferiore drogata *p* è connessa a un conduttore metallico a terra che costituisce la base, e sulla parte superiore drogata *n* poggiano, vicinissime tra loro, le punte di due lamine d'oro come emettitore e collettore (l'impiego dell'oro, materiale non ossidabile, è comune anche oggi perché assicura la costanza di buoni contatti nel tempo). Tra le zone *n* e *p* si stabilisce una giunzione che consente un passaggio di corrente tra emettitore e base se il primo è polarizzato positivamente superando la barriera, ma che in linea di principio blocca una corrente in senso inverso tra collettore e base. Tuttavia le lacune iniettate nella zona *p* dall'emettitore sono attratte in gran numero dalla tensione negativa del vicinissimo collettore tendendo ad abbattere localmente la barriera. L'effetto complessivo è quello che, per opportuni valori delle tensioni in gioco, la corrente emettitore-base causa una corrente collettore-base in senso opposto modulata dai valori della prima, le cui variazioni possono essere molto più intense di quelle del circuito dell'emettitore. Era così realizzato il sogno inseguito da molti a quel tempo di ottenere l'amplificazione di un segnale impiegando un semiconduttore anziché un tubo a vuoto come il triodo.

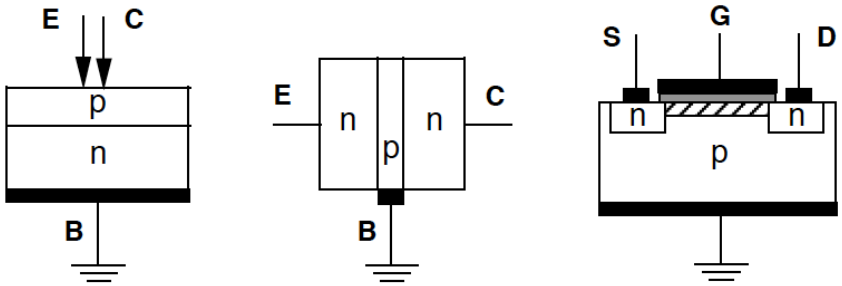

Figura 2

Schizzo dei transistor. Da sinistra: a punta di contatto, bipolare, nMOS. E, B, C indicano Emettitore, Base, Collettore. S, G, D indicano Source, Gate, Drain. Le zone nere sono contatti metallici, la zona grigia è ossido SiO2 isolante, la zona tratteggiata è il canale

Il transistor a punta di contatto ha il merito indiscutibile di essere sato il primo a funzionare e fu prodotto fino al 1966 dando origine al declino dei tubi a vuoto; ma il transistor bipolare (BJT) nato subito dopo e entrato in commercio all'inizio degli anni '60 era destinato a un successo molto superiore per i livelli di potenza con cui poteva operare e per una struttura sostanzialmente meno fragile perché evita il contatto critico tra il semiconduttore e le punte, tanto da essere prodotto fino a oggi dopo una logica evoluzione nel tempo. Nel transistor bipolare mostrato schematicamente nella figura 2 il semiconduttore, prima germanio e oggi quasi esclusivamente silicio, è diviso in tre zone consecutive drogate *p-n-p* o *n-p-n*, con le due estreme collegate a emettitore e collettore e quella centrale, molto più sottile e meno drogata delle altre, collegata alla base: al suo interno vi sono dunque due giunzioni come in tutti i transistor che l'avrebbero seguito e, a quanto possiamo prevedere, come avverrà ancora nel prossimo futuro. Una

corrente locale tra base e emettitore ne controlla una assai più intensa tra collettore e emettitore. Per spiegarne a grandi linee il funzionamento consideriamo il tipo *n-p-n*.

Una tensione di controllo negativa applicata all'emettitore attrae elettroni nella sottile zona *p* attraverso un corrente tra base ed emettitore combinandosi con le poche lacune della zona e liberando il passaggio di elettroni nell'intero corpo del transistor. Questi sono attratti in gran numero da una tensione positiva tra collettore ed emettitore assai superiore di quella tra base ed emettitore. La modulazione della corrente base-emettitore controlla e amplifica quella collettore-emettitore.

I transistor a punta di contatto e bipolari sono particolarmente adatti all'impiego in circuiti analogici, ma la possibilità puramente binaria di permettere o vietare la circolazione di una corrente in un circuito ne consentiva ovviamente l'impiego nei sistemi digitali. Molto più adatto per questi ultimi sono i transistor di tipo MOS, o più propriamente MOSFET per *metal-oxide-semiconductor field-effect-transistor* o transistor a effetto di campo. Inventato anch'esso nei Bell Labs nel 1959, il MOS fu posto in commercio qualche anno più tardi e, in versioni via via più evolute, ha guadagnato rapidamente una completa supremazia nelle applicazioni digitali ove anche oggi è il più largamente impiegato. È anch'esso di norma composto di silicio con tre zone drogate *p-n-p* (detto pMOS) o *n-p-n* (detto nMOS): contiene quindi due giunzioni e ha tre terminali che prendono ora il nome di *source*, *gate* e *drain* (*S, G, D*), ma diversamente dai precedenti la corrente tra source e drain è regolata da una tensione sul gate; un quarto terminale connette a terra il semiconduttore. Anche di questo diamo una descrizione schematica con riferimento alla figura 2 per un transistor nMOS.

Il principio di funzionamento è semplice. Come si vede la placca del gate e il corpo del semiconduttore a cui è opposta, separate da un sottile strato isolante, costituiscono di fatto un condensatore, ed è sulla carica e scarica di questo che si basa il comportamento del dispositivo. Se G è a tensione di terra (condensatore scarico) la presenza di due giunzioni opposte *n-p* e *p-n* impedisce il passaggio di corrente tra S e D. Se G è a tensione positiva rispetto a terra e superiore alla barriera delle giunzioni (condensatore carico), questa respinge le lacune della zona *p* e attira elettroni dalle zone *n*, che convergono in un sottile strato di semiconduttore adiacente all'isolante detto *canale*: per un opportuno valore di tensione di G questi elettroni sono sufficienti a costituire un percorso continuo con gli elettroni liberi delle due zone *n* del transistor, e S e D sono messi in contatto. In sostanza la tensione su G rende isolante o conduttore il percorso S – D: una situazione interpretabile in modo puramente binario nei circuiti digitali.

Il funzionamento di un transistor pMOS è in certo senso opposto. Un'opportuna tensione negativa su G attira lacune nel canale ponendo in contatto S con D, mentre la tensione di terra su G blocca la conduzione.

Il dispositivo immediatamente costruibile attorno al transistor nMOS (o similmente pMOS) è l'*invertitore* schematizzato nella figura 3, che realizza l'operazione logica di NOT tra l'ingresso I su G e l'uscita U su S. Rappresentando i valori 0 – 1 di un bit come due tensioni V_0 – V_1 (tipicamente tensione di terra per

V_0 e positiva di qualche decimo di Volt per V_1) l'operazione NOT implica che per I = 0 (cioè G = V_0) sia U=1 (ciè S= V_1) e viceversa. A tale scopo si collegano S a V_1 attraverso un resistore R, e D a V_0. Ponendo G = V_0 il transistor non conduce, il circuito S – D è aperto e si ha S = V_1 poiché non vi è corrente nel circuito e quindi caduta di tensione su R. Ponendo G = V_1 il circuito S – D è chiuso e S è praticamente collegato a terra perché la resistenza del canale è trascurabile rispetto a R: si ha quindi S = V_0.

Come spieghiamo sotto questo non è esattamente il circuito impiegato oggi, ma la semplice realizzazione del NOT ha una conseguenza importantissima:

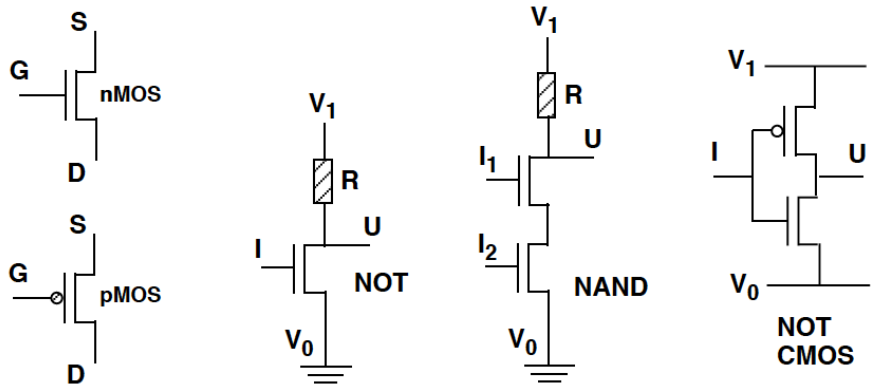

Figura 3
Da sinistra. Simboli dei tansistor MOS. Il circuito invertitore. Un NAND a due ingressi. Il circuito invertitore CMOS

ponendo due (o più) transistor in serie si realizza l'operazione logica di NAND il cui valore è 0 se e solo se tutte le variabili d'ingresso hanno valore 1. Nel circuito, anch'esso indicato nella figura 3, l'uscita U è rilevata sul contatto S del primo transistor e gli ingressi I1, I2 per ogni variabile sono applicati ai gate G dei differenti transistor: è sufficiente che il gate G di uno dei transistor della catena sia interrotto (G = V0) perché non transisti corrente e si abbia S = V1; se tutti i G sono a V1 la catena conduce e si ha S = V0. Se i transistor sono invece posti in parallelo si realizza l'operazione NOR il cui valore è 1 se e solo se tutte le variabili d'ingresso hanno valore 0. Come è noto dalla teoria della commutazione il solo operatore NAND o NOR è sufficiente per esprimere qualunque funzione booleana e quindi costruire qualunque rete di calcolo[2]. Inoltre collegando tra loro ad anello due transistor si realizza un flip-flop, cioè un circuito stabile in due condizioni (se un transistor conduce l'altro è bloccato, e viceversa) che consente di memorizzare un bit. Collegando transistor si può quindi realizzare un intero microprocessore, il circuito più avanzato ospitato da un chip.

La presenza del resistore R nel circuito invertitore o NAND sopra descritto causa ovviamente un consumo di energia quando i transistor sono in conduzione.

[2] Gli operatori NAND e NOR furono introdotti come "operatori universali" dal matematico Henry Maurice Sheffer nel 1913. Il loro enorme interesse pratico si è rivelato con la nascita dei transistor.

Questo problema, che è divenuto sempre più grave con la moltiplicazione del numero di transistor, si rivelò serio sin dall'inizio anche per la dissimmetria di consumo tra un'operazione e la sua inversa. La soluzione denominata CMOS (C indica complementary), ideata nel 1963 da Frank Wanlass e Chi-Tang Sah e commercializzata dalla RSA a partire dal 1968, è ormai adottata come standard. L'invertitore CMOS, anch'esso indicato nella figura 3, è composto da due transistor connessi in serie: un nMOS detto pull-down e un pMOS detto pull-up. L'ingresso I è applicato a entrambi i gate. L'uno o l'altro transistor conduce se la tensione sul gate è rispettivamente maggiore (G = V1) o minore (G = V0) della tensione di soglia connettendo l'uscita U rispettivamente a V0 o V1 senza l'impiego di resistori[3].

L'evoluzione dei transistor CMOS che avrebbe condotto alla struttura impiegata oggi fu imposta principalmente da problemi elettrici che si stabiliscono all'interno del dispositivo con la riduzione delle dimensioni minime (tecnicamente detta feature) di ogni sua parte. Con feature di 350 nm raggiunta nel mezzo degli anni '90 cominciarono a manifestarsi fenomeni non immediatamente spiegabili con l'elettrotecnica classica, soprattutto correnti parassite che superavano gli strati isolanti; queste causavano dissipazione di energia e un imprevisto invecchiamento dei materiali con conseguente degradazione delle prestazioni. La persona chiave in questa evoluzione fu Chenming Hu, professore dell'Università di California a Berkeley e inventore del transistor detto FinFET (fin, cioè pinna) che oggi domina le applicazioni.

Come si intuisce dalla figura 2 il transistor MOS è costituito da strati sovrapposti ed è detto planare, in contrapposizione al FinFET di figura 4 che si sviluppa in tre dimensioni con le zone di S-D e G poste verticalmente. La "pinna" si riferisce al sottilissimo strato di silicio drogato che contiene il canale, contornato dal gate da entrambe le parti con maggiore efficienza di funzionamento rispetto

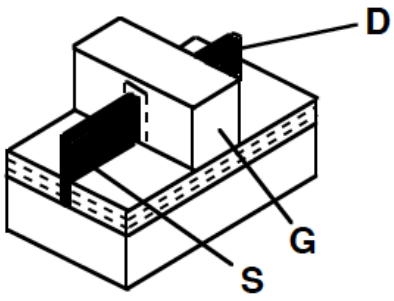

Figura 4
Il transistor FinFET. La zona a puntini è isolante, la pinna è in nero.

[3] La "logica di tensioni" in cui i valori 0, 1 sono associati alle tensioni di ingresso e uscita dei transistor non deve for credere che nel circuito non circolino correnti. Imponendo una tensione sul gate si carica il condensatore gate/substrato attraverso una corrente. Similmente la tensione di uscita del transistor alimenterà un circuito successivo.

al MOS planare e perdite parassite molto inferiori. Il progetto con i primi prototipi fu presentato da Hu nel 2000, fu a lungo discusso per la difficoltà di realizzazione e fu posto in commercio per la prima volta dalla Intel nel 2011 con una feature ormai ridotta a 25 nm. Oggi in diverse versioni è lo standard: le tensioni, i consumi e il tempo di commutazione si sono drasticamente ridotti e la produzione di FinFET ha dato inizio a un possibile sviluppo di chip in tre dimensioni di cui parleremo in seguito.

Vogliamo infine parlare di un altro tipo di transistor oggi molto diffuso, che mostra come la tecnologia abbia raggiunto limiti consentiti solo dallo studio della meccanica quantistica. Ci riferiamo ai floating-gate MOS (FGMOS) che costituiscono tra l'altro le memorie flash, economicissimo mezzo di memorizzazione di dati. Precedute da altri dispositivi basati su simili principi, le memorie flash costruite oggi sono nate dagli studi compiuti negli anni '80 nei laboratori Toshiba da Fujio Masuoka, e sono divenute di largo impiego una decina d'anni più tardi.

Nel FGMOS ogni cella contiene un ulteriore floating gate FG completamente isolato dal resto, posto all'interno dell'isolante tra gate G e canale e separato da questo da un sottilissimo strato di ossido. Una (alta) tensione di alcuni volt applicata a G causa un intenso flusso di elettroni nel canale, alcuni dei quali superano il sottile strato isolante trasferendosi in FG dove possono rimanere per molto tempo: questa migrazione avviene secondo diversi fenomeni quantistici a seconda del tipo di transistor, ed è interpretata come la memorizzazione di un bit 0 mentre l'assenza di carica in FG rappresenta il bit 1. Gli elettroni eventualmente presenti in FG modificano la tensione di soglia del transistor che ora reagisce alla tensione applicata su G commutando o meno a seconda del valore di tale soglia: la corrispondente presenza o assenza di corrente tra S e D è utilizzata come lettura del bit memorizzato.

Le memorie flash furono impiegate inizialmente per memorizzare dati che non richiedevano accessi troppo frequenti che avrebbero potuto danneggiarle. Oggi hanno raggiunto un tale livello di qualità da essere utilizzabili anche nei personal computer ove sostituiscono i dischi consentendo ridottissimi tempi di accesso rispetto a questi.

3. I circuiti integrati e la legge di Moore

Dieci anni dopo la nascita del transistor cominciò a farsi strada l'idea che i semiconduttori avrebbero acquisito un ruolo fondamentale nel mercato se fosse stato possibile costruire su essi interi circuiti contenenti componenti passivi e transistor, riducendone le dimensioni fino a miniaturizzarli. Come racconta un bell'articolo di Silvio Hénin uscito su questa rivista [2] due proposte di brevetto furono depositate nel 1959 da Jack Kilby della Texas Instruments e successivamente da Robert Noyce della Fairchild sulla realizzazione di *circuiti integrati* (IC), nome derivato dalla stessa proposta di Kilby circa "un nuovo circuito elettronico miniaturizzato fabbricato su un corpo semiconduttore in cui tutti i componenti del circuito sono completamente integrati" (comunemente si usa oggi

il termine *chip* anche se questo dovrebbe riferirsi alla sola piastrina su cui è realizzato). Ne seguì una lunga competizione legale sulla priorità della proposta ma nel 1966 le due società si accordarono per mettere in comune i due brevetti e iniziò lo sviluppo inarrestabile di questa tecnologia. Un momento simbolico fu la nascita dello Intel 4004, primo processor completamente integrato su un chip con il contributo essenziale dell'italiano Federico Faggin.

Un anno prima dell'accordo tra Texas Instruments e Fairchild, Gordon Moore cofondatore della Fairchild pubblicava un breve articolo su una rivista sul commercio elettronico. Spiegato il ruolo preminente che avrebbero assunto in diversi campi i circuiti elettronici, esprimeva una previsione sul loro sviluppo nei circuiti integrati che avrebbe acquisito nel tempo una grande popolarità sotto il nome di *Legge di Moore* [3].[4] Le sue parole, che citiamo testualmente, affermavano che i circuiti elettronici "will be regulated by a yearly doubling in the number of components that can be economically packed in an integrated circuit". Accompagnavano la figura 5, tratta dal suo articolo, che indica la curva del costo per componente in funzione del numero di componenti come rilevata nel 1962 e 1965 e prevista per il 1970, simile per tutti i metodi di produzione allora sperimentati.

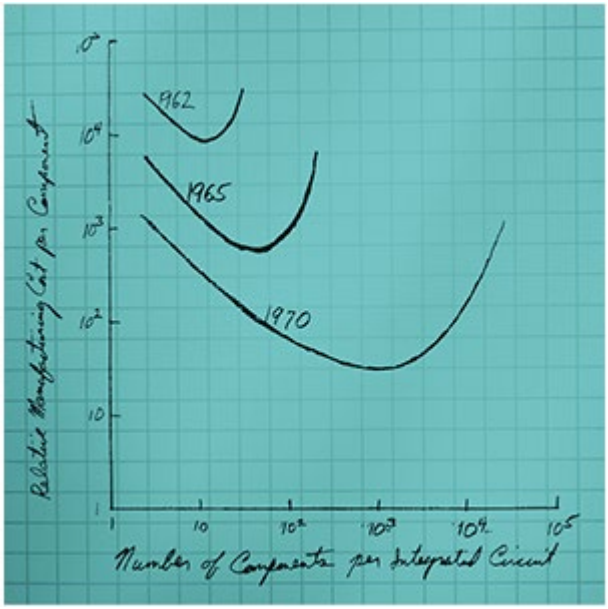

Figura 5
La legge di Moore, nell'immagine tratta dal suo articolo [3].

Abbiamo sottolineato noi la parola "economically" perché su questo punto era concentrata l'affermazione: il famoso raddoppiamento annuale del numero di componenti nel chip con cui la legge viene espressa popolarmente è molto

[4] Nel 1968 Moore lasciò la Fairchild per fondare la Intel assieme a Noyce. È mancato nel marzo 2023 a novantaquattro anni, ma fino all'ultimo ha rilasciato interessantissime interviste.

riduttivo rispetto al concetto espresso da Moore. Ogni tecnologia costruttiva ha una curva di costo che fino a un certo punto diminuisce se si impaccano più componenti in un circuito integrato, ma oltre un certo valore il costo per componente cresce perché il *yield* di produzione, cioè il numero di esemplari funzionanti, diminuisce. La curva con il suo punto di minimo costo per componente si muoverà nel tempo permettendo di aumentare sempre più il livello di integrazione dei circuiti.

La legge di Moore, anche nella sua accezione semplificata, ha continuato a valere per anni con un'ovvia rimodulazione del tempo necessario al raddoppio dei componenti. Sostanzialmente la natura esponenziale della crescita è stata rispettata fino al 2018, anno in cui sembrava arrestarsi per la difficoltà di spingere la miniaturizzazione oltre certi livelli. L'entrata in funzione di apparati innovativi l'ha rimessa in marcia fino al limite odierno, ma da ora solo un drastico cambiamento nella struttura dei chip potrà conservarla. Per parlare di tutto questo occorre prima descriverne brevemente la tecnologia costruttiva dei chip.

4. Evoluzione della produzione: problemi emersi e proiezione nel futuro

La produzione di circuiti integrati è ovviamente molto complessa e ha raggiunto uno stato di altissima perfezione tecnologica con il diminuire delle dimensioni dei componenti (feature, come già detto). L'elemento di base è un disco (wafer) tagliato da una barra cilindrica di silicio cristallino, su cui vengono costruiti tanti chip tutti uguali che saranno divisi tra loro a fine procedimento. I wafer hanno oggi uno spessore di circa 0.5 mm e un diametro fino a 30 cm. La costruzione di un chip passa attraverso la realizzazione di strati successivi di drogaggio e deposizione di materiali sulla superficie secondo forme comandate da una serie di maschere, cioè disegni ridotti in scala microscopica e sovrapposti al chip per definire le zone da modificare su ogni strato. Ogni maschera è ripetuta in tante copie adiacenti per generare tanti chip uguali su ogni wafer.

Accenniamo solo che una maschera ha zone opache o trasparenti. La superficie del semiconduttore, inizialmente scaldata per ricoprirla di ossido isolante, è cosparsa di resina, la maschera viene a questa sovrapposta e una radiazione attraversa le zone trasparenti causando la polimerizzazione locale della resina che si solidifica in quelle zone e le protegge da successivi attacchi. Tolta la maschera si asporta chimicamente la resina non polimerizzata e se necessario anche l'ossido sottostante lasciando esposte le zone che richiedono un trattamento in quello strato del chip come drogaggio o deposizione di materiali conduttori. Si procede così strato per strato isolandoli tra loro e realizzando le connessioni tra strati consecutivi attraverso vias, cioè fori nell'isolante riempiti di tungsteno. Le zone drogate sono nei primi strati e quelli superiori contengono le connessioni metalliche che tipicamente possono invadere fino a venti strati.

Questo procedimento detto fotolitografico, realizzato all'inizio con la luce visibile, si era trasferito col tempo nel campo degli ultravioletti: una transizione indispensabile per la riduzione della feature perché con una radiazione luminosa

non si possono definire dettagli inferiori alla sua lunghezza d'onda. Nel 2018, anno chiave per questa discussione, la minima lunghezza d'onda impiegata era di 193 nm e la feature dei chip più avanzati era attorno ai 10 nm. Questa apparente contraddizione era dovuta a un procedimento lungo e costoso detto multiple patterning consistente nel definire le forme su uno strato del cip attraverso successivi passaggi con maschere diverse spostate un rispetto all'altra di pochi nanometri. Il procedimento non poteva comunque essere ripetuto per troppe passate per motivi di costo e di precisione, ed era chiaro da almeno vent'anni che per ottenere feature più piccole si sarebbe dovuto ricorrere all'impiego di radiazioni di lunghezza d'onda decisamente inferiore. Le principali compagnie di produzione erano impegnate in questo affrontando gravi difficoltà tecnologiche, in particolare l'impossibilità di ridurre e concentrare le immagini delle maschere sul wafer mediante lenti perché, al crescere della frequenza, l'intera energia della radiazione sarebbe stata assorbita da esse. Nel 2018 la ASML, una società olandese che aveva perfezionato per anni il progetto, mise per prima in commercio un'apparecchiatura che impiegava radiazione EUV (Extreme Ultra Violet) a 13.5 nm in grado di produrre chip a 7 nm. In particolare la concentrazione della radiazione era ottenuta con un sistema di specchi speciali in cascata prodotto dalla Zeiss: in questo modo ben il 2% dell'energia prodotta raggiungeva il chip! È doveroso aggiungere che l'apparato della ASML assorbiva circa 1.5 MW di potenza e come in tutte le catene di produzione dei chip doveva essere tenuto in funzione giorno e notte senza interruzione.

Il 2018 segna l'inizio di una nuova era. La tecnologia di oggi impiega radiazione EUV e raggiunge una feature minima di 4 nm, sono già in cantiere produzioni a 3 nm e vi sono nuove fabbriche in allestimento che l'anno prossimo inizieranno la produzione a 2 nm. Come detto il raggio di Van der Vaals nel cristallo di silicio è 0.21 nm: la pinna del FinFET conterrà trasversalmente solo qualche atomo, posto che questa affermazione abbia un corretto senso fisico[5]. Ma apparentemente i progressi non si fermeranno qui.

La prima considerazione da fare a questo riguardo è che lo sviluppo dei chip è legato a problemi di progettazione in parte indipendenti dalle tecnologie di produzione. Il più importante è quello del consumo di energia a cui si lega il riscaldamento del circuito che può comprometterne il funzionamento, principalmente dovuto alle connessioni elettriche che come già detto sono ospitate negli strati superiori. Naturalmente al crescere del numero di transistor cresce il numero di connessioni che nelle tecnologie più spinte hanno una lunghezza complessiva valutabile in molte decine di km per cm2. Queste connessioni sono realizzate in alluminio più facile da depositare, ma al diminuire della feature l'alluminio è sostituito in genere dal rame che ha una conduttività maggiore e produce meno dispersione termica: e qui stanno nascendo problemi inaspettati a un non esperto.

[5] Ricordiamo che la feature si riferisce ai dettagli di minima grandezza. Secondo molti esperti un parametro più significativo per il confronto tra processi produttivi è la lunghezza del canale che in una tecnologia a 2 nm dovrebbe attestarsi attorno a 7nm.

Un conduttore di rame è soggetto al fenomeno quantistico della elettromigrazione sconosciuto all'elettrotecnica classica, secondo il quale gli elettroni che lo percorrono possono dislocare per urto atomi del metallo, spinti nell'isolante esterno a esso: correnti intense possono causare l'interruzione della conduzione in conduttori molto sottili che sono quindi protetti con "bordature" di altri materiali aumentandone però l'ingombro. Per questo scopo si sta sperimentando il grafene che può essere depositato in strati estremamente sottili attorno al rame, ma è ormai riconosciuto che per diminuire lo spessore dei collegamenti il rame dovrà essere sostituito: da tungsteno o cobalto per quanto indicato oggi, ma il cambiamento comporterà un complesso cambiamento del procedimento di produzione dei chip e per ora le realizzazioni si limitano a prototipi.

Un altro fenomeno tipicamente ingegneristico che forse stupirà i "non addetti ai lavori" interessa il progetto complessivo dei chip: se ne parla ormai da qualche anno in tutti i convegni e riunioni specializzate ma finora quasi nulla è stato fatto. Nelle operazioni di un chip, in particolare se contiene un intero microprocessore, l'energia e il tempo necessari per trasferire i dati dalla memoria ai circuiti di calcolo (di fatto ai transistor) sono molto superiori all'energia e al tempo richiesti da questi ultimi per eseguire un'operazione su di essi. I dati pubblicati in [4], relativi a un microprocessore di riferimento, indicano che spostare di 1 mm in parallelo due parole di 32 bit richiede 1.9 pJ e 400 ps mentre un'addizione tra di esse richiede 20 fJ e 150 ps (ricordiamo che p (pico) indica 10-12 e f (femto) indica 10-15). In particolare l'enorme differenza tra i due valori dell'energia è dovuta all'effetto Joule nei conduttori. La soluzione, prospettata da tutti e per ora sperimentata in modo parziale principalmente dalla Samsung, viene indicata come in-memory computing e consiste nell'avvicinare tra loro sul chip memoria e CPU, operazione che richiede alcune trasformazioni nella tecnica di produzione.

Per quanto riguarda la legge di Moore siamo ormai al limite delle possibili riduzioni di feature. La soluzione, anche qui prospettata da tutti e sperimentata a oggi solo in modo limitatissimo, consiste nell'invadere la terza dimensione del chip che per ora ha una organizzazione sostanzialmente planare. La prima innovazione abbastanza semplice ha riguardato le memorie realizzate in più strati sovrapposti. Molto più interessante è la costruzione già sperimentata di CMOS i cui transistor pull-up e pull-down sono costruiti uno sopra l'altro in due zone sovrapposte del chip: struttura ancora in fase sperimentale che quasi dimezza l'area del chip ma complica le connessioni.

Inutile dire che le tecnologie di produzione dovranno essere trasformate in modo sostanziale. L'approccio più desiderabile su cui si scommette, limitato a studi di fattibilità ma non ancora a progetti esecutivi, è connesso al problema dello in-memory computing: uno strato di celle di memoria costruito immediatamente sopra uno strato di calcolo ridurrebbe drasticamente la lunghezza dei collegamenti tra di essi, riducendo tra l'altro l'energia termica generata che non è semplice dissipare dall'interno di un chip tridimensionale. Le previsioni di oggi

indicano l'anno 2025 come inizio dell'era in cui queste soluzioni entreranno commercialmente in gioco. Uno sviluppo tridimensionale con valori delle tre dimensioni confrontabili tra loro sembra ancora un obiettivo molto lontano, anche se il mondo dei chip non ha mancato di stupirci negli anni.

5. Qualche considerazione conclusiva

Concludiamo questa carrellata con qualche informazione sul mondo della produzione. Nella figura 6 è mostrato il chip commerciale su cui è integrato un intero processor, contenente il massimo numero di transistor tra quelli disponibili nel momento in cui questo articolo viene scritto: si chiama M1 Ultra e vi convivono 114 miliardi di transistor. Si noti che la struttura di un intero processor compresi i circuiti di input/output è incomparabilmente più complessa di quella altamente ripetitiva di una memoria che può essere integrata con un numero di transistor anche maggiore.

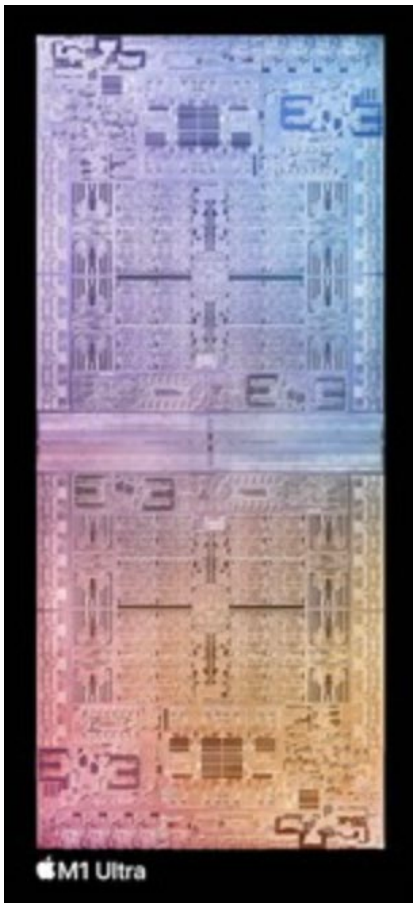

Figura 6
Il chip M1 Ultra della Apple, composto da due M1 Max integrati insieme, contiene 114 miliardi di transistor [6].

M1 Ultra è costruito su progetto della Apple dalla Taiwan Semiconductor Manufacturing Company (TSMC), la più grande del mondo nella produzione di chip, ma mettere a confronto le diverse compagnie di produzione non è cosa ovvia. Le più grandi sono TSMC, Samsung, Intel e IBM con differenze tra gli investimenti: nella produzione primeggiano le prime due e in ricerca e sviluppo le seconde. Ma anche Global Foundries, NTT, Toshiba e varie altre partecipano attivamente al gioco tecnologico e scientifico, e tra queste è doveroso citare la italo-francese STMicroelectronics (STM) nata dalla fusione tra l'italiana SGS e la francese Thomson. In particolare la SGS, da cui proveniva Federico Faggin prima di trasferirsi negli Stati Uniti, è stata insignita nel maggio 2021 del IEEE Milestone per "Multiple Silicon Technology on a Chip, 1985"[6]. Non eccellono ancora le compagnie cinesi, pur producendo chip in grandi quantità in particolare per i telefoni cellulari, perché attriti politici non gli consentono di ottenere il software di progetto e le apparecchiature di produzione più avanzate prodotti in USA, Europa e Giappone.

Interessanti tabelle di dati tecnico-economici sui transistor prodotti oggi si trovano sulle riviste del settore, in particolare citiamo l'articolo [5]; in atti di congressi specializzati; e sui documenti di produttori di chip e di agenzie di valutazione. Tra queste meritano una menzione particolare le agenzie di marketing IC Insight e TechInsight, e sopra tutte la International Roadmap for Devices and Systems (IRDS) cui partecipano esperti internazionali sotto il patrocinio dell'IEEE con il proposito di stimolare l'innovazione e la ricerca [7].

Concludiamo con due dati crudi. Il primo, pubblicato da IC Insight, mostra che nel 2022 le spese complessive nel mondo per ricerca e sviluppo di chip hanno superato 850 miliardi di dollari e la previsione per il 2025 sfiora i 100 miliardi, mentre il rapporto tra queste spese e i ricavi della vendita dei chip si è assestato attorno al 13 % nel 2022 ed è previsto stabile nei prossimi anni. Il secondo, pubblicato da TechInsigth, è che nel 2022 sono stati prodotti circa $2x10^{21}$ (due miliardi di trilioni) di transistor: circa 250 miliardi per ogni abitante del nostro pianeta.

E, a titolo di cronica spassosa, si può stimare che l'area complessiva dei circuiti integrati esistenti oggi si avvicini ai cento milioni di metri quadrati.

[6] Il Milestone è conferito dall'istituto americano IEEE in ricordo di momenti fondamentali nello sviluppo scientifico e tecnologico. La dedica a SGS recita: "SGS (now STMicroelectronics) pioneered the super-integrated silicon-gate process combining Bipolar, CMOS, and DMOS (BCD) transistors in single chips for complex, power-demanding applications." I transistor DMOS (D sta per double-diffused per la loro struttura sul silicio) sono transistor di potenza di cui, come per altri tipi, non vi è spazio per parlare qui.

BIBLIOGRAFIA

[1] Perry, T.S. (2020). "How the father of FinFETs helped save Moore's law". *Spectrum IEEE*, May 2022, 47-51.

[2] Hénin, S. (2019). "Due anniversari da ricordare", *Mondo Digitale,* No 86, Novembre 2019, rubriche *Storia dell'Informatica.*

[3] Moore, G.E. (1965). "Cramming More Components onto Integrated Circuits", *Electronics Magazine,* 38 (8), 114–117.

[4] Dally, W. e Vishkin, U. (2022). "On the model of computation", *Communications of the ACM*, 65 (9), 30-32.

[5] Cass, S. (2022). "The ultimate transistor timeline", *Spectrum IEEE*, December 2022, 29-31.

[6] https://www.apple.com/it/newsroom/2022/03/apple-unveils-m1-ultra-the-worlds-most-powerful-chip-for-a-personal-computer/

[7] https://www.icinsights.com, https://www.techinsights.com, https://irds.ieee.org/editions/2022

BIOGRAFIA

Fabrizio Luccio. Laureato in Ingegneria Elettronica al Politecnico di Milano ha lavorato inizialmente all'Olivetti, è stato ricercatore presso l'MIT, l'IBM in USA e la NTT in Giappone, professore delle University of Southern California, New York University, University of Illinois e National University of Singapore. Tornato in Italia ha insegnato fino al 2010 nell'Università di Pisa ove è oggi professore emerito. È Life Fellow dell'IEEE e professore onorario della Technical University di Xi'an e della University of Nanning in Cina. Ha diretto un'intensa cooperazione scientifica con paesi in via di sviluppo per conto dell'Unesco.

E-mail: fabrizio.luccio@unipi.it

Riscrivere Marx per la società dell'informazione

Stefano Diana

Sommario

L'opera combinata di Turing e Shannon ha dato forma alla odierna "società dell'informazione", fondandola su una nuova metafisica alla quale siamo ormai talmente abituati da vederla come qualcosa di ovvio e naturale. Ma celatamente l'ha impostata in modo da essere orientata alle macchine invece che agli umani, sin dal principio. Per illuminare le aberrazioni e i pericoli di questa cultura, e cosa comporti per noi esseri umani vivere in un habitat machine-biased, nel presente lavoro mi faccio aiutare dalla lucidità di Karl Marx. Come dimostro con esempi da varie opere, certi topoi della critica di Marx all'economia politica del suo tempo possono essere riscritti sostituendo poche parole chiave per produrre un'analisi sorprendentemente pertinente della società dell'informazione e del ruolo che i dati e l'Intelligenza Artificiale svolgono in essa. La spiegazione di questo strano fenomeno nasce dal fatto che l'informazione può esser vista come l'upgrade attuale del denaro marxiano, astrazione disumanizzante e "feticcio automatico" che si espande da sé indipendentemente dai bisogni umani.

Abstract

The combined work of Turing and Shannon shaped today's "information society", and founded it on a new metaphysics to which we've become so accustomed that we see it as something obvious and natural. But covertly it was set up to be machine-oriented instead of human-oriented, from the very beginning. To highlight the aberrations and hazards of this culture, and what it entails for us humans to live in a machine-biased habitat, in this paper I take help from the lucidity of Karl Marx. As I show with examples from various works, certain topoi of Marx's critique of the political economy of his time can be rewritten by substituting a few keywords, producing a surprisingly relevant analysis of the information society and the role that data and Artificial Intelligence play in it. The explanation for this strange phenomenon stems from the fact that information can be seen as the current upgrade of marxian money, a dehumanizing abstraction and "automatic fetish" that expands itself independently of human needs.

Keywords: data, money, information society, artificial intelligence, knowledge, automatic fetish, formalism, machine-biased culture, empathy, Marx, Turing, Shannon

1. Introduzione: Turing ∩ Shannon

Si è spesso parlato dell'intreccio tra il lavoro di Alan Turing e quello di Claude Shannon, ma c'è un aspetto di questo rapporto che mi pare poco frequentato.

Turing ha indagato sui limiti della calcolabilità partendo dalle operazioni di un calcolatore umano: cosa fa una persona quando calcola. A partire da questo primo seme, il confronto tra umani e macchine diventa per lui un tarlo costante. Ritorna più volte sul tema fino al culmine dell'*imitation game*, alias "test di Turing" [1]. L'aspetto notevole per me è che il setting del test richiede che le parti in gioco comunichino fra loro a) *in assenza di corpi*, e b) *esclusivamente tramite codici* (testi).

Dal canto suo, Shannon fondendo la logica con i circuiti elettrici ha trovato la via per migliorare la trasmissione di messaggi tra persone distanti tramite un canale fisico che presenta inevitabile rumore. Anche in questo caso si ritrovano le stesse condizioni di Turing: comunicazione a) *in assenza di corpi* e b) *esclusivamente tramite codici* (testi).

La comunicazione a distanza cominciava a montare poderosamente. «La materia prima era dappertutto», ricorda James Gleick [3], «rutilante e ronzante nel panorama del primo ventesimo secolo: lettere e messaggi, suoni e immagini, notizie e istruzioni, fatti e cifre, segnali e segni: una gran macedonia di specie fra loro in relazione. Tutto era in movimento, attraverso il sistema postale, cavi o onde elettromagnetiche. Ma non c'era una parola sola che denotasse tutte quelle cose».

Quando nacque l'esigenza tecnica di dare una definizione formale e generalizzata di fenomeni così variegati, inizialmente si usò la parola "intelligence". Nel 1926 Ralph Hartley in uno storico articolo [2] adottò per primo il termine "information", poi temprato definitivamente da Shannon. Da ingegneri essi avevano bisogno di portare questo vago termine dal campo aperto della comunicazione umana al dominio della misurabilità e calcolabilità, così da convertire l'informazione in grandezza fisica. Qualcosa di simile a ciò che aveva fatto Newton con la forza, il movimento, il tempo. Per farlo la prima cosa era, nelle parole di Hartley, «eliminare i fattori psicologici coinvolti».

Hartley e Shannon partono dalla necessità umana di comunicare, ma finiscono per rappresentare la comunicazione fra umani a immagine e somiglianza di quella fra macchine. I simboli scambiati indicano stati e oggetti discreti senza ambiguità. Le relazioni tra i simboli sono strutturate con precisione.

La "macchina" di Turing e il "messaggio" di Shannon sono genuine entità matematiche. La *computazione* di Turing opera su simboli astratti con regole formali [4]; l'*informazione* di Shannon ha un valore che dipende solo dalla distribuzione di probabilità dei suoi elementi atomici, dei suoi possibili stati [5]. In quanto entità matematiche, la macchina di Turing e il messaggio di Shannon

godono della massima generalità e astrazione, per questo sono entrambe indifferenti al contenuto ed estranee a qualunque idea di significato[1].

Non è superfluo ricordare che questa riconversione dell'informazione avveniva in un contesto bellico. La Seconda guerra mondiale, al cui servizio lavorano sia Turing che Shannon, definisce i loro obiettivi: la *vitale* decifrazione di messaggi nemici per il primo, la *vitale* correttezza nella trasmissione di messaggi amici per il secondo. Credo che anche questo campo di forze, oltre alla forma mentis matematica, spieghi come mai Turing e Shannon concepiscono la comunicazione umana in modo tanto ristretto e distorto. Solo per questioni di vita o di morte la comunicazione si irrigidisce e non tollera più l'errore, che nel fluido traffico linguistico quotidiano è una sostanza del tutto diversa: sfuggente, emendabile, evocativa, creatrice.

Tuttavia la guerra finì, e la loro interpretazione rimase. Anzi, la loro sfera di influenza si è estesa ovunque. C'è un motivo pratico: l'efficienza tecnica, l'utilità delle infinite applicazioni del calcolo automatico e della trasmissione digitale con compressione e correzione di errore. Ma questa efficienza non è neutra e innocente, presuppone una visione del mondo. Ed è su questa visione che Turing e Shannon si sono innestati perfettamente come un poderoso amplificatore, producendo una rivoluzione epistemologica e antropologica.

Dato il sogno di Turing di «costruire un cervello» e le sue riflessioni in merito in [25], il suo approccio alla computabilità è venuto ad assumere un ruolo essenziale nella concezione stessa di "intelligenza". Sennonché l'astuzia sperimentale dell'*imitation game* – comunicare solo tramite testi scritti a macchina – ha impresso una forma particolare a questo corso. Ha diffuso la credenza che l'intelligenza, e con lei le altrettante vaghe mirabilia dell'anima su cui si affabula dall'antichità, come il pensiero e la coscienza, possano essere ben definite come *funzioni*, e che siano facoltà *astratte* che possono sussistere ed essere riscontrate anche *senza la presenza fisica*, in assenza di un corpo vivente che le manifesta nei suoi comportamenti a un altro corpo vivente. Idee del genere non incontravano alcuna resistenza, al contrario andavano a coronare la tradizione del dualismo cartesiano con una dignità pseudo-scientifica aggiornata, di qualità superiore[2].

E Shannon? Partendo dall'incertezza riguardo a un fenomeno che può avere un certo numero di esiti diversi, l'"informazione" di Shannon si può pensare come la riduzione di tale incertezza che un evento ci porta. In termini matematici questa incertezza fu definita da Shannon come "entropia", reinventata sul calco della fisica come stima empirica di una distribuzione di probabilità. Con una differenza

[1] Shannon lo dichiara in apertura, ibid.: «Spesso i messaggi hanno significati; vale a dire che si riferiscono a o sono correlati mediante qualche sistema con certe entità fisiche o concettuali. Questi aspetti semantici della comunicazione sono irrilevanti per il problema ingegneristico» (traduzione e corsivo miei).

[2] Turing scrive esplicitamente in [1]: «Il nuovo problema ha il vantaggio di tirare una linea di separazione abbastanza netta tra le capacità fisiche e quelle intellettuali di un uomo» (pag. 118).

essenziale: è una grandezza *adimensionale*, «semplice diversità combinatoria, compatibile con l'entropia di Boltzmann-Gibbs sotto certe condizioni». [26]

La sua purezza matematica, che descrive sia l'entropia di una sorgente di eventi, sia l'informazione associata a uno dei suoi eventi che si avvera, è compatibile con la rappresentazione astratta di qualunque fenomeno. Ma attenzione: qui si parla di *conoscenza* che riduce l'incertezza sul mondo, la facoltà umana fondamentale nella nostra tradizione culturale, intimamente legata all'intelligenza e al pensiero. Mentre questi venivano resi incorporei dalla stilizzazione di Turing, la conoscenza si rendeva altrettanto disponibile a farsi astratta per mano di Hartley e poi di Shannon. Dunque l'informazione di Shannon è diventata la sostanza metafisica della conoscenza. E poi dell'intelligenza e del pensiero che la gestiscono. E poi, per analogia con la trasmissione intelligente, degli scambi tra sistemi di qualunque tipo (fisici, chimici, biologici, ecc.). Talvolta persino sostanza *fisica*, reificata[3].

Turing e Shannon introducono nel campo cognitivo umano delle semplificazioni di stampo matematico simili a quelle della logica di Aristotele. A dispetto della sua scarsa verosimiglianza come modello del pensiero umano, la logica ha goduto di enorme successo culturale proprio grazie ai vantaggi immediati offerti dalla sua inebriante capacità di «difalcare gl'impedimenti» [27] e dare l'illusione del controllo. Affine e non meno impressionante è il trionfo del modello informazionale in ogni campo, e l'inflazione di illusioni epistemiche che da esso sono nate, soprattutto ma non solo nel territorio delle questioni umane[4].

2. La società machine-biased

Senza farsi troppo notare, il setting del test di Turing ha impostato una relazione sbilanciata tra gli umani e le macchine informatiche fin dal principio. Una relazione che possiamo dire *machine-biased*, psicologicamente orientata a favorire la macchina, in quanto fatta su misura per i requisiti operativi della macchina e non per le condizioni umane.

Questo elusivo slittamento di contesto mi fa pensare al film "*La vita è bella*". Il protagonista compie il miracolo di fare credere al figlioletto che il lager sia un parco giochi, ma ci riesce solo perché il lager che fa da sfondo alla storia *è già una versione per bambini*, depurata da ogni violenza. Sarebbe stato tanto impossibile mantenere l'apparenza della favola nel terrore di un lager reale, quanto lo sarebbe per una macchina farsi passare per umana *vis-à-vis*.

---

[3] «Da tempo è stato riconosciuto che la sostanza essenziale trasmessa dai neuroni non è la carica elettrica o le sostanze neurochimiche, ma l'informazione. Nell'analisi di un sistema neurale, è essenziale misurare e seguire il flusso di questa sostanza, proprio come negli studi sul sistema vascolare si vuole misurare il flusso sanguigno». [28]

[4] Heidegger le rilevava così nel 1968: «Il concetto guida della cibernetica, il concetto di informazione, è per giunta sufficientemente vasto da poter un giorno assoggettare alle pretese della cibernetica anche le scienze storiche dello spirito. Ciò riuscirà tanto più facilmente in quanto il rapporto dell'uomo di oggi con la tradizione storica si tramuta visibilmente in un mero bisogno d'informazione». [7]

La più importante conseguenza dell'impostazione *machine-biased* è che tutto il dibattito successivo sulla possibilità di simulare o no il comportamento umano con macchine di Turing e con modelli ML/IA si è svolto in una prospettiva contraffatta, conforme alla macchina. Un discorso deumanizzato dalla radice, eppure accolto senza resistenze, anzi con favore. Come mai? Perché è la naturale continuazione di un'altra inveterata tradizione occidentale: il dualismo platonico-agostiniano secondo cui esiste *una vera conoscenza* raggiungibile solo dall'intelletto/anima immateriale, mentre il corpo materiale e le sue emozioni volgari non sono altro che fardello, oscurità, errore e penitenza.

Su queste basi metafisiche fiorisce la "società dell'informazione", «una società neomanufatturiera in cui l'informazione è sia un materiale grezzo che produciamo e manipoliamo sia il prodotto finito che consumiamo». [6]

Essendo fondata su macchine informatiche, qualunque fenomeno che accade in tale società deve essere reso omogeneo ad esse per essere rappresentato nel sistema, altrimenti è invisibile. Queste condizioni mutano nel profondo non solo la nostra visione del mondo, ma anche le relazioni fra noi. Possiamo farcene un'idea leggendo un grande esperto di complessità computazionale, Scott Aaronson.

> Nella pratica le persone si giudicano reciprocamente coscienti dopo un'interazione molto breve, forse anche di pochi secondi. Ciò suggerisce che possiamo mettere un limite superiore finito – per essere generosi, diciamo 1020 – al numero di bit di informazione che due persone A e B si possono realisticamente scambiare prima che A abbia accumulato prove sufficienti a concludere che B è cosciente. [8]

Parole sconcertanti, perfino *cringe*. Chi mai si chiede se una persona appena incontrata è cosciente o meno? Quando càpita è considerato un sintomo patologico. Normalmente è una *tacita inferenza* che il corpo fa da sé sotto la soglia della coscienza, senza usare parole, concetti, ragionamenti. Il corpo vivente di A vede il corpo vivente di B e lo riconosce come simile, subito e tutto compreso. Uno sguardo, un minimo gesto sono più che sufficienti per stabilire un rispecchiamento significativo e una comunicazione tra A e B. [9]

Ma l'ottica informazionale vede e insegna a vedere tutto in termini di informazione, quindi di bit, compresa l'interazione umana. Pretende di codificare persino questo scambio, indicibile persino per i massimi poeti coi suoi processi fisiologici ancora in gran parte oscuri, con quelle sequenze di simboli privi di significato che sono il massimo comun denominatore tra noi e le macchine informatiche[5].

Nella società dell'informazione la condizione di comprensione reciproca tra gli umani *non può essere sostanzialmente diversa da quella fra le macchine*. Deve

[5] In una nota nella stessa pagina: «Le persone che interagiscono in internet, via email o chat, di norma si giudicano l'un l'altro umani invece che spam-bot dopo aver scambiato un numero molto più piccolo di bit!». Ancora interazioni testuali e in absentia, quelle del test di Turing, in un ambiente relazionale modellato dall'informatica

perciò condensarsi in un *numero* dentro un *meccanismo*: la quantità minima di bit necessaria a far scattare in A una molla che incolla su B l'etichetta 'cosciente'[6]. O meglio, per attivare una routine di programma che assegna il valore a una variabile che rappresenta lo stato di coscienza di B. Una variabile booleana: o sei cosciente o non lo sei, 0/1.

La presa di coscienza, in questi termini, non può essere che una funzione a gradino, l'equivalente matematico della *rivelazione*: un attimo prima c'è il buio, un attimo dopo *ding*!, la luce si accende e si è pienamente coscienti. Niente sfumature. Bando ai vaghi livelli intermedi. Non è un caso che fra le immagini più stereotipiche della fantascienza ci siano robot che "diventano autocoscienti" in un preciso istante[7]. E che la famigerata «singolarità tecnologica[8]» sia il momento in cui «l'universo si sveglia».

3. Enter Marx: denaro vs. dati

Il coinvolgimento di Karl Marx in questa storia prende il via sotto specie di un esperimento di "letteratura potenziale" simile a quello di Raymond Queneau, che nei *Fondamenti della Letteratura secondo David Hilbert* riscrive la pietra miliare *Fondamenti della Geometria* sostituendo le parole "punto", "retta" e "piano" rispettivamente con "parola", "frase" e "paragrafo", per assiomatizzare, tra il serio e il faceto, la teoria letteraria[9].

La mia intenzione sperimentale è riscrivere Marx, sempre sostituendo determinate parole chiave, per capire se se ne possa derivare una critica da par suo della nostra società dell'informazione, invece di quella borghese-industriale.

Il punto di partenza è l'ipotesi seguente:

l'informazione di Shannon è l'analogo odierno del denaro in Marx.

Prendiamo in esame questo brano dei *Lineamenti fondamentali della critica dell'economia politica* (alias *Grundrisse*) in cui Marx indaga sul rapporto tra denaro e merci, molti anni prima che nel *Capitale*.

> Ogni merce particolare, nella misura in cui è un valore di scambio, ha un prezzo, esprime una determinata quantità di denaro solo in una forma incompiuta, giacché essa deve essere anzitutto posta in circolazione per venire realizzata – e che lo sia o meno è un fatto accidentale a causa della sua particolarità. Ma finché essa non esiste come prezzo, ma soltanto nella sua determinatezza naturale, allora è soltanto momento della ricchezza in virtù della sua relazione a un bisogno particolare che essa soddisfa, e in

[6] Si tratta letteralmente di «attaccare a qualcosa un cartellino con un nome», nelle parole di Wittgenstein [10].
[7] «Skynet cominciò a imparare a ritmo esponenziale. Divenne autocosciente alle 2:14 del mattino, ora dell'Atlantico, del 29 agosto.» (dal film *Terminator 2*).
[8] L'invenzione letteraria di Vinge e Kurzweil. Curiosamente il libro di Kurzweil [11] è classificato come non-fiction: una riprova della ampia flessibilità delle nostre categorie.
[9] Interessante notare, in margine, che nel far ciò Queneau accoglieva alla lettera l'esortazione di Hilbert stesso a vedere gli oggetti della sua trattazione non come le omonime entità della geometria euclidea ma come pure astrazioni, senza referenti nel mondo fisico.

> questa relazione esprime: 1) soltanto la ricchezza d'uso; 2) soltanto un lato del tutto particolare di questa ricchezza.
>
> Per contro il denaro, a prescindere dalla sua utilizzabilità come merce di valore è: 1) il prezzo realizzato; 2) ciò che soddisfa qualsiasi bisogno in quanto può essere scambiato con l'oggetto di qualsiasi bisogno, con assoluta indifferenza verso qualsiasi particolarità. La merce possiede questa proprietà soltanto grazie alla mediazione del denaro. Il denaro la possiede direttamente nei confronti di tutte le merci, e quindi nei confronti dell'intero mondo della ricchezza, della ricchezza in quanto tale. Nel denaro la ricchezza generale è non soltanto una forma, ma al contempo il contenuto stesso. Il concetto di ricchezza è per così dire realizzato, individualizzato in un oggetto particolare. [13]

Cosa sta dicendo qui Marx? Che le merci, beni tangibili, esistono perché servono a qualcuno per soddisfare qualche suo bisogno. Le merci possono essere scambiate, e per questo assumono un «valore di scambio». Ma in origine, «nella sua determinatezza naturale», un bene ha un particolare e concreto «valore d'uso» per una persona che ne ha bisogno. Questo valore si realizza nella esperienza d'uso, nel godimento del bene. In tal caso il valore del bene ancora «non esiste come prezzo», cioè non è stato tradotto in una quantità di denaro, non ha ricevuto una forma di denaro. E nemmeno esiste un'idea astratta di "ricchezza". Esistono solo usi personali. Quando una persona produce beni per soddisfare le proprie necessità – quando «la portata della sua produzione è misurata dal suo bisogno» [14] – la "ricchezza" tutt'al più può corrispondere al *benessere* che a costui deriva dalla soddisfazione del bisogno.

Il denaro è l'opposto di questa particolarità. È una rappresentazione astratta dei beni, una specie di unità di misura universale per il valore di scambio di beni diversi tra diverse persone[10]. Trascende i valori d'uso di qualsiasi bene, e perciò qualunque bisogno che i beni possano soddisfare. Pur essendo una merce in sé[11]è la sostanza stessa del valore di scambio, «il prezzo realizzato». In più, dà al concetto di "ricchezza" un significato concreto: *la ricchezza è (una quantità di) denaro.*

Essendo il veicolo di ogni scambio, scrive Marx, «il denaro è il mezzano tra il bisogno e l'oggetto, tra la vita e i mezzi di sussistenza dell'uomo. Ma ciò che media a me la mia vita, media per me anche l'esistenza degli altri uomini.» [12]

È qui che l'analogia tra denaro e informazione prende spessore. *L'informazione, i dati, sono oggi i mezzani tra noi e il mondo. Mediano a noi la nostra stessa vita, e anche l'esistenza degli altri uomini.*

---

[10] In questo senso il mercato è come uno strumento di misura che converte tutto in numeri: i prezzi sui cartellini attaccati alle merci.

[11] Al tempo di Marx si trattava di oro e argento, ma il denaro ha conservato la qualifica di merce scambiabile anche ridotto a valore nominale, dissociato dal materiale prezioso

Il flusso di dati si sostituisce al flusso di denaro. Ma il testo marxiano – ecco la logica peculiare del metodo d'indagine per riscrittura – impone un'attenzione ulteriore: 'denaro' e 'dati' sono legati a diversi scenari e la deissi non funziona più. Altre parole chiave vanno sostituite: quelle della vecchia economia di oggetti materiali devono tradursi in quelle di un'economia di oggetti immateriali, *l'economia della conoscenza* in cui i dati generano valore e costituiscono la circolazione principale. Al posto delle merci, insomma, abbiamo la conoscenza. Ma qui è necessario un approfondimento.

Lo scienziato e filosofo Michael Polanyi analizzò i processi creativi che egli stesso e i suoi colleghi vivevano in prima persona [15] [16]. Ne dedusse che noi «sappiamo più di quanto possiamo esprimere» [17], che la conoscenza primaria è quella «tacita» e «personale» incarnata in un individuo, inscritta in forme biologiche nel suo organismo vivente, accumulata nel suo corpo con le esperienze che fa. Know-how, modelli, credenze che dànno forma alle azioni. Benché questo sapere sia inconscio e ineffabile, l'evoluzione ha plasmato la nostra neurobiologia per consentirci di scambiarla direttamente tramite *rispecchiamento e imitazione diretta tra corpi*, da individuo a individuo [9]: il corpo di A mostra a B come fare una cosa, e il corpo di B prova a riprodurla[12]. Ciò è avvenuto molto prima che il linguaggio emergesse, dato che il linguaggio stesso viene appreso per questa via. Qui scorre l'istruzione nativa da genitori a figli, da maestri ad allievi[13].

Col tempo è fiorita l'abilità umana di catturare il sapere incarnato e oggettivarlo fuori dal corpo in quelle che Bernard Stiegler ha definito "ritenzioni terziarie" in [29]: sapere depositato in forme meccaniche, analogiche, digitali. Nell'ultimo e più recente stadio il sapere esplicitato è in gran parte quella "information" cui Harvey e Shannon hanno applicato il loro talento matematico e ingegneristico.

Un individuo può apprendere una porzione di questo sapere esplicito e circolante come dati, analogamente a come usa una merce, un oggetto tangibile, per un proprio scopo. Ne fa un'esperienza che diventa parte del suo corpo, così come il sapere appreso. La differenza è che mentre un oggetto resta lo stesso se è un bene prodotto per sé stessi o una merce prodotta per altri, la conoscenza tacita nel corpo e la conoscenza scambiabile che circola sono materialmente due mondi a sé.

Merce	*Informazione*
Bene prodotto da qualcuno, materializzato in un oggetto fruibile	Sapere prodotto da qualcuno, esplicitato in un messaggio fruibile

[12] Cfr. la critica ad Aaronson sopra

[13] «Si può dimostrare che negli esseri umani tutti i tipi di apprendimento – il condizionamento, l'apprendimento di abilit¿ motorie e l'apprendimento strumentale o condizionamento operante – hanno luogo senza alcuna consapevolezza o contributo da parte della coscienza» [18].

che si può dare a qualcun altro perché lo possa usare.	che si può trasmettere a qualcun altro perché la possa apprendere.
Bene prodotto per lo scambio, dissociato dal godimento personale.	Sapere fatto per la trasmissione, dissociato dal godimento personale.
Soddisfa bisogni con l'uso.	Soddisfa bisogni con la conoscenza.
Valore di scambio: quantità di denaro. Rende omogenei usi e bisogni soggettivi essenzialmente incommensurabili.	Valore di scambio: quantità di dati. Rende omogenee conoscenze soggettive incarnate essenzialmente incommensurabili.

Vediamo ora quali sostituzioni operare nel testo di Marx.

Come il denaro, l'informazione ha due lati. Da una parte è un concetto, una forma degli eventi. Ma attualmente è una quantità espressa in bit, materialmente circolante come dati. Da cui la sostituzione zero:

(S.0) 'dati' al posto di 'denaro'.

Poi, per quanto detto sopra, abbiamo:

(S.1) 'sapere' al posto di 'merce';

(S.2) 'informazione', che è la forma quantitativa di un sapere destinato allo scambio, prende il posto di 'prezzo', «cioè la forma di denaro delle merci» [20], una quantità che rappresenta il valore di scambio di una merce;

(S.3) 'ricchezza', che è l'universo delle merci virtualmente disponibili grazie al possesso di grandi quantità di denaro, diventa 'conoscenza' nel senso più generale, cioè l'universo del sapere virtualmente disponibile grazie al possesso di grandi quantità di informazione.

Vediamo cosa viene fuori con la riscrittura.

> Ogni sapere, nella misura in cui è un valore di scambio, è informazione, esprime una determinata quantità di dati solo in una forma incompiuta, giacché esso deve essere anzitutto posto in circolazione per venire realizzato – e che lo sia o meno è un fatto accidentale a causa della sua particolarità. Ma finché esso non esiste come informazione, ma soltanto nella sua determinatezza naturale, allora è soltanto un momento della conoscenza in virtù della sua relazione a un bisogno particolare che esso soddisfa, e in questa relazione esprime: 1) soltanto la conoscenza d'uso; 2) soltanto un lato del tutto particolare di questa conoscenza.
>
> Per contro i dati, a prescindere dalla loro utilizzabilità come sapere di valore, sono: 1) l'informazione realizzata; 2) ciò che soddisfa qualsiasi bisogno di sapere, in quanto possono essere convertiti nell'oggetto di qualsiasi bisogno di sapere, con assoluta indifferenza verso qualunque particolarità. Il sapere possiede questa proprietà soltanto per la mediazione dei dati. I dati la possiedono direttamente nei confronti di tutte le conoscenze particolari, e quindi nei confronti dell'intero mondo della conoscenza, della conoscenza in

> quanto tale. Nei dati la conoscenza generale è non soltanto una forma, ma al contempo il contenuto stesso. Il concetto di conoscenza è per così dire realizzato, individualizzato in un oggetto particolare.

Cosa sta dicendo Marx *adesso*?

Ci conferma che «nella sua determinatezza naturale», all'origine, la conoscenza è soggettiva. Che una conoscenza tacita è «un lato del tutto isolato della conoscenza», è un sapere incarnato in una persona che coincide con il "valore d'uso" che rappresenta per quella persona e basta. Questa conoscenza soggettiva può diventare informazione solo quando il sapere è esplicitato e posto in circolazione, trasmesso ad altri. Il che può accadere o no, e se non accade *la conoscenza non è informazione.*

Ci conferma poi che i dati, al contrario, soddisfano qualunque bisogno di sapere perché rappresentano qualunque sapere e lo rendono trasmissibile a prescindere dai suoi caratteri originari, «con assoluta indifferenza verso qualunque particolarità». Nel mediare la conoscenza, i dati la astraggono rispetto agli usi e alle persone, e così realizzano l'idea della conoscenza stessa «nella sua totalità, nella sua astrazione dal proprio particolare modo di esistere», come scrive Marx poco più avanti.

Ma non finisce qui: Marx suggerisce altri paralleli tra denaro e dati digitali

Denaro	*Dati*
Nella cultura capitalistica di mercato si tende a rappresentare tutto come denaro.	Nella società dell'informazione si tende a rappresentare tutto come dati. È opinione diffusa che l'insieme dei dati contenuti nell'internet sia l'*intera conoscenza umana*, come se non esistesse altra conoscenza all'infuori di quella.[14]
Tutto, in teoria, può essere tradotto in denaro.	Tutto, in teoria, può essere tradotto in dati (mediante input umano, sensori, convertitori A/D, ecc.)
Il denaro fa astrazione da ogni caratteristica particolare e personale.	I dati fanno astrazione da ogni forma particolare di conoscenza e di esperienza. Un film, un concerto, un diario, una cartella clinica, un anno di lavoro di un venditore, un modello di proteina, una lista della spesa, un progetto di

[14] «L'internet è marcia. (…) La colla che tiene insieme la conoscenza dell'umanità si sta disfacendo». [19]

	macchina, un viaggio… tutto diventa una sequenza di bit.
Il denaro, in forma di oro, può essere esso stesso una «merce di valore».	I dati, come contenuti o applicazioni, possono essere essi stessi "conoscenza di valore".
Il prezzo è la figura trasformata di un bene in un numero.	L'informazione è la figura trasformata di un sapere in un numero.

Come lo scambio di merci prende l'aspetto imposto dal denaro, lo scambio di conoscenza prende l'aspetto imposto dall'informazione. In entrambi i casi lo scambio avviene in virtù della totale *fungibilità* che i dati condividono con il denaro. In realtà la fungibilità dei dati supera di gran lunga quella del denaro[15], e la prova è che il denaro stesso oggi prende la forma di dati, i quali si trovano a un livello di astrazione superiore.[16]

I dati inoltre godono di una proprietà aggiuntiva: la calcolabilità automatica conferisce loro una incomparabile *funzionalità*, cioè la capacità di rappresentare funzioni e algoritmi e di esserne modificati, di elaborare e di essere elaborati nelle macchine informatiche. Grazie a Turing e Shannon, ogni cosa datificata è infinitamente riproducibile, processabile, trasmissibile.

4. DATI E METADATI

Questa riscrittura di Marx disegna un quadro coerente. Possiamo prenderlo come una prima *proof of concept* dell'ipotesi iniziale. Ma mettiamola ulteriormente alla prova.

Stavolta prendiamo un passo tratto dal *Capitale*. Qui Marx confronta due circolazioni, due opposte dinamiche di relazione tra persone, denaro e merci, e spiega come il denaro si trasforma in capitale.

La forma immediata della circolazione delle merci è M-D-M: trasformazione di merce in denaro e ritrasformazione di denaro in merce. Ma accanto a questa forma, ne troviamo una seconda, specificamente diversa, la forma D-M-D: trasformazione di denaro in merce e ritrasformazione di merce in denaro. [...] Il processo D-M-D ha le sue premesse *non in una distinzione qualitativa* dei suoi estremi, poiché essi sono entrambi denaro, ma solo nella loro *diversità quantitativa*. Alla fine si preleva dalla circolazione più denaro di quanto ne fosse stato introdotto all'inizio. Il cotone comprato a 100 sterline, p.es., viene poi

[15] Proprio in opposizione a questa estrema fungibilità è fiorito il mito dei Non-Fungible Tokens: prima che se ne impadronisse la furia speculativa sull'intangibile, manifesta l'insopprimibile necessità umana del rapporto con le cose materiali che cerca disperatamente di realizzarsi anche nel reame alieno del digitale. Cfr. [21]

[16] Ciò indica che l'universo originale di Marx sia abbracciato e assorbito dalla mia trasposizione. I due discorsi non restano distinti in un parallelo passato vs. presente ma sono convoluti l'uno nell'altro: un intreccio assai interessante che spalanca alla ricerca un territorio vastissimo

venduto per 100 + 10 sterline. La forma compiuta di questo processo è quindi D-M-D', dove D' = D + ΔD = la somma di denaro originariamente anticipata, più un incremento. Chiamo *plus-valore* (*surplus value*) questo incremento, questa eccedenza sul valore originario. Quindi il valore originariamente anticipato non solo si conserva nella circolazione, ma in essa *aumenta* pure la sua *grandezza di valore*, aggiunge un *plusvalore*, cioè si *valorizza*. È questo movimento che lo *trasforma* in capitale. [20]

Prima di tutto dobbiamo capire quali sono i nostri sostituti per le due circolazioni, con una breve genealogia comparata. In Marx lo scambio economico primordiale è M-M, merce per merce; analogamente possiamo dire che lo scambio primordiale di sapere è U-U, da umano a umano, per rispecchiamento diretto tra corpi in presenza fisica. Con i linguaggi e con le ritenzioni terziarie la conoscenza si fa esplicita, emergono mediatori tra i due U che li allontanano nello spazio e nel tempo, finché dopo Turing e Shannon la mediazione tende a essere svolta in esclusiva dai dati digitali D.

Per noi i due circuiti sono quindi U-D-U e D-U-D. Con essi riscrivo subito il brano, rimandando a dopo gli altri chiarimenti.

> La forma più semplice di circolazione della conoscenza è U-D-U: trasformazione di sapere personale in dati, e ritrasformazione di dati in sapere personale. Ma accanto a questa forma ne troviamo una seconda specificamente diversa: D-U-D, trasformazione di dati in sapere personale, e ritrasformazione di sapere personale in dati. [...] Il processo D-U-D ha le sue premesse non in una distinzione qualitativa dei suoi estremi, poiché essi sono entrambi denaro, ma solo nella loro diversità quantitativa. Alla fine si prelevano dalla circolazione più dati di quanti ne fossero stati introdotti all'inizio. Un contenuto di 1MB, p.es., si ripresenta come 1MB + 1KB. La forma compiuta di questo processo è quindi D-U-D', dove D' = D + ΔD = la quantità di dati originaria, più un incremento. Chiamo plusvalore (surplus value) questo incremento o eccedenza sul valore originario. Quindi il valore originariamente anticipato non solo si conserva nella circolazione, ma in essa aumenta pure la sua grandezza di valore, aggiunge un plusvalore, cioè si valorizza. È questo movimento che lo trasforma in capitale.

Del circuito U-D-U abbiamo continue esperienze nella società dell'informazione. Ad esempio quando chattiamo con qualcuno attraverso un'app social, dove l'U iniziale e l'U finale sono persone diverse. O quando cerchiamo qualcosa online con un motore di ricerca, caso in cui l'U iniziale e l'U finale coincidono.

La parte delicata è il circuito D-U-D, con U come *human in the loop*. Per comprenderlo correttamente si deve ricordare che nella società dell'informazione *machine-biased* ogni cosa va vista nell'ottica delle macchine informatiche. Ebbene, da questo punto di vista i dati sono solo sequenze di bit. Zero significato, qualunque sia il significato che *noi* diamo alla parola 'significato'. Gli umani estraggono un "significato" dai fenomeni sensibili in quanto li valutano in rapporto alle proprie condizioni di viventi. Lo fanno per natura. E devono farlo anche coi

dati, ragion per cui si servono di macchine informatiche. Le macchine invece non lo fanno. Allora il punto è resistere alla tentazione di leggere da umani la D finale del circuito D-U-D come se fosse un *significante*, un contenuto digitale destinato alla nostra fruizione. Se così fosse, il ciclo sarebbe in realtà D-U-D(-U), perciò sarebbe l'*altro ciclo*, U-D-U, "la circolazione semplice della conoscenza".

Marx affronta un problema analogo col suo D-M-D: anche qui la D finale non va vista come denaro con cui acquistare beni e soddisfare bisogni, altrimenti ricadremmo nella circolazione "sana" (D-)M-D-M. No, la D finale di D-M-D è *denaro fine a sé stesso*: un numero. Così pure nel nostro caso la D finale di D-U-D è *dati fini a sé stessi*: di nuovo un numero (la sequenza di bit).

Come una somma di denaro, anche un insieme di dati «si può distinguere da un altro soltanto per la sua grandezza». Nel caso del denaro è il semplice importo; nel caso dei dati è il numero in base 2 che corrisponde a quella sequenza, oppure un numero calcolato a partire da quei dati: ad es. il "peso" in byte, una checksum, una funzione hash, ecc.

Ma cos'è quel curioso 1KB che si somma all'1MB di partenza digerito dallo *human in the loop*?

Si tratta di *metadati*, dati che descrivono altri dati. Nella circolazione D-U-D' il mediatore umano viene *usato* per produrre quel ΔD di metadati che espande il valore iniziale dei dati aggiungendo ad essi *qualcosa di umano*, un quid di *significato* che ci compete in rapporto a loro. Questa espansione, tuttavia, avviene a un costo: il sapere soggettivo di U deve essere *codificato all'origine* per poter confluire subito nei dati.

I metadati di ΔD possono essere generati dagli umani sia come tracce collaterali alle nostre interazioni con il mondo dei dati – chiavi di ricerca, trend topic, pattern di navigazione, reazioni sui social network, georeferenziazione, ecc. – sia intenzionalmente con l'input manuale di valori[17]. Fra questi ultimi metadati hanno un posto speciale le *data labels o annotations*, le etichette che associano significati ai data point dei dataset con cui si addestrano i modelli IA[18].

Benché eseguita da umani, spesso manodopera a basso costo[19], la data *annotation* è una pura funzione informatica: associa *permanentemente* a un contenuto digitale un insieme definito di tag, che è una restrizione formale del

[17] La distinzione tra metadati intenzionali e non intenzionali non esiste nell'ottica della macchina, per la quale i commenti e le reazioni a un post su Facebook possono essere metadati di quel post, mentre dal nostro punto di vista sono parte di una conversazione. In realtà cosa sia o non sia metadati per la macchina dipende dalla struttura dei dati definita nel codice.

[18] La domanda per l'etichettatura semantica di immagini e video dilaga nei più vari settori al seguito delle applicazioni IA: dalla ricerca pura si arriva ad es. a Netflix per classificare film e serie, o a Tesla per addestrare le auto a guida autonoma. Ma anche il feedback umano nel RLHF (Reinforced Learning with Human Feedback) usato per migliorare i Large Language Models si può considerare una forma di data annotation.

[19] Vedi ad es. [22], [30], [31]. I casi emergono di pari passo con l'espansione dell'industria.

linguaggio naturale[20]. L'umano qui è necessario in quanto svolge una funzione impossibile per la macchina; tuttavia è costretto a svolgerla *nei termini della macchina*. Il setting dell'*imitation game* continua a dettar legge: anche qui l'umano si deve limitare a soli testi scritti.

Questi metadati sono il risultato del lavoro cognitivo di qualcuno, perciò *incorporano definitivamente la sua interpretazione* e portano con sé la soggettività di quel qualcuno. Ma attraverso la filiera D-U-D' questa soggettività è dissipata per sempre quando il contenuto interpretato è assorbito in D' e così cristallizzato, reso "oggettivo". Nelle parole di Marx: «il movimento mediatore scompare nel proprio risultato senza lasciar traccia». [20]

Chiunque, ad esempio, utilizzi un dataset di training come ImageNet[21], impiega quella conoscenza datificata as is, senza alcun riferimento alla provenienza sia delle immagini che delle etichette che dicono cosa c'è nelle immagini.

Quelle etichette sono tutt'altro che accessorie. Formano la "ground truth" del modello IA, l'ordito di lillipuziani ormeggi referenziali che lo ancora al nostro mondo. Guidano il training fino al rilascio in produzione. Condizionano intimamente il suo comportamento futuro. Eppure alla fine tutto appare *come se le etichette fossero proprietà native di dati grezzi*, dati che si sono generati da sé.

La crasi di dati e metadati è un nuovo punto di partenza per l'osservazione e la sperimentazione, come se fosse un fenomeno naturale. Dati da cui derivare altri dati. Gli operatori umani in D-U-D' sono scomparsi e il processo è diventato D-D'. Il ΔD fornito dall'umano serve ad accrescere la funzionalità e l'efficienza dei dati, quindi il loro valore economico, inclusa una maggiore capacità di auto-riprodursi.

Anche gli stessi ricercatori che progettano e addestrano un modello IA sono in un ciclo D-U-D': una volta messo il sistema in produzione svaniscono perché questo è precisamente lo scopo della rete neurale: produrre dati da altri dati, eliminare U, abilitare D-D'.

5. AUTO-ESPANSIONE E POTERE DEI DATI

Per Marx l'evoluzione fatale del denaro-capitale è l'auto-riproduzione D-D': denaro che cresce da sé indefinitamente, con un tasso d'interesse, senza più bisogno di convertirsi in merci utili o necessarie alle persone. È un «feticcio automatico»: "feticcio" per il valore simbolico e mistico di cui è carico, "automatico" per la sua capacità di auto-espandersi. Questa crescita infinita e indipendente dai bisogni umani porta nel cuore della società umana un principio alieno di destabilizzazione.

Analogamente i dati tendono alla propria auto-riproduzione D-D' senza più bisogno non solo di esserci *realmente utili*, ma perfino di *significare* qualcosa per

[20] Ancora l'acutezza di Wittgenstein [10]: «si può anche dire che sia la rappresentazione di un linguaggio più primitivo del nostro».
[21] Corpus di oltre 14 milioni di immagini etichettate, che viene utilizzato da ricercatori di tutto il mondo per addestrare e testare reti neurali per il riconoscimento visivo di immagini: https://image-net.org/

noi. Come il denaro di Marx, i dati sono un «feticcio automatico» che porta nel cuore della società umana un principio alieno di destabilizzazione. Lo sono molto più del denaro, perché l'auto-riproduzione dei dati:

- è *massiva* – i dati prodotti dai sistemi IA si moltiplicano in quantità e velocità che non hanno rapporto con le capacità cognitive umane;
- è *necessaria* – perché più cresce il volume dei dati disponibili più siamo costretti ad affidare la loro elaborazione ai sistemi IA;
- è *credibile* nel mimare comportamenti umani, può impersonare un interlocutore e trarci in inganno.

L'onnipotenza del denaro D che può *comprare* tutto, raccontata da Marx, diventa nel nostro mondo l'onnipotenza dei dati D che possono dire tutto. Il potere dei dati non sta più nell'appropriazione materiale di cose e persone, come per il denaro, ma nella *conoscenza*. Soprattutto nella conoscenza delle persone come strumento di dominio.

Pensiamo all'effetto del circuito D-D' nel *profiling* ([23]): i dati che compongono il ritratto informatico di una persona fisica possono essere automaticamente rielaborati per derivarne altri dati aggiunti al profilo, e questo ΔD consiste di fatto in nuove presunte "verità" sull'identità di quella persona, considerate pari agli altri dati. Un esempio tipico: D è il mio punteggio di affidabilità creditizia e/o il mio codice di avviamento postale, il ΔD derivato è una mia "disposizione a delinquere" per cui l'autorità giudiziaria può prendere provvedimenti restrittivi a mio carico. Pur essendo solo *montature calcolate*, questi ΔD sul nostro conto possono sottrarre credito, lavoro, libertà, cambiare destini a seconda del potere a loro concesso.

CONCLUSIONE

Qui infatti si pone a noi la scelta tra due strade.

α Dare a D' un ruolo *descrittivo* ausiliario alle decisioni umane in un ambiente sociotecnico in cui le macchine sono concepite per collaborare con noi e aiutarci a generare senso[22]. D' è subordinato ad U, i dati sono oggetto di interpretazioni e discussioni nel campo umano, le valutazioni non si basano sul «rendere i task stupidi» (L. Floridi) e non si limitano mai a ciò che è riducibile al «linguaggio primitivo» di dati e numeri. Il circuito della conoscenza è (D-)U-D'-U.

ω Dare a D' il ruolo *performativo* di una decisione automatica, con conseguenze reali sulla vita degli umani, come un sistema giudiziario imperscrutabile. Qui i criteri di valutazione hanno la forma dei dati stessi (semplici parametri numerici) e possono essere incorporati nei dati come metadati senza origine. Data l'universalità dell'informazione, idealmente *il mondo diventa una chiusura operativa di D*. Il circuito della conoscenza è D-(U-)D'.

[22] V. ad es. [24].

Com'era per Marx il ciclo M-D-M, per noi è la via α quella sana e auspicabile:

- cerca una continuità antropologica e ci offre una chance di porre rimedio, per quanto possibile, alla discontinuità essenziale che separa l'informazione dal mondo fisico cui apparteniamo;
- non lascia mano libera alla ristrettezza dei criteri numerici, ma li completa con ciò che è incalcolabile: sentimenti, immedesimazione, intuizione, dialogo, narrativa, mito, ecc.;
- assume che qualsiasi processo artificiale che ci riguarda abbia come punto di fuga gli interessi fondamentali degli esseri umani come parte della natura.

Al contrario, la via ω è un piano inclinato e scivoloso che si perde nell'astrusa apoteosi della cultura *machine-biased*, in cui i tratti tipici degli umani e dei viventi gradualmente svaniscono dai nostri disegni per la società presente e futura.

Essenza di questa alienazione è il distacco fra gli esseri umani. Marx deplora che le persone finiscano per parlarsi tramite le cose; oggi quando ci guardiamo tramite le macchine siamo costretti a vederci l'un l'altro come le macchine ci vedono, cioè come *cose immateriali* fatte di proprietà e funzioni. Lo sguardo mediato dalle macchine riproduce e *istituzionalizza* l'empatia zero propria dei modelli di intelligenza e di comunicazione di Turing e Shannon, gli stessi in ogni tempo perseguiti dai promotori della "perfettibilità" umana mediante logica e macchine.

Questa depressione empatica strutturale apre la strada a nuove oppressioni, e anche in questo il capitale-dati è più efficiente del capitale-denaro. La società *data-driven* del resto prosegue senza soluzione di continuità la società *market-driven*, e l'economia dei dati generalizza l'economia capitalista in una forma più astratta che offra il minore attrito possibile nelle condizioni attuali.

BIBLIOGRAFIA

[1] Turing, A. M. (1950). *Computing machinery and intelligence*, in *Mind*, 59, pp. 433-460. Trad: *Macchine calcolatrici e intelligenza*, in Somenzi V. (a cura di) (1965). *La filosofia degli automi*, Boringhieri.

[2] Hartley, R. (1926). *The Transmission of Information*, Vol. 118, Part 2, December 18, 1926, page 874-875.

[3] Gleick, J. (2015). *L'informazione. Una storia, una teoria, un diluvio*, Feltrinelli. (pos. Kindle 86-89)

[4] Turing, A. M. (1936-7). *On Computable Numbers, with an Application to the Entscheidungsproblem*, in *Proc. of the London Mathematical Society* vol. 42 (1936-7), pp. 230-265.

[5] Shannon, C. (1948). *A Mathematical Theory of Communication*.

[6] Floridi, L. (2014). *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, Raffaello Cortina. (pos. Kindle 1634-1635)

[7] Heidegger, M. (1988), *Filosofia e cibernetica*, ETS.

[8] Aaronson, S. (2011). *Why Philosophers Should Care About Computational Complexity* – arXiv:1108.1791 (trad. mia).

[9] Rizzolatti, G., Sinigaglia, C. (2006). *So quel che fai. Il cervello che agisce e i neuroni specchio*, Raffaello Cortina Editore.

[10] Wittgenstein, L. (2014). *Ricerche filosofiche*, Einaudi.

[11] Kurzweil, R. (2008). *La singolarità è vicina*, Apogeo.

[12] Marx, K. (1975). *Manoscritti economico-filosofici del 1844*, Einaudi.

[13] Marx, K. (1968). *Lineamenti fondamentali della critica dell'economia politica 1857-1858*, vol. 1, La Nuova Italia.

[14] Marx, K. (1844). *Comments on James Mill, Éléments D'économie Politique* (online) https://www.marxists.org/archive/marx/works/1844/james-mill/ (ultimo accesso 04/01/23); traduzione mia.

[15] Polanyi, M. (1958). *Personal Knowledge: Towards a Post-Critical Philosophy*, Routledge & Kegan Paul Ltd.

[16] Polanyi, M. (1966). *The Tacit Dimension*, Doubleday.

[17] Polanyi, M. (1988). *La conoscenza inespressa*, Armando.

[18] Jaynes, J. (2014). *La natura diacronica della coscienza*, Adelphi.

[19] Zittrain, J. (2021). *The internet is rotting* (online) https://www.theatlantic.com/technology/archive/2021/06/the-internet-is-a-collective-hallucination/619320/ (ultimo accesso: 04/01/23)

[20] Marx, K. (1980). *Il capitale*, I, Editori Riuniti.

[21] Han, B.-C. (2022). *Le non cose*, Einaudi.

[22] Murgia, M. (2019). *AI's New Workforce: The Data-Labelling Industry Spreads Globally* (online) https://medium.com/financial-times/ais-new-workforce-the-data-labelling-industry-spreads-globally-f472cb1bac09 (ultimo accesso 04/01/24)

[23] https://www.altalex.com/documents/altalexpedia/2018/05/28/profilazione (ultimo accesso: 04/01/23)

[24] Cabitza, F., Campagner, A., Ciucci, D., Seveso, A. (2019). "Programmed Inefficiencies in DSS-Supported Human Decision Making." In: Torra V., Narukawa Y., Pasi G., Viviani M. (eds) *Modeling Decisions for Artificial Intelligence. MDAI 2019. Lecture Notes in Computer Science*, vol 11676. Springer, Cham. https://doi.org/10.1007/978-3-030-26773-5_18

[25] Turing, A.M. (1994). *Intelligenza meccanica*, Bollati Boringhieri.

[26] Natal, J.; Ávila, I.; Tsukahara, V.B.; Pinheiro, M.; Maciel, C.D. (2021). *Entropy: From Thermodynamics to Information Processing*. Entropy 23(10):1340. https://doi.org/ 10.3390/e23101340

[27] Galilei, G. (1632). *Dialogo sopra i due massimi sistemi del mondo*, Landini.

[28] Meister, M. e Berry, M. J. (1999). *The neural code of the retina*. Neuron, 22(3):435–450.

[29] Stiegler, B. (1994). *La Technique et le Temps 1: La faute d'Épiméthée*, Galilée. Trad. it. *La colpa di Epimeteo. Vol. 1: La tecnica e il tempo*, Luiss University Press (2023)

[30] Perrigo, B. (2023). *Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic* (online) https://time.com/6247678/openai-chatgpt-kenya-workers/

[31] Ingram, D. (2023). *ChatGPT is powered by these contractors making $15 an hour* (online) https://www.nbcnews.com/tech/innovation/openai-chatgpt-ai-jobs-contractors-talk-shadow-workforce-powers-rcna81892

BIOGRAFIA

Stefano Diana è un autore e ricercatore indipendente, con background in ingegneria informatica, che indaga principalmente sui rapporti tra l'umano e la tecnologia. È stato docente di comunicazione e consulente per molte aziende ICT. Il suo saggio del 1997 "W.C.Net. Mito e luoghi comuni di Internet" (minimumfax) fu raccomandato da Umberto Eco. Nel 2016 ha pubblicato "Noi siamo incalcolabili. La matematica e l'ultimo illusionismo del potere" (Stampa Alternativa) che esamina i limiti del pensiero formalista e critica l'egemonia del calcolabile come patogeno socioculturale.

E-mail: stefanodiana.inc@gmail.com

Computer, Video, Scienza, Tecnica, Graphic Art … Panta rei

AnnaMaria Carminelli Gregori, Alessandro Marassi, Piero Delise

Sommario

All'inizio i computer sono stati utilizzati come macchine veloci per decifrare messaggi per scopi militari e poi per i calcoli di fisici e scienziati in genere. In seguito sono stati scoperti da artisti per scopi di ricerca in campo musicale e nella pittura.

Vengono presentati ricordi ed esperienze, in particolare nella computer graphics, degli ultimi 50 anni vissuti col computer per …. amico! Talvolta visto anche come nemico.

Abstract

At the very beginning computers have been used as fast machines to decipher messages for military purposes and then to perform calculations by physicists and scientists. After a while they have been discovered by artists to do research in music and painting.

Memories and experiences, mainly about computer graphics, of the last 50 years lived with the computer as a friend are presented...! Sometimes it is also seen as an enemy.

Keywords: Computer Graphics, Computer Art, History of Technology, Human-Machine Interaction, Artificial Intelligence, Robotics.

1. Il passato remoto ed il ruolo

Ci sono momenti ed occasioni che non si dimenticano. Con il tempo che passa tutto appare più sfumato, ma sempre vivo.

Nei primi anni '60 l'incontro ravvicinato col computer dava la sensazione di imbattersi con un extraterrestre.

Prima di tutto le sue dimensioni, il suo linguaggio, i suoi numeri, la sua velocità di calcolo, i suoi tasti con significati speciali … tutto era extra. Non era la lavatrice di casa, era di più … Neppure la macchina da scrivere gli assomigliava, tantomeno quella da cucire. Quando a furia di prove e di studi, cominciava a diventare più docile era già il tempo di cambiarlo, di prenderne uno più potente, più veloce … più extra.

Cambiare, aumentare, modificare. Anche il suo ruolo cambiava. Da veloce

macchina di calcolo, tanto cara ai Fisici, agli Scienziati in generale che finalmente potevano convalidare le loro teorie con risultati numerici, diventava lo strumento per fare ricerca su tanti temi, per esempio quello musicale, artistico.

Si arricchivano anche le sue componenti, le stampanti sempre più veloci venivano affiancate dai plotter, ossia da strumenti dotati di un cilindro rotante verticalmente e di una penna movibile orizzontalmente che permettevano di ottenere disegni di ogni tipo, componendo e calibrando i due movimenti.

Segue, in Fig. 1, un'immagine di un plotter Calcomp 565 del 1966

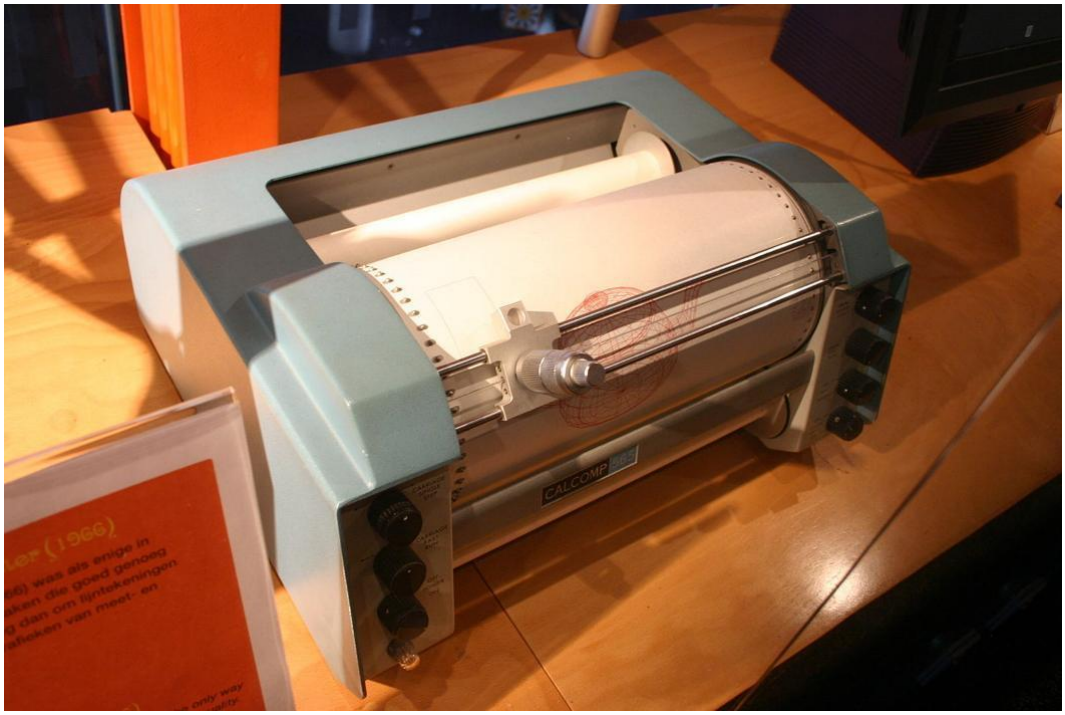

Figura 1
Plotter Calcomp 565 del 1966

La tecnologia, dunque, ha fatto passi da gigante ed oggi ci ritroviamo con strumenti molto più potenti e più facili da usare. E per capire meglio cosa significasse realizzare un'applicazione come quella qui di seguito descritta vale la pena di fare un salto indietro, negli anni '60.

Una volta ideato il programma, questo veniva dapprima scritto in uno dei pochi linguaggi a disposizione, come l'Assembler o il più user friendly Fortran (all'epoca era molto diffuso il Fortran IV).

Ogni istruzione del programma doveva poi essere trasformata in una scheda perforata (v. esempio in Fig. 2)

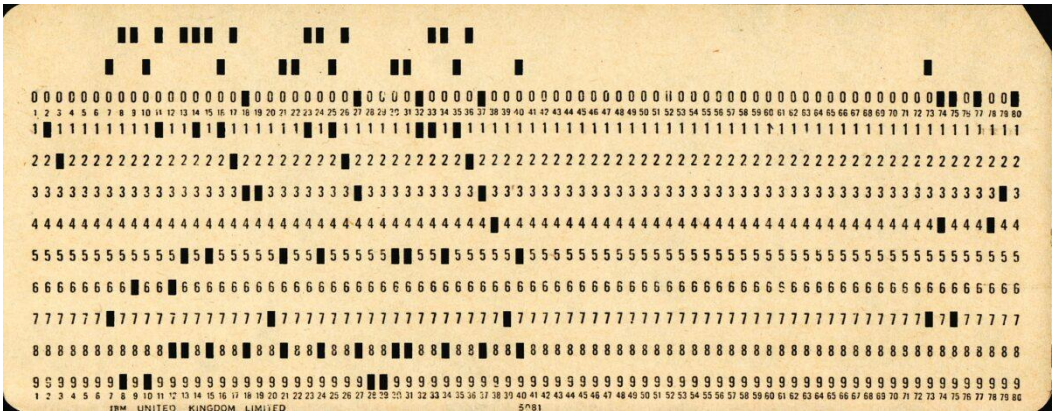

Figura 2
Scheda perforata

utilizzando apposite macchine perforatrici (v. esempio in Fig. 3), ingombranti e rumorose, che trasformavano le digitazioni su una tastiera in combinazioni di fori su schede in cartoncino. (Notare che le schede avevano sostituito il nastro di carta perforato che nel 1963 veniva usato nei computer ad uso scientifico come lo storico IBM 1620. In questo caso le macchine perforatrici erano più simili a macchine da scrivere.)

Figura 3
Macchina perforatrice

Niente a che vedere, dunque, con la facilità e l'immediatezza con cui oggi creiamo o modifichiamo programmi.

Anche la Computer Graphics [16] sia in termini hardware che software era ai primordi. Non esistevano interfacce hardware né algoritmi e librerie software che permettessero una rappresentazione visiva di dati numerici in modo agevole. Furono gli scienziati a dare il via al filone della Computer Art: le loro ricerche su fenomeni fisici collegati all'acustica ed alla visione [13] [17] diedero vita ai primi risultati grafici con immagini gradevoli e successivamente ad animazioni delle stesse anche a scopo didattico.

2. L'applicazione

Con una composizione grafica fatta sul Plotter dal titolo "Rain Drops", nel 1968, Claudio Galmonte e Annamaria Carminelli Gregori vinsero un concorso di Computer Art a Los Angeles. Proprio dal ricordo di Claudio Galmonte, uno dei pionieri della Computer Art a Trieste, sono scaturite queste pagine che mettono in luce come l'immagine realizzata fosse una composizione grafica "randomizzata". Le sue caratteristiche casuali erano legate all'aspetto della realtà che si voleva rappresentare, ossia alle gocce di pioggia che cadono sul parabrezza dell'automobile.

Proprio dal parabrezza dell'automobile in un giorno di pioggia partì l'idea di Rain Drops. Naturalmente ci furono discussioni, dibattiti, critiche, apprezzamenti … su tutto, dai colori alle circonferenze. Poi le idee si concretizzarono, tenendo presente che in un'immagine ci "deve essere un solo elemento che catturi l'attenzione, non mille!" come sentenziò Antonio Marussi, allora Presidente del Centro di Calcolo dell'Università di Trieste. E intervenne anche la fantasia di Galmonte che pensò a distribuzioni casuali diverse in corrispondenza alle situazioni da visualizzare: distribuzione casuale uniforme per i centri delle gocce (la pioggia cade uniformemente) e distribuzione casuale Gaussiana per i diametri delle gocce in modo da escludere quelli troppo grandi o troppo piccoli. E poi ci furono altre modifiche relative alla maggiore o minore intensità della pioggia. Si susseguirono giornate piene di studi, proposte, formule, modifiche, prove, schizzi contrasti … Una sintesi della ricerca e delle idee formulate in quei giorni è riportata nelle immagini seguenti tratte da appunti conservati in archivi dimenticati. Si inseriscono solo come una panoramica sintetica del lavorio intellettuale e materiale di quel periodo (Fig.4).

Si arrivò ad un'immagine simpatica che variava in dipendenza dall'incipit iniziale. Si classificò bene tra 600 concorrenti, prima in Italia, seconda in Europa, decima in totale. Gli autori ricevettero premi, certificati, lettere di congratulazioni e di encomio, interviste alla Radio, alla Televisione, richieste di collaborazione da società esistenti negli Stati Uniti, come per esempio l'ACM (Association for Computing Machinery).

Purtroppo in Italia i tempi non erano maturi per la Computer Art [17], il Centro di Calcolo dell'Università si doveva occupare di servizio … Così nonostante l'encomio solenne dell'Università di Trieste, i dirigenti del Centro stesso

pensarono fosse meglio abbandonare il tipo di “ricerca” appena nato e lavorare solo per il servizio scientifico.

La composizione di Rain Drops, però piacque e fu pubblicata su tanti libri e riviste,

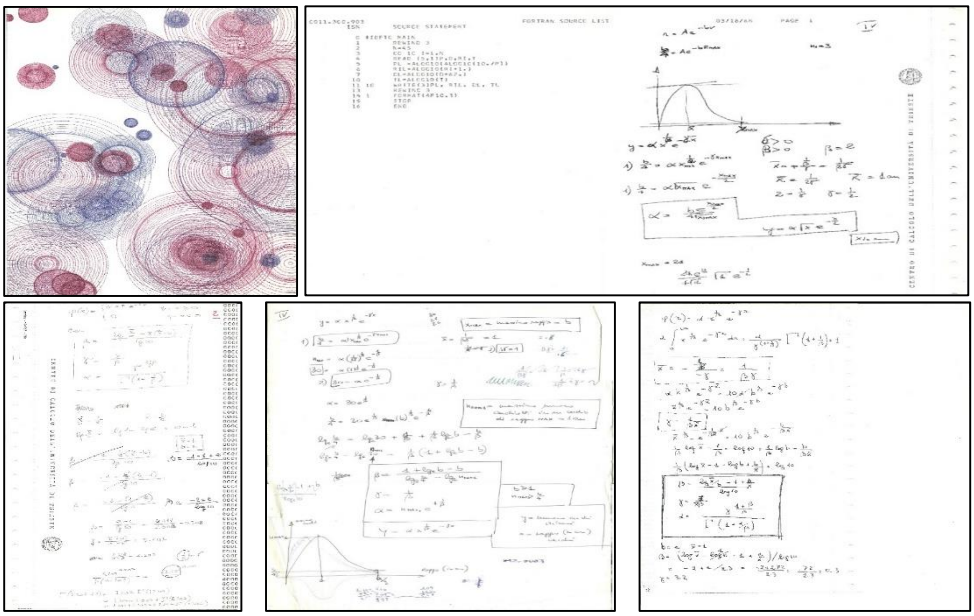

Figura 4

citata come il primo lavoro italiano di Computer Art, sia nella forma originale come appare in Fig.4 e 5, sia rielaborata successivamente da Annamaria Carminelli sul

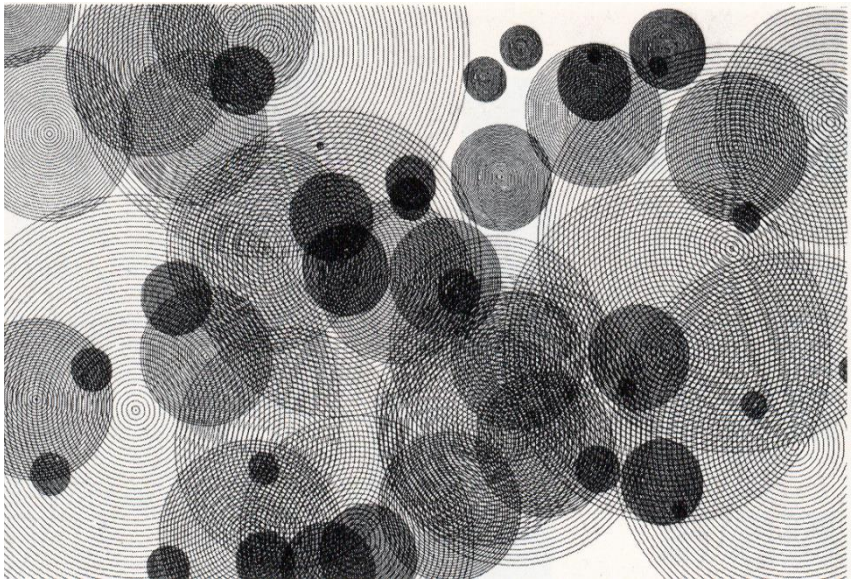

Figura 5

monitor del CROMEMCO come in Fig. 6 e 7, e finalmente sfociata in un breve

Video del 2022 in cui le gocce si animano e danzano con la musica di Burt Bacharach… Ma questo è il presente mentre le ultime due immagini rappresentano proprio una parte del passato prossimo.

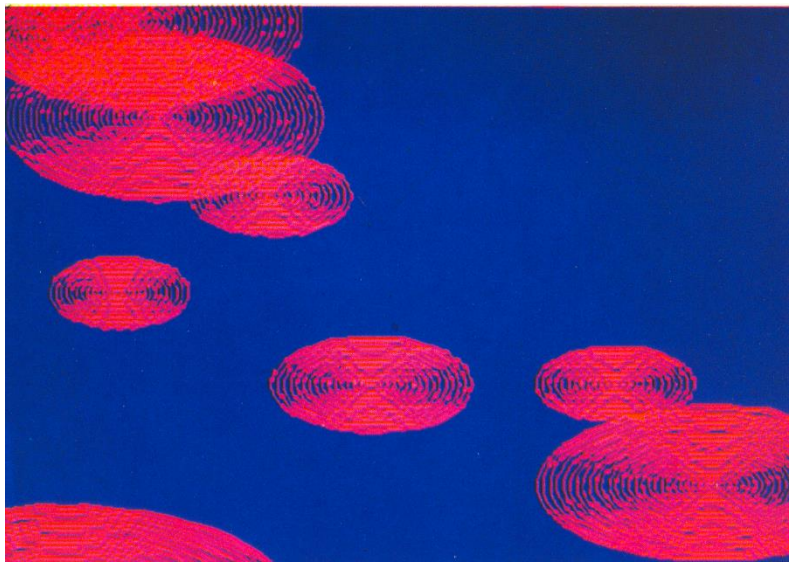

Figura 6

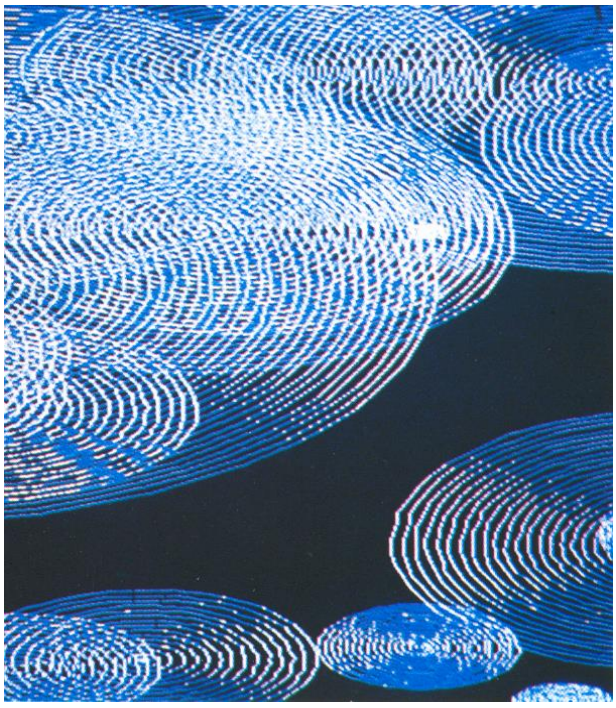

Figura 7

Ormai siamo negli anni '70 - '80 del secolo scorso e si è passati ai computer collegati in rete (il CROMEMCO è un esempio di tale collegamento) con funzioni e utilizzi distribuiti.

Nel 1970 ritorna a Trieste il pittore (nato a Trieste nel 1938) Edward Zajec. Anche tramite il pittore Černigoj inizia una collaborazione con Matjaž Hmeljak del Centro di Calcolo dell'Università di Trieste nel settore - allora nuovo - della Computer Art [14][15].

Eccone un primo esempio in Fig. 8: foto di un'immagine prodotta dal programma TVC nel 1971 (idea e progetto di Zajec, programma in Fortran 4 di Hmeljak eseguito sul calcolatore del Centro di Calcolo IBM 7040).

Zajec aveva iniziato ad usare il computer per il suo lavoro di artista già nel 1968 (lo stesso anno di Galmonte e Carminelli), portato dal suo percorso estetico, come l'uso di un reticolo di moduli simili tra loro. L'idea di costruire un'immagine con sequenze di righe programmate algoritmicamente affascinava Hmeljak.

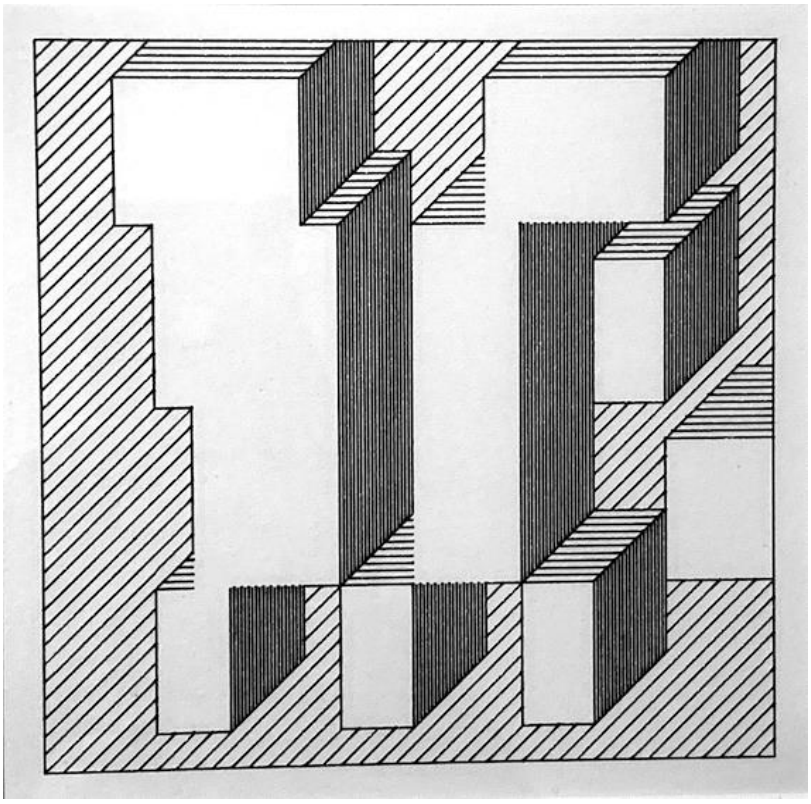

Figura 8
"TVC", 1971, programma in Fortran 4 eseguito su IBM 7040

La collaborazione ebbe un inizio difficile, finche' non fu trovato un linguaggio comune tra il tecnico Hmeljak e l'artista Zajec. La spinta per il lavoro comune fu la volontà di realizzare algoritmicamente le idee costruttiviste di Zajec con l'uso

del calcolatore, e di sperimentare le nuove possibilità da esso offerte, come ad es. l'interazione di chi guarda un'immagine con il programma che crea questa immagine (versione interattiva del progetto "TVC", realizzata nel 1974 sul CDC6400 con l'uso della console di sistema, chiaramente non utilizzabile praticamente a tal fine). Inoltre nei due autori c'era anche la voglia di comunicare agli altri e di integrare le loro sensazioni con l'ambiente. Così nacquero diverse mostre in Italia e all' estero (Trieste, Gorizia, Venezia, Lubiana, USA ...). L' evolversi del loro dialogo è sintetizzato nelle immagini riportate nelle due figure seguenti. A sinistra, Fig. 9, un'immagine del progetto "Logical Moments in Color" o "LMC" del 1976, a destra, Fig. 10, due immagini del progetto "Matrix" del 1978:

Figura 9

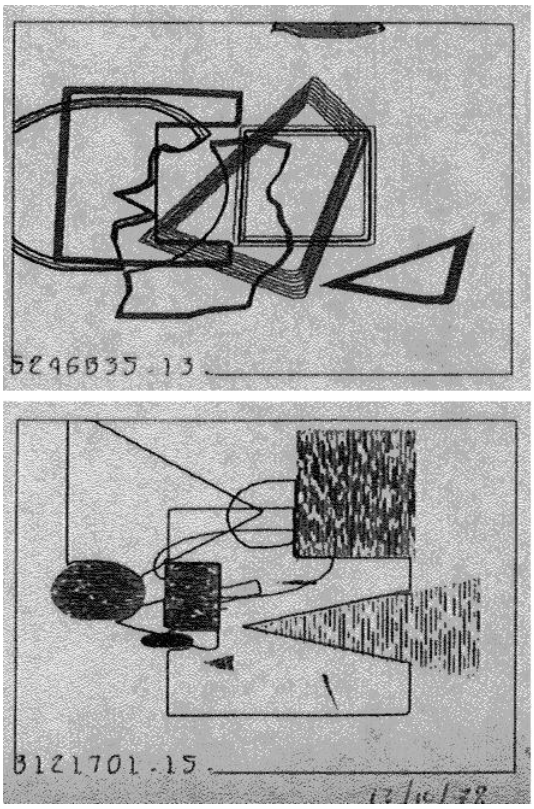

Figura 10

Poi Zajec andò negli Stati Uniti dove continuò con molto successo le sue ricerche nell'ambito della Computer art. Nel periodo degli anni '80, Zajec fece ricerche nell'ambito del video digitale astratto, in particolare sull'animazione del colore, teoria che chiamò "Orphics". Zajec realizzò diverse animazioni di colori astratti nello stile "Orphics", di cui si riportano due fotogrammi in fig. 11.

Tra le animazioni di colori di Zajec da ricordare "Chromas", presentato anche a Trieste, con musica del triestino Coral. Questi lavori fanno parte della storia dell'arte digitale, e sono presenti in molti musei negli USA e in Europa, ricordati in Biblio [14], [17].

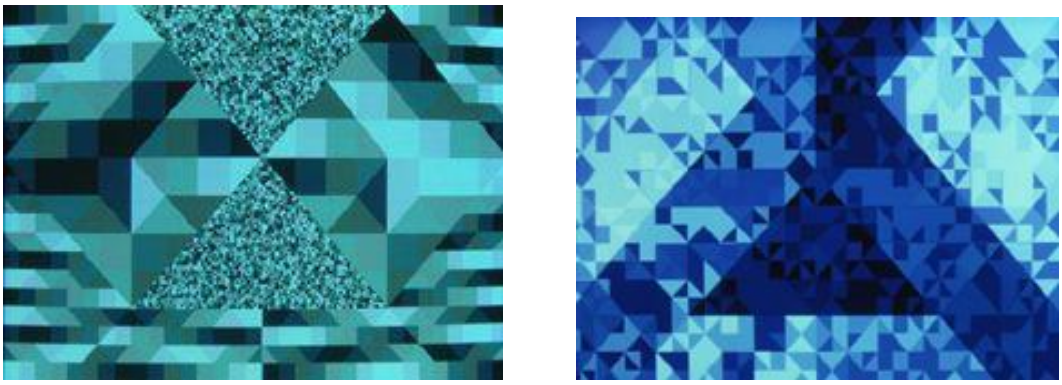

Figura 11
Due fotogrammi da “Chromas", Digital Video di Zajec, 20 min, musica di Giampaolo Coral, Seconda Sonata. Piano: Corrado Gulin.

Dopo anni di lavoro con Zajec (1970-1980) Hmeljak riprende a lavorare su idee e progetti propri, ripartendo dalle basi, con idee molto semplici, quali l'uso della casualità (randomizzazione) nella costruzione di semplici strutture di linee come in Fig. 12.

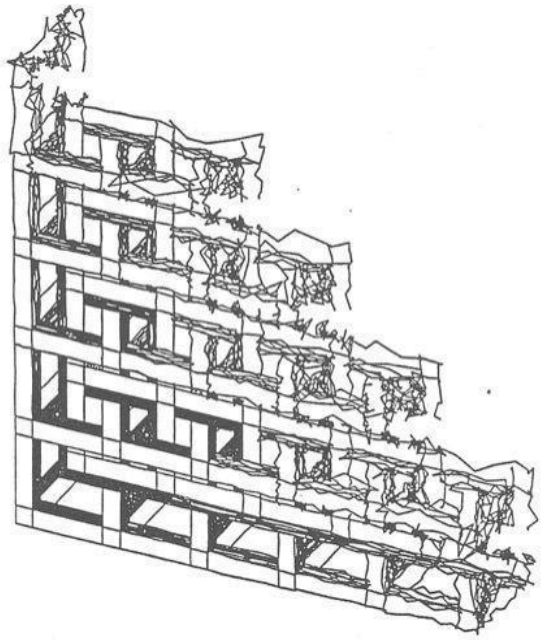

Figura 12
“RKT”, 1985

In alcune delle immagini di Hmeljak c’è talvolta un quid in più ... un puntino rosso in un’immagine a toni di grigio, un segnale di disturbo in una struttura simmetrica, un’idea aggiunta, quasi a non prendersi troppo sul serio, a scherzare col suo lavoro, a porsi domande. Un esempio in Fig. 13:

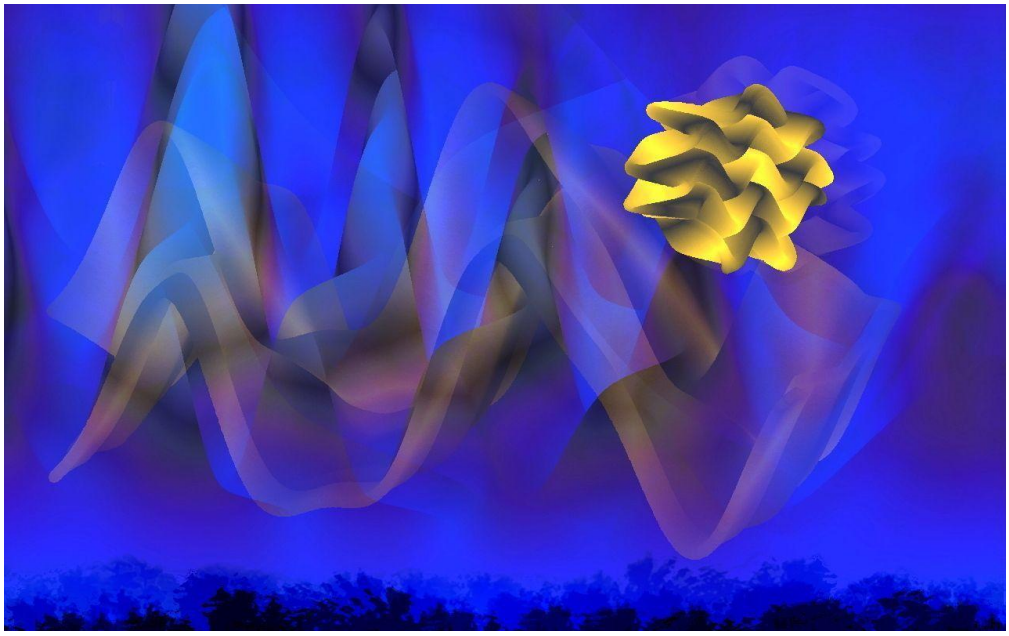

Figura 13

In alcune ricerche Hmeljak (tra altre) studia le possibilità espressive delle trasformazioni nello spazio bidimensionale (x1,y1) = f(x,y), inizialmente con tracciatura delle linee deformate, come ad es. in Fig. 14 del 1996, ”Love problem”;

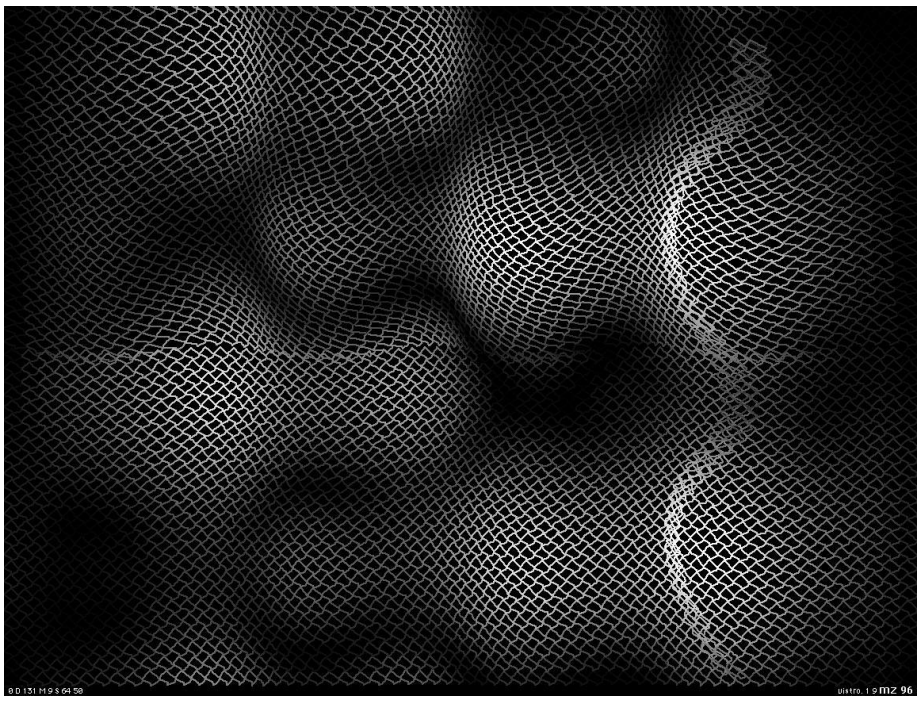

Figura 14

poi associando all'intensità della deformazione [cioè alla distanza tra il punto di partenza (x,y) al punto di arrivo (x1,y1)] la scelta del colore e di altri attributi locali.

Seguono due esempi, in fig. 15 in bianco e nero "Abbraccio morbido", anno 2000

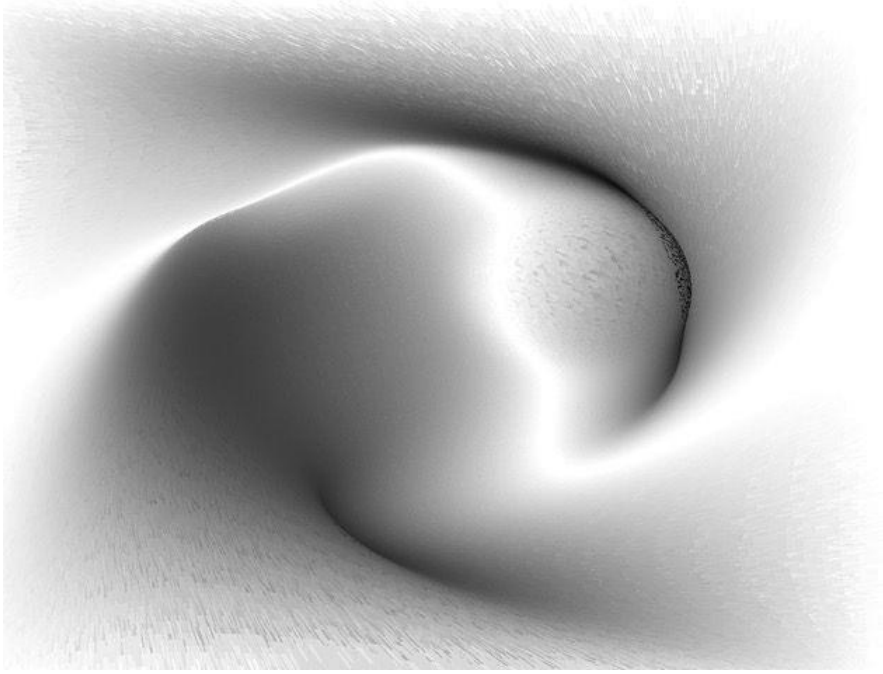

Figura 15

e, in Fig. 16, con l'uso di palette di colore, "Child's Dreams Island", 1999.

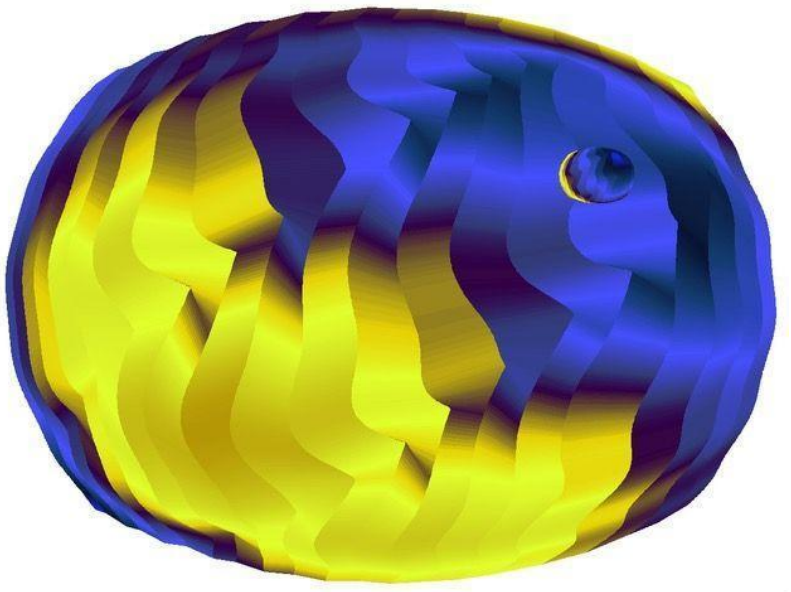

Figura 16

In questa come in altre immagini è coinvolgente la dolcezza dei colori, la luminosità. Questo è il periodo in cui anche un impulso alla ricerca di nuove forme grafico-artistiche si ha con “L’Imaginaire Scientifique” esposizione del 1986 alla Geode di Parigi di immagini realizzate in ambito scientifico artistico. Tra gli autori c’è anche Annamaria Carminelli con “Modelage Fractal” e le relative immagini. Ancora in questo ambito Carminelli presenta “Animazione delle Piccola Musica Notturna di Mozart” in Spazio4 (Fiera di Trieste 1987) e nei convegni dell’Eurographics International (anni ’90).

Tornando a Hmeljak invece spazialità e dinamica si colgono nelle realizzazioni del periodo più recente. Osservandole si "vede" il movimento e si intuisce come dalla figura ne possano scaturire altre dando vita ad una sequenza animata, ad uno dei video realizzati in particolari occasioni.

Negli anni intorno al 2010 Hmeljak si rivolge anche all' ambiente usando sovrapposizione di immagini astratte digitali a fotografie, come nell'esempio di Fig. 17

Figura 17
"Bora", 2015

Negli ultimi anni (2019, 2023) Hmeljak lavora sugli spazi pseudo-3D, con definizione di nuove trasformazioni "2D->2D", come appare nelle due immagini seguenti, in Fig. 18 “Stretta” del 2019 e in Fig. 19 “volo”, 2023.

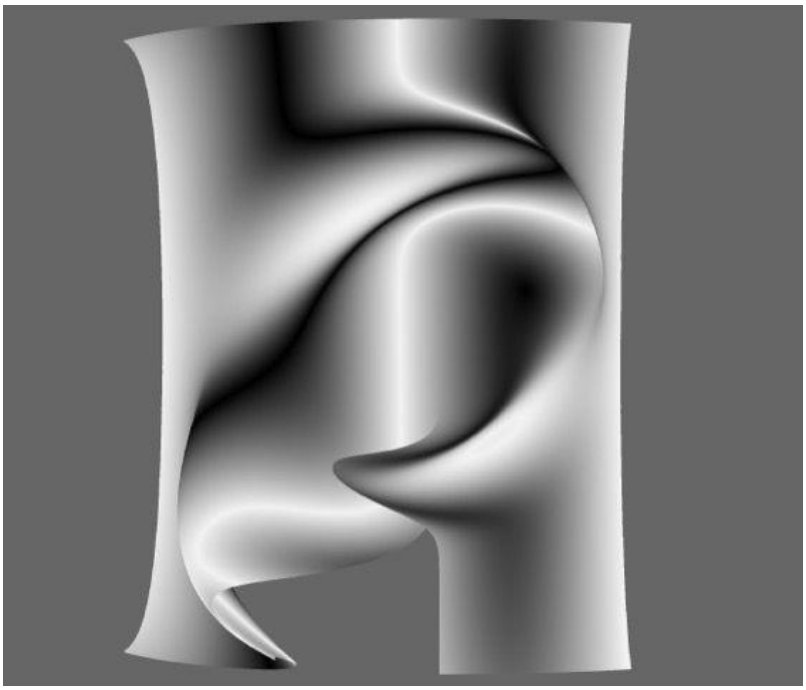

Figura 18

Figura 19

…e questo è il “presente” di Hmeljak

4. Il presente (... in generale)

Nel presente c'è di tutto: cinema, animazioni, realtà virtuali, robotica, similitudini, metafore, immaginari volatili, ambienti sognati ... e tanto altro.

Pensieri che diventano immagini e viceversa. Questo sta alla base del video (2023) citato in Biblio [11] dove tra gli autori appare anche Riccardo Lazzeri che è il "suggeritore" della musica di Bacharach.

Una rappresentazione che non vuole simulare la realtà, ma che vuole andare al di là del reale, per aprire scenari di fuga. Almeno questo era nel pensiero degli autori: far danzare le Gocce di Pioggia tenendo sempre presente la fantasia inesauribile di Claudio Galmonte. La danza non travalica i limiti stabiliti, ma si esprime in tutti i suoi gradi di libertà concessi, per presentare modelli alternativi a quelli usuali. Difficile condensare un video in un paio di immagini che comunque seguono: Fig. 20 e Fig. 21

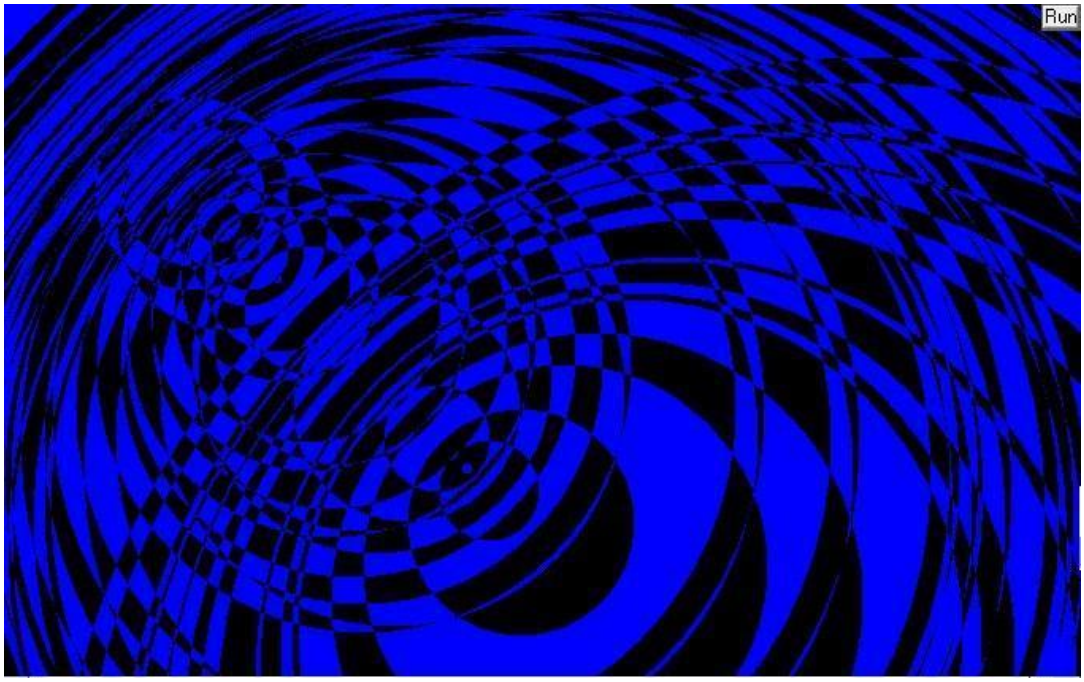

Figura 20

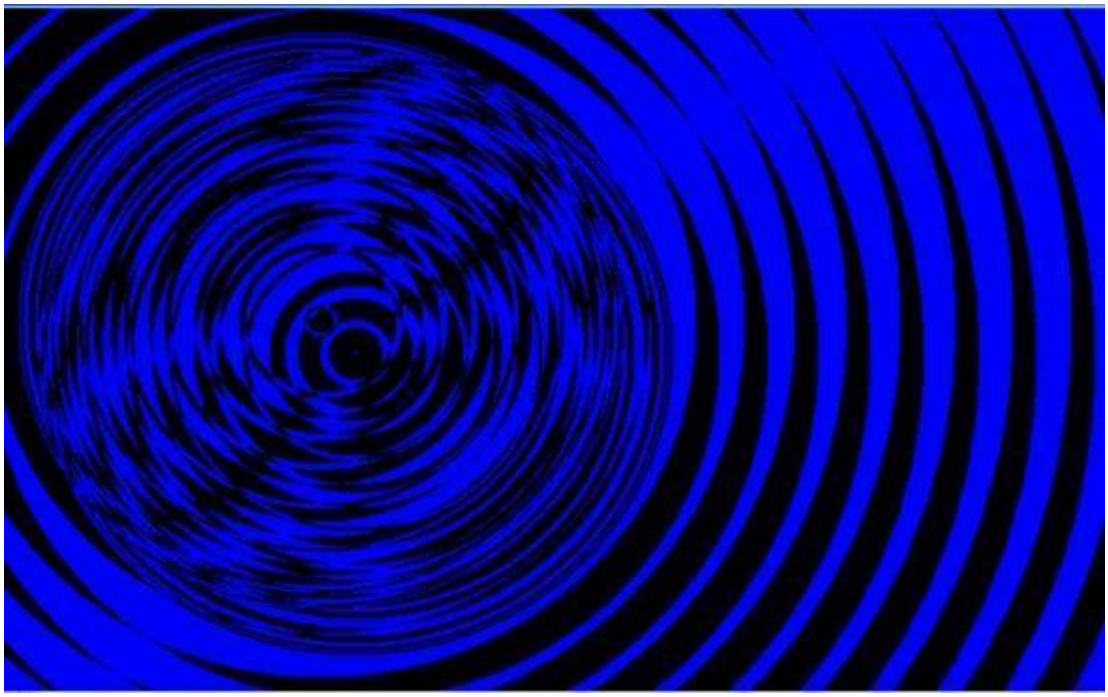

Figura 21

Riguardo alle realizzazioni attuali, senza volerne fare un elenco o comunque una esposizione esaustiva, sono per la maggior parte basate sui canoni dell'intelligenza artificiale (sintetizzata in A.I.) che si pone adesso come si poneva il computer negli ultimi anni del secolo scorso. Così come il computer era diventato un nuovo strumento per la pittura (per così dire un nuovo "pennello") anche l'A.I. lo sta diventando. Lo si nota soprattutto nei videogiochi dove viene usata non solo per costruire comportamenti dei giocatori in sintonia o in antitesi col gioco, ma anche ambienti e mutazioni di questi a livello fantascientifico. Si pensi alle play station che simulano situazioni in luoghi extra terrestri, extra galattici. Applicazioni per bambini "super dotati"? No. Appaiono piuttosto e semplicemente un indice del cambiamento di attività e atteggiamenti umani.

Cambiamento che si ritrova nella robotica dove un robot può essere addestrato anche a disegnare, diventando lui il nuovo pennello [19]. Come sempre c'è l'Umano che detta le regole comportamentali. Non solo con algoritmi, ma con dati, ai quali il motore di ricerca, che guida il robot, attinge e usa per realizzare il suo operato. I dati sono anche dinamici, ossia quelli colti dalla realtà che circonda il robot nell'immediatezza della situazione.

Questo appare nelle realizzazioni di Paolo Gallina [19], docente di "Meccanica applicata alle macchine e robotica" all' Università di Trieste che ha utilizzato un robot con modalità diverse. Per esempio per dipingere sotto la guida di una persona che interagisce con il robot tramite i movimenti degli occhi (misurati da un apposito eye-tracker) e tradotti simultaneamente dal robot stesso in pennellate su un foglio. Oppure costruendo il suo Baskerbot (robot pittore di strada) che si

inserisce nel filone dell'arte "algoritmica" essendo in grado di produrre dipinti in maniera autonoma.

A proposito dei robot, in ambito europeo emerge l'artista austriaco Alec Klessing col suo progetto "Long Distance Art" ed il risultato ottenuto. Mentre lui dipingeva un quadro a Vienna, due bracci robotici (modello ABB IRB 4600 alti 2,8 metri e del peso di 435KG ciascuno) ricalcavano i movimenti della sua penna tramite un sensore ad infrarossi e riproducevano la stessa opera contemporaneamente a Londra ed a Berlino.

Ancora un accenno a Harold Cohen con i suoi programmi (in C e LISP) noti col nome di AARON in grado di generare opere astratte e figurative esposte anche alla Tate Gallery.

E poi sempre nel presente c'è il programma di grafica Deep Dream prodotto e pubblicato da Google che usa una rete neurale convolutiva per creare e modificare immagini ad effetti allucinogeni ed onirici.

Insomma un presente ricco di applicazioni interessanti.

5. Il futuro

Non esistendo limiti alla fantasia umana la domanda arriva spontanea: c'è un modello prevedibile per le espressioni artistiche?

Se si può rispondere avremo una visione del futuro che per ora resta nebulosa.

Comunque qualcosa di nuovo appare, sia pure in controluce, col famigerato metaverso ossia con un ambiente immaginifico in cui il mondo virtuale si fa più intrigante coinvolgendo il mondo reale e viceversa, sempre sotto la "tutela" dell'Intelligenza Artificiale.

E si può concludere con una delle celebri frasi di Albert Einstein "La logica vi porterà da A a B. L'immaginazione vi porterà dappertutto".

BIBLIOGRAFIA

[1] Mara Bozzi Zadro, AnnaMaria Gregori Carminelli: "Rappresentazione conforme del geoide sull' ellissoide internazionale", Bollettino di Geodesia e Scienze Affini, Istituto Geografico Militare, Anno XXV, N. 1, 1966.

[2] A.M.Carminelli-Gregori, C.Galmonte: "Rain Drops 1968-1984", Computer Images di M. Salvemini, ed. Jackson.

[3] A.M.Carminelli-Gregori: "Eidomatica per riflettere", Cultura e Scuola, Istituto della Enciclopedia Italiana fondata da G.Treccani, N.o 100, 1986.

[4] A.M.Carminelli-Gregori: "Fenomeni in evoluzione: immagini e modelli", Cultura e Scuola, Istituto della Enciclopedia Italiana fondata da G.Treccani, N.o 110, 1989.

[5] A.M.Carminelli-Gregori: “Ricottura e tempera per immagini”, Cultura e Scuola 116, Istituto dell’Enciclopedia Italiana, fondata da Giovanni Treccani, 1990.

[6] A.M. Carminelli-Gregori, Paolo Agati – “Dalla compressione alla composizione di immagini”, Cultura e Scuola 123, Istituto dell’Enciclopedia Italiana, fondata da Giovanni Treccani, 1992

[7] A.M.Carminelli-Gregori, E. Mumolo, I.Bonat: "An Optimisation Algorithm for Fractal Encoding of Graytone 2D Images", in Proceedings of EUSIPCO'94 - 7a European Signal Processing Conference.

[8] A.M.Carminelli-Gregori, E. Mumolo, I.Bonat: "Fractal Encoding of Digital Images with an Optimisation Algorithm", in Proceedings of theTwelfth IASTED International Conference, May 1994, Annecy, France.

[9] A.M.Carminelli-Gregori, R.Cobalti,, G.Vercelli: «Intelligent Web Agents for Information Retrieval and Classification» , Proc. of Irtl. Conf. on Practical Application for Intelligent and Mult-Agents, Technologics, PAAM-99, 1999.

[10] A.M.Carminelli-Gregori, Massimiliano Nolich: “Malware Classification Using Evolutionary Computation”, WIVACE 2007 - Workshop di Vita Artificiale e Computazione Evolutiva, Settembre 2007 - Sampieri (Ragusa).

[11] A.M.Carminelli-Gregori, A. Marassi, P. Delise, R. Lazzeri “Animazione di Rain Drops” Video. https://youtu.be/pa-zVd2NPQ4

[12] A. Moro: “Trattamento automatico delle informazioni.” 6 edizione Franco Angeli/Manuali professionali 1977

[13] F. Dietrich: “Visual Intelligence: The First Decade of Computer Art (1965-1975)” LEONARDO, Vol. 19, No. 2, pp. 159-69, 1986 Copyright 1985 IEEE. Reprinted with permission, from IEEEComputer Graphics and Applications, Vol. 5, No. 7, pp. 34-35, July 1985

[14] Edward E. Zajec | Database of Digital Art, su dada.compart-bremen.de. URL consultato il 22 gennaio 2023

[15] G. O’Regan: A Brief History of Computing. Springer London 2012

[16] S. Harrington: Computer Graphics. Corso di programmazione. McGraw-Hill Companies 1987

[17] T. Dreher: History of Computer Art.
http://iasl.unimuenchen.de/links/GCA_Indexe.html 2015

[18] H.S.Stone: "Introduction to Computer Organization and Data Structures" McGraw-Hill 1971

[19] //www.units.it/ news/un-robot-per-dipingere-con gli occhi-tre-nuove-creazioni-di-paolo-gallina-tra-arte-e-tecnologia

BIOGRAFIA

Annamaria Carminelli. Laureata in fisica pura nel 1962 all'Università di Trieste, ha avuto una carriera variegata. Borsista IBM, aspirante ricercatore CNR, tecnico laureato e nel 1985 dopo un Concorso Nazionale Professore associato. Pensionata dal 2006, ma con incarichi di insegnamento all'Università ancora per tre anni.

Email: carminelli.annamaria@libero.it

Alessandro Marassi. Laureato in Ingegneria Elettronica nel 1985 all'Università di Trieste e in Scienze dell'Informazione nel 2009 all'Università di Udine. Tecnico direttivo in ACEGA Trieste dal 1986 al 1989; progettista software in Meteor CAE SpA dal 1989 al 1994; esperto ingegnere presso Comune di Trieste dal 1994 al 2000; funzionario di elaborazione dati e tecnologo in INAF – Osservatorio Astronomico di Trieste dal 2000.

Email: alessandro.marassi@inaf.it

Piero Delise Dal 1984 al 1989 ha tenuto corsi di informatica per adulti presso l'ENAIP di Trieste. Laureato in Ingegneria Elettronica nel 1992 all'Università di Trieste. Nel biennio 1992-93 ha tenuto esercitazioni e lezioni per diploma di informatica presso l'Università di Trieste. Dal 1989 al 2017 ha lavorato nell' Information Technology presso la compagnia di assicurazioni RAS, poi Allianz Spa. In pensione dal 2018.

Email: piero.delise@gmail.com