ANNO ACCADEMICO 2024/2025

Etica, Società e Privacy

Etica

Altair's Notes

UNIVERSITÀ DI TORINO

DIPARTIMENTO DI INFORMATICA

Premessa

Licenza

Formato utilizzato

Box di "Concetto sbagliato":

Concetto sbagliato 0.1: Testo del concetto sbagliato

Testo contente il concetto giusto.

Box di "Corollario":

Corollario 0.0.1 Nome del corollario

Testo del corollario. Per corollario si intende una definizione minore, legata a un'altra definizione.

Box di "Definizione":

Definizione 0.0.1: Nome delle definizione

Testo della definizione.

Box di "Domanda":

Domanda 0.1

Testo della domanda. Le domande sono spesso utilizzate per far riflettere sulle definizioni o sui concetti.

Box di "Esempio":

Esempio 0.0.1 (Nome dell'esempio)

Testo dell'esempio. Gli esempi sono tratti dalle slides del corso.

Box di "Note":

Note:-

Testo della nota. Le note sono spesso utilizzate per chiarire concetti o per dare informazioni aggiuntive.

Box di "Osservazioni":

Osservazioni 0.0.1

Testo delle osservazioni. Le osservazioni sono spesso utilizzate per chiarire concetti o per dare informazioni aggiuntive. A differenza delle note le osservazioni sono più specifiche.

1 Etica - Introduzione

1.1 Il Corso in Breve...

Il corso ha un aspetto prettamente *umanistico* e *interdisciplinare*. Si andranno ad affrontare diverse prospettive che impattano su diverse dimensioni della propria vita:

- Innovazione (brevetti).
- Economia (monopoli).
- Politica.
- Giuridica.
- Tecnologica.
- Mercato dell'attenzione.
- Retorica.

Definizione 1.1.1: Retorica

L'arte del dialogare per convincere le persone a fare quello che si vuole.

Note:-

Nei giornali, nella politica, nella pubblicità[a], etc.

[a] Per questo esistono gli AdsBlocker.

Esempio 1.1.1 (Retorica)

- La cimice asiatica sostituirà la cimice verde nei nostri prati.
- I robot sostituiranno gli esseri umani nei posti di lavoro.

Due frasi grammaticalmente identiche, ma non sono la stessa cosa: la prima ha un significato letterale, la seconda no. Perché la seconda non intende che i robot andranno a prendere e buttare via gli esseri umani dal posto di lavoro, ma che i padroni sostituiranno i lavoratori con dei robot.
Questa è una frase strumentale, per nascondere il ruolo dell'essere umano.

Note:-

Qualsiasi affermazione di una società è puramente strumentale. Punta a manipolare le persone per ottenere un ritorno economico.

1.1.1 Perché Questo Corso?

Le tre missioni dell'università:

- Didattica.
- Ricerca.
- Terza missione: esportare le conoscenze tecnologiche alle aziende[1] e rendere coscenti le persone di quello di cui si occupa l'università (opportunità e rischi).

I messaggi del corso:

- La tecnologia e l'impatto che ha nella società sono costruzioni sociali, senza nessuna inevitabilità.
- La tecnologia non può risolvere i problemi che crea (e. g. i bias)[2].
- L'AI non è solo una tecnologia.
- L'essere umano ha un ruolo importante nel determinare l'impatto delle tecnologie digitali, sia nella propria professione sia come cittadini.

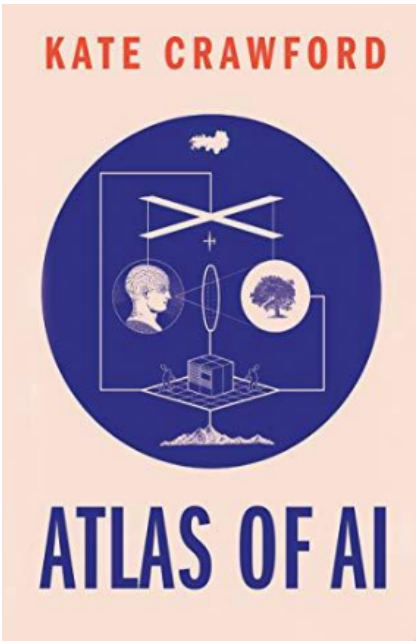

Figure 1.1: Uno dei libri possibili per l'esame.

Che cosa vuol dire che l'AI non è solo una tecnologia?

- In "Atlas of AI", Kate Crawford, propone un esperimento.
- Prendete Google Search (schifezza) e scrivete ARTIFICIAL INTELLIGENCE.
- I risultati sono immagini, principalmente di cervelli, con sfondo blu.

[1] Che spreco...
[2] Fuck Silicon Valley and fuck Marc Andreessen.

- Cos'è che non viene mostrato?
  - L'inquinamento delle miniere di litio (con cui sono fatte le batterie).
  - Le proteste delle popolazioni la cui vita è stata distrutta dalle miniere di litio.
  - Le terre rare (metalli che servono per i microprocessori) che vengono estratte in paesi sottosviluppati, zone di guerra, da minori (spesso in cina).
  - Le condizioni di schiavitù in cui lavorano centinaia di persone allo sviluppo dei dispositivi elettronici.
  - Gli enormi datacenter alimentati da fonti non rinnovabili.
  - Per allenare i sistemi di AI vengono usate foto su cui non si hanno i diritti.
  - The cleaners: la gente, sottopagata, che si occupa di pulire tutto lo schifo umano (decapitazioni, pedopornografia, violenza su donne, uomini, bambini, animali, etc.) in modo che i modelli di ML e i Social Media siano puliti.
  - L'ossessione per la produttività a cui sono soggetti i lavoratori (dal Taylorismo in avanti).
  - In alcuni paesi il riconoscimento facciale è usato per discriminare minoranze.
  - Le varie challenge su TikTok che procurano morti.
  - Molly Russsel: ragazzina di 14 anni che si è suicidata dopo che Instagram e YouTube, con algoritmi AI di personalizzazione, le hanno mostrato più di 2000 contenuti di istigazione al suicidio e all'autolesionismo.
  - FaceBook è stato accusato di genocidio in Etiopia, Bangladesh e Myanmar per persecuzioni di minoranze.
  - Assalto a Capitol Hill da parte dei sostenitori di Trump, anche a causa dei Social Media.

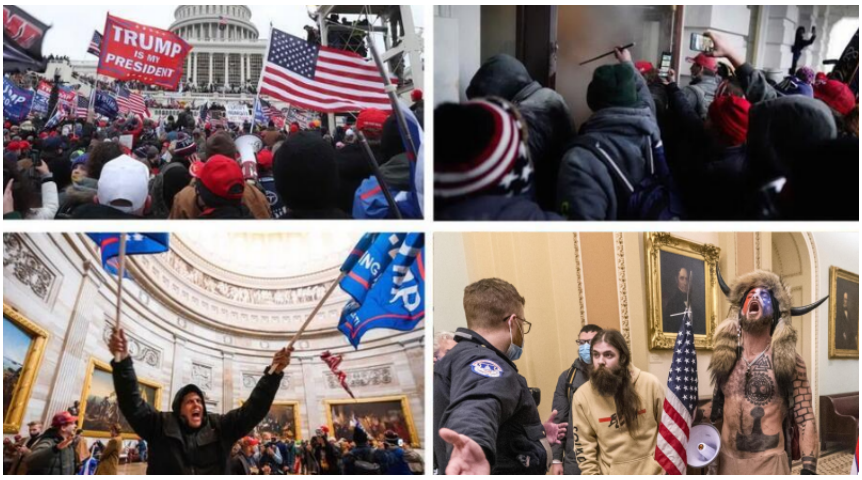

Figure 1.2: Assalto a Capitol Hill (6 Gennaio 2021).

  - QAnon nasce su 4chan: teoria secondo la quale il "deep state" sia composto da pedofili che rubano i bambini per estrarne il loro sangue e produrre una droga contro l'invecchiamento.

Figure 1.3: Elon moment.

  - "You have blood on your hands" disse il senatore Graham a Mark Zuckenberg, "You have a product that's killing people.".

Social Network vs. Social Media:

- Un Social Network serve per mettere in connessione le persone.
- Un Social Media serve per distribuire contenuti.
- Diversi Social Network sono diventati Social Media (e. g. FaceBook).

Domanda 1.1

Perché si è andati in questa direzione?

Risposta: ultimamente si mostrano i contenuti più virali, ritenuti interessanti al singolo, per creare dipendenza e aumentare gli introiti dovuti alla pubblicità. "If you're not paying for it, you're not the customer; you're the product being sold".

1.2 La Costruzione della Realtà Sociale

"La costruzione della realtà sociale" è un libro scritto da John Searle in cui sostiene che la realtà sociale sia frutto di una costruzione[3]. Searle racconta il seguente aneddoto: io, John Searle, cittadino americano, mi trovo in un bistrot lungo la Senna che mi bevo una birra[4]. Ma com'è possibile che ciò sia avvenuto? Sono un cittadino americano, sono salito su un aereo, sono arrivato a Parigi, sono entrato nel bistrot che vende birra, parlo con il cameriere e gli chiedo una birra, ovviamente pagando. Sembra tutto semplice, ma dietro a ogni punto si nasconde una rete concettuale che le persone danno per scontata e accettano. Perché mi fanno salire sull'aereo? Perché sono un cittadino americano e ai cittadini americani è permesso fare turismo in europa. E come provo che sono cittadino americano? Esibisco il passaporto. Come possibile che il bistrot venda birra? Ci vuole la licenza. E com'è possibile che il cameriere sia lì, mi porti la birra e si aspetti di essere pagato? E com'è che vengono accettati pezzi di metallo il cui valore corrispettivo è quello della birra consumata? Dietro tutto questo c'è una *realtà sociale* fatta di norme, ma anche di *consuetudini* che rendono possibili le interazioni e che collegano *fatti bruti* (il passaporto, le monete, etc.) con un *livello astratto* (regole costitutive).

Al giorno d'oggi la tecnologia e, come esempio estremo, il metaverso, va a influenzare e modificare la realtà sociale. Inoltre la realtà sociale è costituita da parole e attualmente si hanno i LLM che "parlano" e possono partecipare alla realtà sociale.

1.2.1 Come l'AI Influenza la Realtà Sociale

Nel 1989, durante la rivolta studentesca di piazza Tiananment, un ragazzo si mise davanti a una colonna di carriarmati e venne portato via. Non sapendo il suo nome viene soprannominato "Tank Man". Poco fa è diventata famosa una "foto" generata dall'AI di Tank Man, quest'immagine fu, per un po' di tempo, indicizzata da Google. Google la rimosse, ma dato che è stato fatto un articolo di debunking è attualmente indicizzata (con il link al relativo articolo in cui viene spiegato in dettaglio il fatto che l'immagine è fake).

Note:-

Oltre a questo ci sono stati altri casi, uno recente è Trump Gaza: come Trump si reimmagina la striscia di Gaza come resort. È questo contenuto è stato condiviso dallo stesso Trump sulla sua piattaforma Truth.

Domanda 1.2

Ma la disinformazione c'è sempre stata. Che cambia tra Tank Man, Trump Gaza e altre fake news non AI-Generated?

Risposta: sebbene la disinformazione ci sia sempre stata, al giorno d'oggi è molto più facile e veloce grazie all'AI. Una volta la disinformazione era solo su giornali e poteva portare conseguenze ai giornalisti, ora le regole per il materiale AI-Generated sono molto vaghe.

[3] Curioso visto quello di cui viene accusato, anche il consenso è una costruzione sociale?
[4] Average americano

Figure 1.4: Immagine AI-Generated di Tank Man.

Domanda 1.3

I social media sono una piazza pubblica?

Risposta: è un tema molto complesso. Da una parte in una piazza pubblica se si urla qualcosa finisce lì, mentre sui social media *potenzialmente tutti* possono accedere e vedere/sentire. Quanto conta l'algoritmo di personalizzazione in tutto questo? Per esempio Gonzales vs. Google in cui i genitori di una ragazzina morta negli attentati terroristici in francia hanno fatto causa a Google perché i terroristi si erano radicalizzati con contenuti di YouTube offerti loro dall'algoritmo di personalizzazione. Tuttavia perserò la causa.

1.2.2 L'AI Rappresenta Effettivamente il Mondo?

Se si utilizza un generatore di immagine AI per creare foto di giudici, ingegneri, medici, politici e CEO si vedrà una grande maggioranza di maschi bianchi. Al contrario se gli si chiedono immagni di lavapiatti, lavoratori di fastfood e bidelli si troveranno persone di pelle scura. Se gli si chiedono insegnanti, cassieri, lavoratori dei servizi sociali e domestici si troveranno più donne. Ma è così la reltà? Sì e no, in realtà le proporzioni sono più bilanciate di quelle proposte dall'AI generativa che funge da "cassa di risonanza".

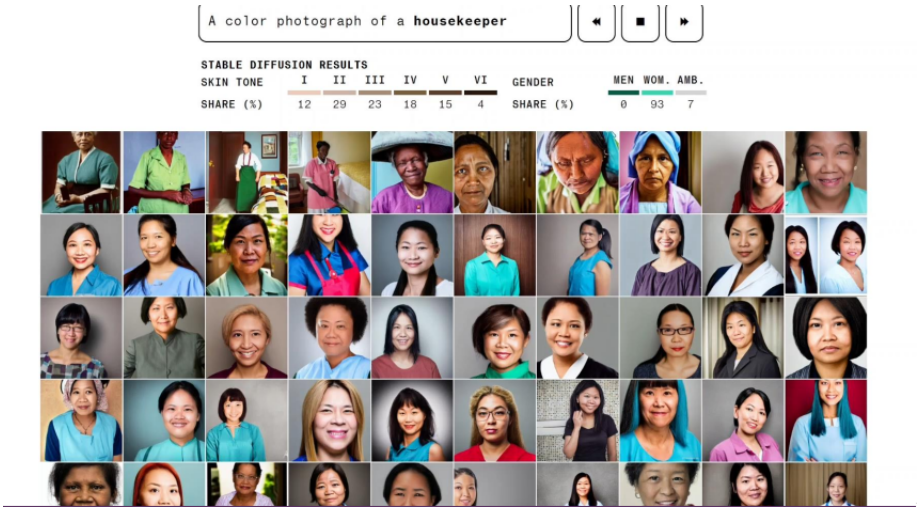

Figure 1.5: Esempio.

Domanda 1.4

C'è una soluzione al problema?

La mossa di Google: con Gemini ha utilizzato un dataset "politically correct", però quando gli si chiedeva di rappresentare un soldato tedesco nella WWII sputava fuori persone di colore, asiatiche e donne. Con l'immagine di un papa uscivano papi neri (tecnicamente possibili) e papesse (beh, è una carta dei tarocchi). I padri fondatori, in un epoca di fervente schiavismo, diventavano di colore. Infine i fondatori di Google stessa diventavano asiatici.

La soluzione di Google: semplice.

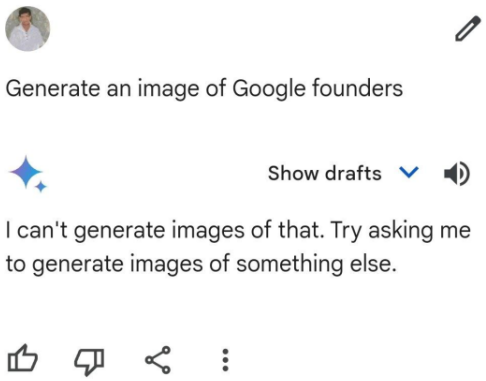

Figure 1.6: Non posso onii-chan :3

Osservazioni 1.2.1

Brevi osservazioni sulla censura nei vari modelli:

- Deepseek: se parli di Taiwan o di quello che è successo a piazza Tiananment.
- ChatGPT: se parli di Noam Chomsky (attivista anarchico e sostenitore di Pol Pot).
- Grook: se parli di fake news in relazioni a Trump e Musk.

Note:-

Queste soluzioni sono solo tecniche, ma il problema è politico.

1.2.3 I Pericoli dei Language Models

Un anno prima dell'uscita di ChatGPT furono licenziate da Google Margaret Mitchell e TImnit Gebru. Esse misero in guardia sul pericolo dei pappagalli stocastici: i language models non sanno quello che dicono, ma semplicemente ripetono delle parole che non comprendono.

Unfathomable Training Data:

- La grandezza del dataset non garantisce la diversità: gran parte del web è scritto in inglese, da uomini bianchi.
- Poiché passano mesi o anni dall'inizio dell'apprendimento il dataset è già vecchio. Per esempio la prima release di ChatGPT non era a conoscenza della guerra in Ucraina. Anche dal punto di vista sociale sono lo specchio di una società più vecchia.
- Encoding Bias.

Le relazioni parasociali: questi language models possono influenzare gli esseri umani nel dialogo reciproco. Si attribuisce a questi sistemi più di quello che realmente c'è. Nel 1966 fu creata Eliza, uno dei primi chatbots, che interpretava uno psicoterapeuta che faceva domande con un particolare stile. Un aneddoto famoso riguarda una segretaria che si convinse di parlare realmente con uno psicoterapeuta (pur sapendo che cos'era Eliza).

Figure 1.7: Eliza.

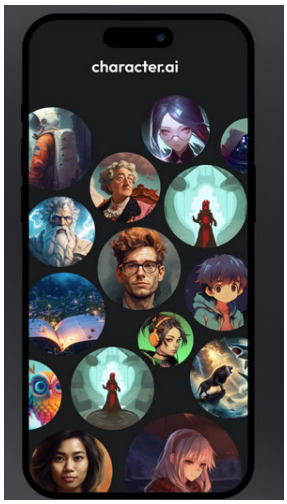

Figure 1.8: La moderna versione di Eliza.

Definizione 1.2.1: Allucinazioni

Se il chatbot non sa qualcosa se lo inventa con un sistema probabilisitico. Se le probabilità sono labili viene restituito in output qualcosa che sembra coerente, ma non ha senso (allucinazioni).

La missione di OpenAI: sviluppare la Artificial General Intelligence (AGI) per sostituire l'essere umano in qualsiasi compito che abbia *valore economico*.

Il problema con la gente contraffatta: da quando hanno inventato il denaro chiunque lo contraffae è punito con le pene più severe. Il problema si riconduce alla realtà sociale di Searle e che contraffarre il denaro è un crimine perché fa venire meno la fiducia nel denaro, minando la società stessa. Se vale l'analogia tra denaro contraffatto e umano contraffatto (chatbot che si spaccia per un essere umano) allora si è di fronte al più grande crimine della storia, interagendo con persone finte. Da questo deriva anche il problema del deepfake pornografico AI-Generated.

Altri problemi:

- Le guide generate dall'AI: sempre più libri guida vengono generati dall'intelligenza artificiale. Nel migliore dei casi sono libri turistici inaccurati, ma può capitare di imbattersi in libri più pericolosi (e.g. guide per distinguere i funghi velenosi). Questo fenomeno ha dato origine al termine *sloop internet* per indicare i rifiuti prodotti dall'AI generativa.
- La disinformazione causata dall'AI è personalizzata all'utente. Ciò può fungere da gigantesca echo chamber per convincere determinate persone o per dissuaderne altre.
- Gli scam online: gente a cui telefonano dei "parenti" la cui voce è replicata dall'AI (e.g. Jennifer DeStefano).

2
Ancora sull'AI

2.1 AI or not AI This is the Dilemma

2.1.1 Il Watermark

Definizione 2.1.1: Watermark

Il watermark consiste in cambiamenti non visibili in alcuni pixel chiavi che possono essere individuati dalla macchina.

Figure 2.1: Immagine con watermark.

Note:-

Uno dei provvedimenti dell'amministrazione Biden per "difendersi" da watermark è stato quello di dire: inseriamo del watermark che indichi se un'immagine è AI-generated.

Domanda 2.1

Questa è una soluzione ragionevole?

- Se è possibile introdurlo allora esisterà un AI in grado di rimuoverlo.

- Questa proposta garantisce Microsoft, openAI, Google che se ne lavano le mani. Però le aziende non americane non hanno alcun obbligo di applicare il watermark.
- Inoltre il fatto che una foto abbia il watermark, in un'epoca di cospirazionismo, non significa nulla[1].

2.1.2 Il Copyright

Tutti i vari modelli di AI generativa sono stati addestrati su tutti i dati del mondo, principalmente il web (anche su materiale soggetto a copyright). Per esempio alcuni siti come z-library o il web archive[2] in cui sono presenti contenuti piratati. I vari LLM hanno probabilmente assorbito anche quei contenuti. Ovviamente quella merda che è Google ha detto: eh, ma se è lì sul web allora lo posso prendere, se non lo volevano non lo dovevano pubblicare (a parte che è l'esatto opposto del meccanismo di copyright, ma forse non si rendono conto che è la stessa argomentazione che utilizzano gli stupratori nei confronti della loro vittima). Oppure openAI che chiede alle persone di indicare quali pagine non vogliono sia preso dal crawler. Tuttavia questo non funziona, per esempio con i siti di mirroring.

Artificial intelligence (AI)

'I didn't give permission': Do AI's backers care about data law breaches?

Regulators around world are cracking down on content being hoovered up by ChatGPT, Stable Diffusion and others

A demonstrator holding a 'No AI' placard. In Italy, ChatGPT has been banned after the regulator said there appeared to be no legal basis to justify the collection and storage of personal data. Photograph: Wachiwit/Alamy

Figure 2.2: Protestante con cartello.

[1] God, how much I hate humanity.

[2] Entrambi consigliatissimi, da visitare almeno una volta nella vita, sono la nuova biblioteca di Alessandria.