Etica e intelligenza artificiale. Intervista a Federico Cabitza

di Giacomo Bottos

L'intelligenza artificiale può concorrere alla creazione di nuove potenzialità per l'umano oppure a un ecosistema nel quale l'umanità resti ai margini. Nell'intervista si affrontano questi scenari e il ruolo dell'etica in tale chiave, anche a partire dal recente volume Intelligenza Artificiale. L'uso delle nuove macchine, scritto insieme a Luciano Floridi e edito da Bompiani. Federico Cabitza è professore associato di interazione uomo-macchina all'Università degli Studi di Milano-Bicocca, dove è responsabile del laboratorio di Modelli di incertezza per decisioni e interazioni e direttore del nodo locale del laboratorio nazionale CINI Informatica e società.

In che senso e con quali limiti il concetto di etica può essere applicato alle macchine e agli artefatti?
Federico Cabitza: Le macchine e gli artefatti tecnologici sono semplicemente estensioni della nostra capacità di agire, anzi parte di essa e delle nostre capacità di trasformare e governare l'ambiente che ci circonda per renderlo più favorevole e adatto alle nostre esigenze. Anche le macchine digitali, che ci sembrano operare con un certo grado di indipendenza e autonomia, sono comunque programmate da qualcuno o inserite da qualcuno in un contesto sociale e lavorativo e in un ambiente con determinate finalità, che spesso non riguardano tanto l'utente di quel contesto quanto invece riguardano quell'attore che opera nello sfondo e tendiamo a non notare. Quindi il tema etico relativo alle macchine è semplicemente quello che riguarda il valore (ad esempio giusto/sbagliato, benefico/dannoso) delle azioni che esse rendono possibili, e il valore relativo alle conseguenze delle azioni che esse compiono, autonomamente, o che permettono agli esseri umani di compiere più agevolmente o più efficacemente.

Che cosa si intende per algoretica?
Federico Cabitza: Diverse cose, ma principalmente la riflessione sull'uso etico delle macchine computazionali (robot o software) in contesti dove la loro azione e ciò che producono - ad esempio contenuti, informazioni, suggerimenti e, perfino, decisioni - possono avere un effetto giuridicamente rilevante su qualcuno o ledere in qualche modo i diritti fondamentali di un essere umano. Con il termine algoretica alcuni vogliono anche indicare un programma intellettuale - sia filosofico, si veda il lavoro di Benanti, che ingegneristico, si veda Marcus - che mira ad inserire negli algoritmi capacità di giudizio valoriale e morale in modo tale che le macchine possano essere autonome e, si spera, più oggettive e giuste degli esseri umani, in valutazioni di carattere etico; su questo piano del discorso, però, io devo esprimere una posizione decisamente scettica se non proprio contraria, perché ritengo che non si debbano dare ulteriori alibi agli esseri umani rendendo i loro strumenti in grado di fare a meno del loro controllo e della loro vigilanza, oppure in grado di 'calcolare' giudizi di valore rispetto ad una morale di riferimento - anche assumendo, ma non concedendo, che fosse possibile esprimere tale morale in qualche modo non ambiguo e computabile. Il rischio di essere eterodiretti o influenzati eccessivamente da questi strumenti sarebbe secondo solo al rischio di disabituarci alle valutazioni di carattere etico e alla presa in carico delle responsabilità delle proprie azioni (o non azioni).

Che rapporto reciproco esiste tra approccio ethical by design e approccio ethical in design?
Federico Cabitza: Per semplificare, 'etico per specifica di progettazione' è una caratteristica di un

sistema il cui comportamento progettato e programmato è tale da non ledere i diritti fondamentali delle persone (affected people); 'etico nel design' è un'espressione che ho proposto in contrasto e parziale polemica con una certa deriva con cui viene usata l'espressione precedente - quella che ritengo porti ad assolvere il progettista o comunque l'attore umano perché all'obbligo di comportarsi eticamente o meno sono chiamati appunto i sistemi tecnici e non gli esseri umani. L'approccio ethical in design è anche una provocazione per auspicare un atteggiamento (umano) più consapevole da parte di chi richiede, progetta, programma, installa, adotta e alimenta sistemi computazionali dall'effetto giuridicamente rilevante, affinché questi attori considerino attentamente tutte le implicazioni e le conseguenze delle loro scelte progettuali e interventi prima di intraprenderle e, non lo si dimentichi, dopo averle intraprese. Si consideri anche il concetto, collegato, di tecno-vigilanza, definita come scienza basata sulle evidenze relative alla raccolta, al rilevamento, alla valutazione, al monitoraggio e alla prevenzione delle conseguenze negative legate all'uso delle nuove tecnologie digitali a supporto della cognizione umana basate su AI/ML.

Quali responsabilità dovrebbero essere in capo a chi progetta, controlla e usa i sistemi di intelligenza artificiale?
Federico Cabitza: Dipende dal dominio applicativo e dai rischi che possono essere ricondotti all'uso dei sistemi AI in quei contesti. Il regolamento che le istituzioni dell'Unione Europea stanno definendo con i rappresentanti delle comunità di portatori di interessi e dei parlamenti nazionali adotta convintamente un approccio basato sulla valutazione dei rischi e prevede per tutti coloro che progettano o usano sistemi considerati ad alto rischio una serie di obblighi di verifica di conformità e vigilanza, affinché siano garantite tutte le tutele e i controlli necessari secondo standard riconosciuti (molti dei quali in via di definizione proprio in questi mesi, si vedano le ISO/IEC TR 5469, ISO/IEC 4213, ISO/IEC JTC 1/SC 42). Il tema aperto è comunque quello che purtroppo il diritto si evolve sempre un po' più lentamente rispetto agli ambiti di azione che possono conquistare le tecnologie digitali con il continuo progresso delle tecnologie e l'accesso al mercato di nuove soluzioni - al momento, senza alcun obbligo di certificazione o validazione e in un contesto dove pare a molti che 'frenare' l'innovazione sia inopportuno o autolesionistico.

Quali sono i limiti della teoria della singolarità e perché esercita un fascino così significativo in ambito tecnologico?
Federico Cabitza: La teoria della singolarità teorizza (o dovrei dire immagina) un momento, nella storia dell'umanità, in cui saremo stati in grado di costruire un sistema che sarà non solo più intelligente di chi l'ha sviluppato, ma anche di ogni singola persona e di tutta l'umanità nel suo complesso. Gli atteggiamenti riguardo cosa possa succedere allora variano molto, come presumibile. C'è chi ipotizza che la maggior parte dei problemi che affliggono le società umane e il pianeta che ci ospita potranno essere risolti grazie agli interventi di questa superintelligenza che, si immagina, avrà accesso ad un mondo iperconnesso di dispositivi e infrastrutture critiche e quindi a tutto ciò che può rendere la sua azione tempestiva ed efficace, quindi risolutiva. C'è anche chi ritiene, ad esempio il noto filosofo Nick Bostrom, che questa superintelligenza possa ritenere gli esseri umani parte, se non causa diretta, dei problemi che è chiamata a risolvere, e quindi adoperarsi per soluzioni drastiche che porrebbero fine alla nostra cultura, se non proprio alla nostra specie. Quindi l'idea di singolarità ripropone, e aggiorna all'epoca della tecnoscienza, le speranze di carattere messianico e i timori di natura più apocalittica, in un sincretismo che io fatico a distinguere da certa fantascienza iperottimistica o, al contrario, distopica. In realtà, io ritengo pericoloso affidarsi a speranze messianiche per vedere risolti problemi che sono invece totalmente in carico alla responsabilità degli esseri umani, e in particolare di una cerchia relativamente ristretta di essi. Solitamente i pensieri apocalittici e le speranze messianiche sono alimentati dalla perdita della

fiducia in se stessi e dal sogno che una provvidenza possa intervenire per ribaltare una situazione completamente compromessa e sostanzialmente irrimediabile: a riguardo Fabian Scheidler ha scritto pagine molto chiare. Mi sembra quindi che la singolarità si possa vedere come una specie di profezia di impotente e fatalistica resa alla tecnologia e a chi la controlla, che porta con sé il rischio di autoavverarsi e minare quella buona volontà che è invece necessaria per comprendere che è necessario invertire la rotta e impegnarsi a fondo per modificare il nostro rapporto con gli altri esseri umani - ad esempio nelle dispute e controversie internazionali -, gli altri esseri viventi, e l'ambiente in cui viviamo.

Che cosa si intende per approccio consequenzialista?
Federico Cabitza: È un approccio che valuta la bontà di un'azione dai suoi effetti, o da quelli che si può ragionevolmente prevedere siano i suoi effetti principali. Si tratta di un approccio non alternativo ma complementare a quello basato sui principi, sulle virtù o sui valori di riferimento.

Ritiene che ci sia una sottovalutazione dei rischi e dell'impatto dello sviluppo di tecnologie di intelligenza artificiale?
Federico Cabitza: Penso che vi sia una vera e propria rimozione del concetto di rischio, almeno nella pubblica opinione e in gran parte della produzione accademica e scientifica e la cosa non deve sorprendere, perché per pensare al rischio, o pianificare mettendolo al centro del proprio progetto, è necessario avere un rapporto maturo e consapevole con i concetti di probabilità, di impatto e, soprattutto, di incertezza. Il concetto di incertezza, in particolare, si articola sia in termini di ignoranza, che non è un disvalore in sé se è consapevolezza dei propri limiti - si ripensi all'insegnamento socratico -, sia in termini di inconoscibilità, che del resto è la vera molla di ogni impresa scientifica. Come detto, ritengo molto positivo che a livello politico e di regolamentazione la consapevolezza dei rischi sia invece molto alta, soprattutto nel contesto dell'Unione Europea. Però dobbiamo parlare di più ai cittadini di incertezza, esporli al dubbio dei tecnici e al valore della sperimentazione, per renderla una dimensione del discorso (di ogni discorso, sia di quelli politici che di quelli scientifici) in cui ciascuno di noi possa trovare sia stimoli di ricerca intellettuale, che di confronto con tesi diverse e anche opposte e, soprattutto, sviluppo di atteggiamenti di prudenza e umiltà, quando la posta in gioco è semplicemente troppo importante per affidarsi al miraggio della verità oggettiva.

Che cosa significa automation bias?
Federico Cabitza: È un termine che indica un bias, cioè una distorsione cognitiva sistematica, nei confronti dell'output di sistemi di intelligenza artificiale che agiscono come supporti decisionali. Se una persona che deve prendere una decisione, e che si avvale di un sistema computazionale per prendere questa decisione più efficacemente o più velocemente, si fida del sistema anche quando non dovrebbe, cioè quando questo fornisce un suggerimento o una raccomandazione erronea, e non attua tutte le strategie cognitive che adotterebbe in assenza del supporto, cioè in poche parole si fida e si affida all'intelligenza artificiale, allora si parla di automation bias. È un fenomeno tutt'altro che raro e al quale gli esseri umani possono essere detti naturalmente portati perché l'evoluzione ci ha selezionati come esseri straordinariamente capaci di capire quando possiamo trascurare certe informazioni, a favore di altre, e quindi compiere certi compiti con un minore dispendio di energie. Ciò che rende interessante questo fenomeno è che esso si innesca più frequentemente in situazioni che potremmo denotare come paradossali, cioè apparentemente inaspettate, perché il rischio di automation bias è tanto più alto quanto migliore è il supporto decisionale e minore è la probabilità che esso possa compiere un errore. Per questo, quando un errore capita - perché capita sempre - non ce lo aspettiamo e quindi possiamo fallire nel capire che dovremmo cavarcela da soli e che,

anziché adagiarci nelle scorciatoie cognitive offerteci dall'intelligenza artificiale, dovremmo invece vigilare e mettere in campo energie e attenzioni aggiuntive. Mitigare il rischio di automation bias significa progettare sistemi che siano in grado di valutare correttamente la probabilità di sbagliarsi, cioè siano adeguatamente calibrati, e che siano in grado di farsi capire da chi deve prendere le decisioni, vale a dire siano più 'spiegabili' o interpretabili.

Che cosa si intende per ethical by undesign e in quali casi tale approccio potrebbe essere utile e necessario?
Federico Cabitza: È un'espressione che ho coniato, ispirato dal lavoro di Pierce, per indicare un atteggiamento di estrema prudenza per lo sviluppo, la produzione e l'adozione di tecnologie di cui non conosciamo ancora il potenziale trasformativo, ma possiamo anche dire distruttivo, nei contesti organizzativi e sociali in cui venissero adottate su larga scala: tale atteggiamento consiste nell'astenersi dall'azione - quindi, per certi versi, ha elementi di affinità con la dottrina dell'inazione teorizzata nel taoismo oppure con il principio ippocratico del primum non nocere - e, nei casi dove reputiamo o riscontriamo che i rischi siano troppo alti, anche nel parziale smantellamento e riduzione della componente tecnica e digitale. È una provocazione proposta per suggerire che, a volte, il vero progresso consiste nell'opporsi all'ideale della 'innovazione per l'innovazione', a tutti i costi, cioè dell'inseguimento del nuovo ad ogni costo, e nel riconoscere che quello che si può perdere cambiando abitudini e adottando acriticamente la tecnologia è troppo importante per correre il rischio che questo avvenga. L'invito ad essere ethical by undesign è anche un invito ad opporsi al soluzionismo tecnologico ad oltranza, e cioè a quell'atteggiamento - a cui si è opposto tra i primi Evgeny Morozov - che per ogni problema della nostra società, incluso il diffondersi delle pandemie o il cambiamento climatico, concepisce e promuove soluzioni razionali, tecnologiche e da progettare a tavolino, finendo per trascurare la complessità dei contesti socio-tecnici in cui quelle soluzioni dovrebbero essere applicate efficacemente e per sottovalutare il ruolo di ciò che è imprevedibile e non governabile: cioè, di nuovo, dell'incertezza.

Quali scenari alternativi possibili vede per il futuro dell'intelligenza artificiale?
Federico Cabitza: Come diceva il famoso fisico danese Niels Bohr, è difficile fare previsioni, soprattutto sul futuro e, io aggiungo, lo è a maggior ragione in ambito di evoluzione culturale, perché l'evoluzione dell'intelligenza artificiale e l'impatto che essa avrà sulle nostre vite non riguarderà altro che l'evoluzione e continua trasformazione delle nostre pratiche di condivisione di informazioni e di collaborazione, usando l'automazione per renderci la vita più comoda e per ottimizzare i nostri processi produttivi. Se devo però dare sfogo alla fantasia, posso immaginare due possibili scenari: uno in cui delegheremo sempre più aspetti della nostra vita a queste forme di automazione e le considereremo sempre più scontate, con una consapevolezza sempre minore di quanto le nuove tecnologie digitali potranno condizionare e influenzare il nostro lavoro, le nostre attività quotidiane e il modo in cui impariamo, produciamo e condividiamo conoscenza; e uno scenario in cui ci saremo resi conto di quanto più vulnerabile diventerebbe la nostra società se affidassimo aspetti critici del suo funzionamento e delle nostre attività cognitive alle macchine e ai gruppi di potere che le progettano e le mantengono. Lascio capire quale scenario mi auguro si realizzi, nonostante la sua minore probabilità. Il futuro è aperto, e dipende da noi e da quello che vogliamo fare, come società, delle tecnologie che i nostri laboratori di ricerca producono incessantemente - sotto la spinta di leve di incentivazione spesso legate esclusivamente al profitto di chi li finanzia o al prestigio di chi ci lavora -, a volte con non molta più lungimiranza o saggezza di un apprendista stregone. Nel mondo reale, però, a differenza della famosa storia, è improbabile esista un maestro stregone che ci tiri fuori dai guai in cui dovessimo infilarci illudendoci di controllare e saper gestire ciò che invece non possiamo più controllare.