“Noi, robot” – Update 3.0 al Capitolo 5 (da pag. 165), 31/12/2010

In ultimo, il problema etico

Esiste, al di là della trattazione delle problematiche squisitamente cognitive, linguistiche e operative, e parallelamente alla questione attinente alla produzione artistica, un altro tema piuttosto spinoso che deve essere tenuto in considerazione quando si riflette sul reciproco accostamento di cui sono protagonisti uomo e macchina. Si tratta dell'etica, una categoria dello spirito, e una caratteristica del sentire e agire umano – spiccatamente legata al suo essere sociale, elemento di un gruppo collettivo relazionale – estremamente rilevante, su cui peraltro la discussione filosofica si è lungamente soffermata nell'arco dell'intera tradizione del pensiero.
Contrariamente a quella propriamente detta, l'etica *robot* sta vivendo oggi i suoi albori, peraltro in modo molto dubitativo, ipotetico, tentennante. Il motivo è chiaro: definire i contorni di una teoria etica non può prescindere da volizioni e stati mentali, o almeno dalla consapevolezza di avere di fronte macchine cognitivamente all'altezza di prendere decisioni sulla base di considerazioni valoriali, e quindi solo avendo un approccio possibilistico alle versioni *artificiali* dei detti fenomeni si può pensare di cominciare a tessere un discorso fondato sul tema.
Tutto ciò che può scaturire dalla situazione attuale, in cui una soluzione definitiva all'enigma dell'IA non è giunta, è una semplice lista di regole *comportamentali* – ma sarebbe meglio dire *pattern funzionali* – legati a obiettivi e desideri eteronomi, in quanto umani (e in questo risulta ancora chiara la posizione di subordine in cui le macchine si trovano, fatto naturale in quanto esse sono ancora totalmente eterodirette).
Insomma: definire una serie di regole e impedimenti all'azione di un macchinario, in funzione della necessità di non danneggiare e/o avvantaggiare le persone non significa strettamente costruire un'etica, ma solo (di nuovo, come sempre nel campo della programmazione di macchine, è solo la funzione a cambiare) creare un algoritmo orientato a un obiettivo, senza di fatto andare a discutere il valore che quell'obiettivo ha verso *tutti* i soggetti coinvolti nell'azione.
Anche questa versione depotenziata dell'etica, comunque, presenta una serie di questioni – specie di natura pratica – da risolvere.

Inoltre, anche un'etica propriamente detta si presenterebbe comunque in due vesti: presupponendo la consustanzialità dei due soggetti, essa sarebbe un'estensione trasformativa dell'etica (più precisamente, di una delle varie teorie etiche disponibili nella tradizione filosofica) umana, ripensata per tenere in considerazione i nuovi attori. Ipotizzando invece una radicale alterità, sarebbe necessario costruire una teoria parallela e strettamente dedicata al soggetto robotico – ed è questo, a grandi linee, ciò che sta facendo ad esempio Gianmarco Veruggio con la sua neonata *roboetica*.

Ci sono, quindi, tre macrocategorie del concetto di *etica*, a nostra disposizione. Un'etica *spuria* (la serie di regole operative per la macchina, finalizzate a massimizzare il bene umano), la *roboetica* (un codice comportamentale nuovo che parte dall'assunto secondo cui macchine e uomini sono irriducibili le une agli altri) e un'etica totalmente nuova all'interno della quale macchine e uomini sono parimenti soggetto e oggetto dello stesso tipo di valutazioni – di fatto quest'ultima è ancora lontana nel tempo (ed è ancora estremamente complicato discuterne), e secondo la grande maggioranza degli studiosi non avrà mai ragione di essere, dato che parte da una versione radicale dell'IA forte che è ampiamente minoritaria nel panorama scientifico.

Sul numero di dicembre 2010 della rivista "Le Scienze", appare un articolo molto interessante, sul tema, redatto dalla filosofa Susan Leigh Anderson e dall'informatico Michael Anderson. Il concetto di etica preso in considerazione è il primo, quello più pragmaticamente utile da discutere oggi come oggi, in quanto la progettazione e l'implementazione di un codice etico coerente con quella definizione sembra quantomeno plausibile.

L'idea parte dalla rievocazione della stagione dell'utilitarismo britannico di Bentham e Mills, convinti – per usare la sintesi dei due Anderson – che "prendere decisioni morali dipende da calcoli di *aritmetica morale*", ovvero che "l'azione giusta è quella che probabilmente renderà massimo il *piacere netto*, calcolato addizionando le unità di piacere e sottraendo le unità di dolre vissute da tutti coloro che ne sono affetti".

Ammesso che questa posizione sia accettabile, e che quindi sia possibile pensare a una teoria etica fondata sulla computabilità – conditio sine qua

non dell'implementazione all'interno di un sistema informatico o robotico – si aprono subito delle questioni fondamentali.

Innanzitutto, bisogna ribadire quanto affermato precedentemente, ovvero che questo sistema può dirsi *etico* solo in senso traslato, poiché nel calcolo del "piacere" comune il soggetto dell'agire, ovvero la macchina, non rientra.
Può esistere un calcolo (operato dall'esterno, ovvero valutato da parte dell'uomo, che però è principalmente oggetto) delle condizioni favorevoli alla sopravvivenza, al non danneggiamento, alla migliore operatività, ma questo non esaurisce, nella tradizione etica, lo spazio delle categorie legate al piacere o, più precisamente, al bene.
Inoltre, bisogna ricordare che nella discussione di un sistema etico il soggetto agente ha sempre avuto un ruolo centrale, anche se spesso tenuto presente da terzi – nel caso specifico, i teorici che ne discutevano – capaci in ogni caso, grazie ad empatia, analogia biofisica e quant'altro, di affiancare al puro rilievo statistico di ciò che viene considerato bene desiderabile dai propri simili una maggior profondità di comprensione, e quindi di elaborazione.
In altre parole, il calcolo generale di questa "aritmetica morale" è un calcolo necessariamente parziale. Stiamo quindi affrontando il discorso

Tale osservazione, tuttavia, rappresenta una via d'uscita al successivo punto critico che si genera nel prosieguo della riflessione sull'etica delle macchine, punto critico messo in evidenza anche nell'articolo dei due Anderson, che scrivono:

"Altri dubitano che una macchina sarà mai capace di scelte etiche, perché le macchine non hanno emozioni e non possono apprezzare i sentimenti di chi potrebbe subire le conseguenze delle loro azioni".

La prima, e più immediata, risposta è fornita proprio dai due studiosi, più avanti nel passo:

"Gli esseri umani sono così inclini a farsi trasportare dalle emozioni che spesso finiscono per comportarsi in modo tutt'altro che etico. Questa caratteristica, come la tendenza a favorire noi stessi e i nostri cari, spesso ci rende soggetti men che ideali quando si tratta di decisioni di natura etica."

Insomma: se si parla di decisioni razionali, di analisi freddamente aritmetica, appunto, si suppone che i calcoli siano da operare a priori, per quanto possibile – e la variabilità del sentire umano risulta essere una sorta di impedimento fisiologico, più che una caratteristica intrinseca all'etica.
Diviene tale, peraltro, solo nel momento in cui si debba quantificare il grado di desiderio nei confronti di un bene o di un evento di un soggetto investito dall'azione: trattandosi tuttavia di macchine e avendo stabilito inizialmente che esse non sono soggetti possibili del discorso, nella misura in cui non teniamo in considerazione loro desideri e loro volizioni, il problema si riduce semplicemente all'inserimento, a fianco di regole generali, della volontà del singolo individuo umano, all'interno di ogni situazione nella sua specificità.

Questo pare, più che un problema legato al concetto di etica, uno scoglio cognitivo: un essere umano si presuppone capace (e qui il discorso è complesso, dato il numero di possibili impedimenti clinici, caratteriali, patologici, ecc.) di *leggere* (e di qui: comprendere) le manifestazioni emotive e il grado di desiderio espresso da un suo simile.

Una macchina, allo stato attuale, non può farlo, e può semplicemente basarsi su parametri (quanto più possibilmente precisi) presettati. La dinamica di lettura della disposizione d'animo altrui è quindi il vero problema da affrontare. Ma si tratta, appunto, di un nodo più squisitamente cognitivo.
Un esempio interessante, in questi termini, ci arriva dalla cinematografia, dal citato “I, Robot” di Alex Proyas, tratto dall'omonima opera di Asimov, pionieristica proprio in termini di introduzione del problema etico nel mondo dei robot.
Il protagonista Will Smith, nel raccontare la sventura occorsagli durante un tragico incidente automobilistico, sottolinea come sulla base di un calcolo freddamente razionale, il robot giunto in suo soccorso abbia deciso di salvare la sua vita, e non quella di una bambina altrettanto coinvolta nell'incidente.
A muovere quella scelta, ci sarebbe stata la valutazione delle probabilità di sopravvivenza dei due: essendo maggiori per Smith, la macchina avrebbe quindi tratto lui dall'automobile accartocciata, penalizzando così la ragazzina, poi deceduta. Smith, nel film si scaglia animosamente

contro questa freddezza, e sembra plausibile, di primo acchito, condividere la sua posizione, ritenendo inaccettabile il decorso degli eventi.
Questo, tuttavia, non dimostra che la condotta del robot sia stata *universalmente* antietica: semmai, ci troviamo in presenza di una condotta che contraddice alcuni sistemi etici (di cui Smith, evidentemente, è fautore).
Ma *forse* da una prospettiva benthamiana, l'atto si sarebbe potuto giudicare giusto. Insomma, si tratta di un problema di definizione dei valori, ma non sembra possibile dedurre da questo che il robot non abbia *assolutamente* agito eticamente. Peraltro, nel prosieguo della scena, Smith fornisce un'ulteriore argomentazione (involontaria) a favore della possibilitdi un'etica robot: sostiene accoratamente che il robot non è stato in grado di pensare al futuro della bambina, o ai molti affetti cui era legata, al modo in cui i suoi familiari avrebbero sofferto.
A ben vedere, si tratta proprio di...*nuovi parametri*, e anche lo stesso discorso di Smith, pur emotivamente umano nella maniera in cui viene espresso, da un punto di vista etico altro non fa che fornire una argomentazione razionale a quella che secondo lui sarebbe stata la scelta eticamente più corretta. Se dunque il robot, oltre ad aver calcolato la probabilità di sopravvivenza (che è un singolo parametro della scelta, anche se dipendente da molte variabili), avesse inserito tra le variabili anche l'aspettativa di vita più lunga, il possibile numero di relazioni familiari, la sofferenza di questi ultimi, il minor quantitativo di vita percorso, avrebbe forse agito diversamente?
Insomma: il punto, qui, sta proprio nella quantità (potenzialmente infinita) di fattori in gioco. Il robot, nell'opinione di Smith, non ha semplicemente fatto un freddo calcolo, bensì ha fatto un calcolo molto limitato, e limitato a categorie etiche considerate poco umane. Fosse stato programmato per tenere in conto anche valori intrinsecamente più vincolati alla costellazione dell'emozione, dell'affettività, ecc., avrebbe plausibilmente agito diversamente.

Quindi nei termini della discussione possibile, il modello proposto di etica non appare impossibile da elaborare – con la sola restrizione legata alla definizione di etica, che tuttavia, come sostenuto precedentemente, risulta già intrinsecamente deformata dalla natura del discorso intavolato. Nelle *condizioni d'esistenza* poste, insomma, l'obiezione rimarcata nell'articolo non sembra decisiva.

In ultimo, poi, va sottolineato come la variabilità imprevedibile dell'uomo (che precedentemente nel libro si è constatato essere simulabile anche, semplicemente, attraverso delle routine che randomizzino, in modo raro e non prevedibile, la risposta a un singolo stimolo), quell'inclinazione a essere *soggetto etico non ideale*, altro non sia che il carattere distintivo cui Philip Dick faceva riferimento nel fornire il suo personale parere sulla indissolubile differenza tra la natura umana e quella delle macchine.

Il problema legato alla costruzione di un'etica destinata alle macchine, si riduce così a due difficoltà principali: la prima è l'effettiva capacità, da parte dei robot, di prendere in considerazione (secondo i loro standard computazionali, che sono funzionalmente diversi da quelli umani) un numero molto ampio di variabili e regole comportamentali, svincolandole dalla logica binaria; la seconda è relativa, sostanzialmente, alla *semantica* etica, ovvero all'assegnazione effettiva dei singoli valori ai rispettivi parametri unitari. Insomma, occorre alfabetizzare eticamente i robot.
Il primo punto, di fatto, riporta direttamente all'effettivo progresso tecnologico: macchine sufficientemente potenti, per di più dotate dei necessari strumenti percettivi, potranno immettere all'interno di un calcolo etico un numero incredibilmente alto di parametri differenti, ognuno dotato di un valore desunto dalla singola situazione reale (e affrontato scacchistica mente tenendo in conto un certo intorno causale temporale), per poi computare il tutto attraverso i nuovi e fondamentali modelli legati alla logica *fuzzy*, indispensabile nel manipolare fattori mai binari, salvo negli eventuali casi di *valore assoluto*: possibile esempio, il valore del parametro *conservazione della vita.*
Ecco: valore assoluto, conservazione della vita. In base a quale principi accettare o rigettare quest'ultima come bene inalienabile, incancellabile? O addirittura: dove fondare l'eventuale necessità di immettere all'interno di un calcolo etico dei *valori assoluti*, ovvero dei confini non valicabili, degli *infiniti attuali* in espressioni in cui tutti gli altri fattori sono invece finiti?
Questi, tuttavia, non sono problemi legati all'etica robot, bensì all'*etica generale*, ovvero alla progettazione teoretica del sistema – processo in cui, come abbiamo visto, le macchine non rientrano attivamente.
Arriviamo allora al punto critico numero due, ovvero l'assegnazione dei singoli valori, cioè all'elaborazione di una teoria etica. In questo caso,

sostanzialmente, il fatto che ad essere destinatari di regole siano delle macchine non cambia – anzi: semplifica – il processo di elaborazione: viene rimosso un gravoso fattore da considerare, ovvero lo stato d'animo dell'agente, e si procede alla discussione, squisitamente umana, di quali siano i rapporti valoriali tra le varie istanze del *bene* che possono occorrere.
Diventa allora una questione operativa, rendere il singolo robot in grado di apprendere, laddove ci sono spazi indefiniti nella serie di leggi in suo possesso, nuove regole e nuovi pattern d'azione.
Ancora una volta, tuttavia, si rientra in un macroproblema già visto, ovvero la questione cognitiva dell'apprendimento di regole dall'esperienza – e in questo caso, che siano regole etiche o regole motorie, o di altro genere, poco importa.
La macchina le assimila e le utilizza a prescindere dalla classe filosofica cui decidiamo di farle appartenere.
Il fatto che si stia parlando di dimensione etica, quindi, è piuttosto un problema di catalogazione filosofica, riguarda più che altro l'universo dei programmatori.
Questo è ancor più vero se si considera che nell'arco della storia del pensiero, sono state innumerevoli le teorie proposte, così come i sistemi di valore offerti: diviene allora semplice, nel momento in cui si stabilisce che si sta elaborando un insieme di regole volte alla massimizzazione del bene, accettare che a questo sistema, indipendentemente dal soggetto chiamato in causa, si possa dare l'etichetta di *etica*.