DOPPIOZERO

L’etica dei robot

Giuseppe O. Longo 9 Novembre 2012

Esaminiamo il concetto di “roboetica”, cercando di esplicitarlo nei suoi significati possibili. Dalle considerazioni fatte nei precedenti articoli emerge una prima accezione, molto generale: “roboetica” è semplicemente “l’etica nell’epoca dei robot”, cioè l’insieme dei comportamenti dell’umanità (e la valutazione di questi comportamenti) quando dell’ambiente in cui vivono gli uomini facciano parte anche i robot.

Ma “roboetica” potrebbe anche significare l’insieme (più ristretto del precedente) di quei nostri comportamenti nei confronti dei robot che consentono di mantenere un giusto equilibrio dinamico tra noi e loro.

Poiché i robot posseggono una certa autonomia e una certa capacità di apprendere dall’esperienza, “roboetica” può anche indicare l’insieme dei comportamenti utili, o almeno innocui, dei robot nei nostri confronti. Infine, ed è il significato più avveniristico, potrebbe significare il complesso dei comportamenti che i robot adottano tra loro e verso il loro ambiente, di cui fanno parte anche gli umani.

Riassumendo, la roboetica può significare:

a) L’etica umani -> ambiente (ambiente in cui ci sono altri umani e anche i robot).

b) L’etica umani -> robot.

c) L’etica robot -> umani.

d) L’etica robot -> ambiente (ambiente in cui ci sono anche i robot e gli umani).

Mi rendo conto che si tratta di definizioni approssimative e discutibili, ma da qualche parte bisogna pur cominciare.

La terza accezione si deve conformare al precetto generale e tradizionale per cui le macchine non debbono danneggiarci (*primum non nocere*). È certo questa accezione che aveva in mente Isaac Asimov quando propose le sue famose “Tre Leggi della Robotica”, le quali, cablate in modo inestirpabile nel cervello positronico dei robot, dovrebbero impedire un loro comportamento ostile o dannoso verso di noi:

1) Un robot non può recar danno a un essere umano e non può permettere che, a causa di un suo mancato intervento, un essere umano riceva danno.

2) Un robot deve obbedire agli ordini impartiti dagli esseri umani, purché tali ordini non contravvengano alla Prima Legge.

3) Un robot deve proteggere la propria esistenza, purché la sua autodifesa non contrasti con la Prima o con la Seconda Legge.

Asimov attribuì le Tre Leggi allo scrittore di fantascienza John W. Campbell a seguito di una conversazione avuta con lui un paio di giorni prima del Natale 1940. Ma, secondo Campbell, Asimov aveva già in testa le leggi, che avevano solamente bisogno di una formulazione esplicita.

Benché Asimov avesse fissato la data di creazione delle leggi, il loro progressivo ingresso nelle sue opere richiese un periodo di tempo piuttosto lungo. Così nel 1941 Asimov scrisse due racconti senza menzionare esplicitamente le Tre Leggi, le quali apparvero insieme esplicitamente nel racconto *Girotondo* del 1942.

In apparenza le Tre Leggi sono semplici, chiare, univoche: dovrebbero bastare per regolare perfettamente almeno il punto c). In realtà quando le regole di Asimov fossero calate nel mondo reale non mancherebbero di suscitare problemi e ambiguità. Che cosa vuol dire “danno”? E chi lo stabilisce, chi lo quantifica? Chi ne è responsabile? Il concetto di danno sembra legato al concetto di male (non solo fisico) e sul problema del male si sono arrovellate generazioni di filosofi, teologi, letterati e artisti. Il cervello positronico, razionale e rigoroso, saprebbe impostare e risolvere le “equazioni del male” grazie a un’edizione aggiornata del

calculemus leibniziano, secondo cui ogni problema trova una soluzione qualora ne sappiano impostare i termini in modo rigoroso e quantitativo? C'è da dubitarne...

In effetti la nozione di danno che compare nelle Leggi presenta molte ambiguità: se un umano sta per recare danno a un altro essere umano (per esempio sta tentando di ucciderlo), come si deve comportare il robot? Se interviene reca danno all'assassino, ma il suo mancato intervento reca danno alla vittima.

Inoltre noi uomini siamo contraddittori: come si deve comportare un robot che riceva un ordine contraddittorio (dallo stesso uomo o da due uomini diversi) che sotto il profilo logico metta in crisi il suo sistema di valutazione? Di fronte a una contraddizione gli umani se la cavano quasi sempre con scelte che li fanno "uscire dal sistema" all'interno del quale si annida la contraddizione. Ma questa evasione può avvenire grazie a una certa dose di irrazionalità o di follia creativa. Per consentire al robot di non paralizzarsi di fronte a una contraddizione, si potrebbe forse immaginare di iniettargli un pochino di pazzia... ma con quali conseguenze?

Si può continuare a speculare: se si affidasse lo sviluppo della "specie" robot a un processo evolutivo analogo a quello biologico (o a quello bio-culturale), essi potrebbero compiere - in sostanza fuori del nostro controllo - progressi tali da consentir loro valutazioni etiche più raffinate e precise delle nostre. Per esempio potrebbero, prima o poi, cavarsela meglio di noi in tema di bene e di male (anche se il bene e il male sono sempre riferiti a un soggetto: bene per chi? male per chi?) e potrebbero sviluppare una "teodicea" più rigorosa e soddisfacente della nostra, cioè potrebbero avvicinarsi alla soluzione di un problema teologico e metafisico che ci assilla da sempre: se il creatore del nostro mondo è bontà infinita, perché nel mondo c'è il male? E i creatori del mondo, per i robot, potremmo essere noi...

Ma a quel punto dovrebbero ancora sottostare alla Prima Legge, una legge formulata da creatori imperfetti, incapaci di costruire un mondo privo di male e di insanabili contraddizioni? Oppure sarebbero *loro* a dettarci leggi nuove e ad assumere il bastone del comando, come solerti genitori nei confronti dei loro vivaci e stolti frugoletti? Del resto nel film di Stanley Kubrick *2001: Odissea nello spazio* (1968) il calcolatore *Hal 9000* si comporta proprio così: prende il comando della nave e tenta di uccidere gli umani che intralciano il compimento della missione, invertendo l'ordine d'importanza delle Leggi, cioè subordinando la Prima e la Seconda alla Terza.

Asimov si era certo posto problemi di questo tipo, tanto che in seguito, in un racconto del 1950, *Conflitto evitabile*, aggiunse la Legge Zero:

0) Un robot non può recar danno all'umanità e non può permettere che, a causa di un suo mancato intervento, l'umanità riceva danno.

Questa Legge suppletiva è interessante per il suo carattere "meta" e conferma che le prime tre non sono sufficienti a costituire un'etica di tipo c) sicura. Infatti se un folle minacciasse la distruzione in massa dell'umanità, la Legge Zero autorizzerebbe il robot a eliminarlo, cioè lo autorizzerebbe a infrangere la Prima Legge. Si apre qui il problema della valutazione quantitativa dei danni, ragionevole anche se molto discutibile secondo la morale tradizionale: l'uccisione di molti è (sarebbe) più grave dell'uccisione di uno.

Ma neppure con quest'aggiunta le leggi di Asimov riuscirebbero a proteggerci da comportamenti robotici dannosi, perché le conseguenze ultime di un'azione, pur rispettosa delle Quattro Leggi, potrebbero alla lunga essere nocive per l'umanità o per singoli esseri umani. Infatti l'analisi di queste conseguenze di lunga portata sfiderebbe la più potente intelligenza (naturale o artificiale) immaginabile: troppe sono le ramificazioni e le interazioni con la mutevole complessità del reale. Sappiamo benissimo che anche le azioni umane dettate dalle migliori intenzioni del mondo sfociano spesso in disastri.

Inoltre ci si può chiedere: per valutare se un'azione sia stata buona o cattiva quando ci si deve arrestare nell'esame della catena delle sue conseguenze? Nella società umana solo alcune azioni "cattive" sono giudicate tali esplicitamente e sono sanzionate in un momento preciso grazie a un procedimento giudiziario che interrompe (o almeno vorrebbe interrompere) la catena delle causazioni.

La maggior parte dei nostri atti non sono oggetto di giudizio formale a un istante dato e continuano a provocare conseguenze, positive e negative, nel mondo ben al di là delle nostre intenzioni e per un tempo potenzialmente illimitato.