**Governance digitale e tutela dei diritti: effettività o rischio di "rights washing"?**

*Marta Tomasi*

Convegno annuale dell'Associazione Gruppo di Pisa
"L'effettività dei diritti oltre le corti" – Perugia, 12 giugno 2026

## 1. Introduzione. Dalla crisi dell'effettività alla proliferazione del linguaggio dei diritti

Negli ultimi anni si sono susseguite numerose analisi che hanno descritto, quasi unanimemente, un costituzionalismo sottoposto a pressioni forti, forse inedite. Si tratta di tensioni che hanno dato vita a un vivace dibattito tra coloro che vedono affacciarsi sulla scena un costituzionalismo in veste nuova, quella digitale, e coloro che – rifiutando le specifiche aggettivazioni – ritengono che le coordinate tradizionali siano sufficienti a leggere, interpretare e affrontare le nuove sfide. Nuove forme di potere, nuove sedi della decisione, nuove fonti di normazione, nuove configurazioni della responsabilità pubblica. In questo scenario, anche il tema dei diritti fondamentali è stato investito da un duplice movimento: da un lato, l'emersione di nuove pretese di tutela connesse alle tecnologie digitali e, in particolare, ai sistemi di IA; dall'altro, una progressiva ridefinizione delle modalità attraverso cui i diritti vengono protetti, interpretati e governati.

La questione, in effetti, non riguarda soltanto la tipologia e il contenuto dei diritti, ma la stessa architettura della loro effettività. Come è stato osservato, la peculiarità dei sistemi di IA risiede nella loro capacità di operare su scala, di integrarsi in processi decisionali ordinari e di automatizzare valutazioni che incidono in modo diffuso e spesso invisibile sulle condizioni di vita delle persone (Shany). In tale contesto, mutano radicalmente velocità, estensione e reversibilità del danno.

I rischi prodotti dall'AI non si esauriscono più nell'errore puntuale o nella decisione individuale illegittima, ma assumono dimensioni strutturali e collettive: *pattern* discriminatori, effetti di *chilling*, danni diffusi difficilmente imputabili a un singolo decisore. Proprio questa trasformazione mette in crisi il paradigma tradizionalmente reattivo di tutela dei diritti fondamentali, costruito attorno all'idea che la violazione si manifesti in un evento identificabile, riconoscibile dal soggetto leso e successivamente sindacabile davanti a un giudice o a un'autorità indipendente (Munjić).

L'assetto rimediale classico mostra, infatti, profonde difficoltà quando la causalità è diffusa, la responsabilità distribuita tra molteplici attori e le ragioni della decisione risultano tecnicamente opache. È precisamente in questo spazio che la governance algoritmica tende a produrre una frattura crescente tra riconoscimento formale dei diritti e concreta possibilità di esercitarli (Quintavalla, Temperman).

Da qui deriva il progressivo spostamento del diritto da una logica prevalentemente reattiva a una logica anticipatoria e preventiva. Le recenti evoluzioni della governance europea dell'AI –

dagli obblighi di *risk assessment* alle valutazioni di impatto – vengono spesso presentate come il segno di una transizione metodologica: dalla tutela ex post alla protezione proattiva dei diritti fondamentali (Munjić).

Se l'effettività, la conformità fra enunciazione e attuazione, è da intendersi come fattore che condiziona l'essenza e l'esistenza stessa di un sistema giuridico (Fuller), è attraverso questa lente che l'apparente espansione del linguaggio dei diritti nella governance digitale deve essere indagata. La questione, dunque, è se i nuovi approcci che si stanno sviluppando, orientati all'anticipazione del rischio e alla responsabilizzazione degli attori del digitale, consenta un rafforzamento concreto degli standard di tutela dei diritti o se, viceversa, non si affacci il rischio di forme di "rights washing". L'espressione è ripresa e adattata dalle più diffuse formule di "green -" e "ethics washing" e tende ad evidenziare il rischio di appropriazione del linguaggio dei diritti da parte di attori privati e la sua cattura all'interno di dispositivi tecnico-procedurali.

## 2. L'appropriazione del linguaggio dei diritti nella governance digitale

Il linguaggio dei diritti fondamentali possiede innegabilmente (e nonostante le recenti dinamiche di arretramento) una straordinaria forza di legittimazione. Come è stato osservato anche nel dibattito internazionalistico, il riferimento ai diritti umani costituisce al tempo stesso un quadro normativo e una risorsa simbolica capace di conferire autorevolezza, credibilità e consenso a istituzioni, pratiche e modelli di governance assai eterogenei (Fasoli/ Başak Çalı). Non sorprende, allora, che proprio tale lessico venga sempre più frequentemente adottato anche da soggetti privati operanti nel settore tecnologico (cfr. *ethics bluewashing* di Floridi). Il fenomeno è particolarmente evidente nella *platform governance* e, più recentemente, nei modelli di autoregolazione dell'IA, nei quali categorie tradizionalmente appartenenti al costituzionalismo democratico vengono progressivamente riutilizzate all'interno di architetture normative private. E anche là dove l'attenzione ai diritti è imposta dal potere pubblico, attraverso strumenti normativi, permane il rischio di dinamiche distorsive e forme di "contaminazione" proprie della governance privata, poiché l'attuazione concreta di tali obblighi si sviluppa comunque entro assetti organizzativi, incentivi e razionalità tipici degli attori di mercato.

Le pagine che seguono intendono soffermarsi su due casi nei quali tali dinamiche emergono con particolare chiarezza.

### 2.1. Claude's Constitution

Emblematico, sotto il primo profilo, è il caso della cosiddetta "Constitution" di Claude elaborata da Anthropic per il proprio LLM. Il documento, presentato dall'azienda come una descrizione dettagliata dei valori e dei comportamenti cui il modello dovrebbe conformarsi, costituisce l'evoluzione del progetto di "constitutional AI" sviluppato dalla società a partire dal 2022. Si tratta di un tentativo ragionato, che fa seguito a una articolata e approfondita riflessione filosofica e interdisciplinare e che lascia trasparire il bisogno, l'ansia e insieme la difficoltà di governare un progresso tecnologico pervasivo e trasversale. In base a questo "documento normativo fondamentale", Claude dovrebbe agire secondo una gerarchia di valori

(sicurezza, etica, conformità alle linee guida aziendali, utilità), accompagnata da specifici divieti (*no-go areas)*.

È evidente che il ricorso al linguaggio costituzionale non basta ad avvicinare queste architetture normative private ad autentiche costituzioni. Si tratta, anzi, come è stato osservato, di espressioni di uno "pseudo-costituzionalismo" radicalmente distinto dal costituzionalismo novecentesco (Betzu). Le differenze sono tanto ovvie, quanto profonde. Principalmente, questi dispositivi tendono a legittimare assetti di potere già esistenti, piuttosto che limitarli; sono privi di un catalogo di diritti organizzati, oltre a essere del tutto assente un mandato democratico capace di giustificare la selezione e gerarchizzazione dei valori incorporati nel sistema.

Un aspetto significativo, però, che qui preme segnalare, ha a che fare proprio con la struttura discorsiva del testo: in linea con il funzionamento di modelli non *rule-based*, Anthropic costruisce la propria "Costituzione" non come un semplice insieme di regole tecniche, bensì come un quadro di valori ai quali la macchina dovrà adeguarsi, sviluppando *alignment*, *reasoning* e auto-valutazioni che evocano attitudini di radice aristotelica, simili alla capacità di "giudizio" e alla "saggezza pratica".

Da questo punto di vista, è difficile non rinvenire una assonanza con il linguaggio costituzionale dei diritti (Pino), il quale positivizza il valore che intende tutelare, attraverso clausole che non risolvono l'indeterminatezza di quel valore, chiamando in causa opzioni valutative. Due elementi, tuttavia, differenziano in modo significativo – soprattutto in punto di ricadute effettive sulla tutela dei diritti – il mondo del costituzionalismo reale dal mondo degli algoritmi.

La prima questione riguarda la selezione dei valori elevati a parametro. Pur dichiarando apertamente la propria non-piena neutralità etica e filosofica, infatti, la "Costituzione" di Claude ritiene sufficiente che i giudizi di valore svolti possano risultare ragionevolmente accettabili per ampie categorie di soggetti interessati ("we aim to make value judgments that wide swaths of relevant stakeholders can feel reasonably comfortable with"). Questo approccio, che dichiara il punto cieco, ma non lo affronta oltre, porta alla luce un aspetto spesso occultato nella governance algoritmica: ogni sistema normativo incorpora inevitabilmente una specifica comprensione del bene, storicamente e culturalmente situata (Taylor "background understandings").

La "Costituzione" di Claude sembra oscillare precisamente tra questi due poli: da un lato la consapevolezza della propria non neutralità; dall'altro la tendenza a presentare la parte come il tutto (gli antropologi parlano di una *synecdochic attitude* dei sistemi algoritmici – Hejazi, Franch, Roberti, Blanzieri), a universalizzare un particolare orizzonte assiologico – quello della "cultura tech progressista" della Silicon Valley – trasformandolo in *benchmark* apparentemente neutro di ragionevolezza, sicurezza o utilità.

Il secondo aspetto problematico riguarda il modo in cui l'indeterminatezza dei valori viene resa governabile e i casi di conflitto valoriale portati a composizione. Mentre nel costituzionalismo democratico ciò avviene attraverso l'operato di istituzioni (*in primis* le corti) che impiegano

tecniche argomentative e pratiche professionali condivise e verificabili, limitando i rischi di "opzioni assiologiche sfrenatamente soggettive" (Pino), nel mondo algoritmico la decisione viene incorporata direttamente nell'architettura tecnica – opaca e priva di responsabilità democratica– del sistema. In questo senso, il problema non consiste soltanto nella privatizzazione della normazione, ma nella progressiva sostituzione della mediazione democratica con forme di razionalità tecnico-funzionale che tendono a presentarsi come neutrali proprio mentre incorporano e rendono operative profonde scelte valoriali. Emblematico, in questo senso, è il fatto che il parametro evocato dal documento sia quello di "a thoughtful senior Anthropic employee - —someone who cares deeply about doing the right thing", figura assunta quale criterio generale di giudizio, ma che resta inevitabilmente collocata entro un preciso contesto culturale, economico e politico.

Questo esempio di appropriazione del linguaggio costituzionale – che pure è frutto di una "riflessività interna" (Taylor) – rischia di conferire una sorta di patente di democraticità a forme raffinate e sottili di egemonia privata.

### 2.2. Fundamental Rights Impact Assessment

Un rischio distorsivo analogo, sebbene diversamente configurato, emerge dall'analisi di alcuni degli strumenti ai quali la regolazione pubblica contemporanea – in particolare nel contesto europeo – affida la tutela dei diritti fondamentali nella governance digitale. Se nel caso del Constitutional AI il lessico dei diritti viene mobilitato spontaneamente da attori privati come fonte di legittimazione e orientamento para-normativo, nei meccanismi introdotti dal legislatore europeo il movimento sembra procedere nella direzione opposta: è il potere pubblico a imporre obblighi formalmente finalizzati alla protezione dei diritti fondamentali. Il riferimento è ai modelli di *due diligence* che caratterizzano una parte crescente della regolazione europea delle tecnologie digitali, che rappresentano interessanti esperimenti di co-regolazione dell'universo digitale, e, più in particolare, del meccanismo di *Fundamental Rights Impact Assessment* (FRIA) previsto dall'AI Act. Questi strumenti si collocano entro la più ampia logica dell'*accountability*, orientata almeno nelle intenzioni, a promuovere forme di tutela sostanziale dei diritti attraverso processi continui di identificazione, valutazione e mitigazione dei rischi. Tuttavia, la crescente centralità di siffatti dispositivi solleva interrogativi circa la possibilità che, anche in questo caso, il ricorso al linguaggio dei diritti finisca per operare come una strategia di legittimazione delle funzionalità algoritmiche, senza che tale centralità simbolica si traduca necessariamente in un corrispondente rafforzamento di forme di tutela effettiva.

La FRIA rappresenta probabilmente l'esempio più significativo di questa tendenza. Lo strumento si ispira dichiaratamente alla *Data Protection Impact Assessment* (DPIA) prevista dal GDPR, riprendendone la logica preventiva e il tentativo di integrare la tutela dei diritti all'interno dei processi decisionali delle organizzazioni. Tale filiazione va tuttavia indagata con maggiore attenzione sotto almeno due profili. Da un lato, la DPIA è stata elaborata nell'ambito di una disciplina – il GDPR – che ha nella tutela dei diritti fondamentali il proprio obiettivo primario, mentre la FRIA viene introdotta all'interno dell'AI Act, un atto che combina finalità di

tutela con obiettivi di armonizzazione del mercato e promozione dell'innovazione tecnologica. Dall'altro lato, la stessa esperienza delle DPIA ha già mostrato alcuni limiti strutturali di siffatti strumenti, che possono essere interpretati come meri esercizi procedurali di gestione del rischio. In particolare, è stato osservato come approcci eccessivamente centrati sulla dimensione individuale possano risultare inadeguati a cogliere forme più ampie e sistemiche di lesione dei diritti, quali quelle che incidono sull'eguaglianza, sulla non discriminazione o sulla distribuzione del potere sociale.

Limitazioni analoghe sembrano emergere anche con riferimento alla FRIA. Sebbene essa rappresenti uno strumento potenzialmente rilevante per rafforzare la tutela dei diritti fondamentali nei sistemi di IA ad alto rischio, la sua configurazione normativa attuale appare ancora caratterizzata da un elevato grado di vaghezza metodologica e da una marcata proceduralizzazione. Come ha osservato attenta dottrina (Mantelero), il Regolamento individua l'obbligo di valutazione, senza fornire però una metodologia sufficientemente robusta per orientarne concretamente lo svolgimento, lasciando ampi margini di discrezionalità agli operatori chiamati ad applicarlo.

Il rischio è che uno strumento concepito per favorire una valutazione sostanziale degli impatti sui diritti fondamentali finisca per ridursi a un adempimento prevalentemente documentale. Non a caso, parte della letteratura ha osservato come la FRIA rischi di operare soprattutto come requisito di documentazione e tracciabilità, senza impedire o limitare necessariamente l'impiego di sistemi che presentano rischi significativi per i diritti fondamentali. In questa prospettiva, la tutela dei diritti tende a essere tradotta – al meglio – in procedure, indicatori, matrici di rischio e meccanismi di reporting che, pur aumentando la visibilità, la formalizzazione e dunque la leggibilità dei processi decisionali, possono favorire una logica di compliance maggiormente orientata alla dimostrazione della conformità che alla trasformazione sostanziale delle pratiche applicative. Tale dinamica richiama il più generale problema della "seduzione della quantificazione" (Sally Engle Merry), animata dalla promessa della misurazione di rendere conoscibili e governabili fenomeni complessi attraverso dati, indicatori e metriche apparentemente oggettive. Tuttavia, proprio nel passaggio di traduzione in categorie misurabili, i diritti rischiano di essere progressivamente privati del loro contesto (sociale, politico e culturale) di riferimento e, dunque, del loro significato più pieno. Separata da valutazioni qualitative capaci di cogliere la complessità dei contesti nei quali gli impatti si producono, l'assorbimento entro pratiche di conformità rischia di sostituirsi a processi di effettiva comprensione e gestione.

A ciò si aggiunge l'ulteriore rischio della "selettività". Le valutazioni, infatti, tendono a concentrarsi sui diritti maggiormente consolidati nelle pratiche regolatorie della governance digitale o più agevolmente "misurabili", perché dotati di una matrice maggiormente "tecnica". Non sorprende, in questo senso, che questioni relative alla protezione dei dati personali o alla discriminazione algoritmica occupino spesso una posizione centrale, mentre altre dimensioni potenzialmente rilevanti – quali gli impatti ambientali, le condizioni di lavoro lungo la filiera tecnologica o gli effetti sistemici sulle relazioni sociali e democratiche – rischiano di ricevere

un’attenzione minore. Le recenti interviste condotte dalla *Fundamental Rights Agency* dell’Unione Europea – sintetizzate nel rapporto *Assessing High-risk Artificial Intelligence: Fundamental Rights Risks* del 2025 – con *stakeholders* coinvolti nell’attuazione dell’AI Act confermano tali preoccupazioni, evidenziando il rischio di una progressiva standardizzazione delle valutazioni e di una riduzione della complessità dei conflitti valoriali a categorie predeterminate di rischio. In tal modo, ciò che nasce come strumento volto a rafforzare la tutela dei diritti fondamentali rischia di contribuire alla loro amministrazione burocratica, privilegiando la gestione del rischio rispetto al confronto sostanziale con le ragioni e i valori che quei diritti intendono proteggere.

**3. Il rischio dell’effettività apparente**

I due esempi discussi, pur nella loro diversità, mettono in luce un rischio comune: che il crescente ricorso al linguaggio dei diritti contribuisca alla costruzione di una sorta di *rights-compliance façade*, di evaporazione dell’effettività, una cosmesi nella quale la tutela dei diritti fondamentali viene, da un lato, ricondotta a specifiche visioni valoriali elaborate all'interno di contesti privati, prive di una chiara legittimazione democratica e spesso difficilmente scrutinabili; dall'altro, tradotta in *checklist*, procedure e adempimenti tecnici che rischiano di privilegiare la conformità formale rispetto alla protezione sostanziale dei diritti stessi.

Del resto, come ampiamente evidenziato dalla riflessione sul carattere contestuale del linguaggio normativo, il significato dei concetti giuridici non può essere separato dalle pratiche istituzionali e dagli assetti di potere entro i quali essi vengono impiegati. Ne consegue che la “traduzione” del lessico dei diritti dal contesto del costituzionalismo democratico a quello della governance digitale non possa essere considerato un’operazione semanticamente neutra, ma richieda un costante scrutinio critico delle trasformazioni che tale trasposizione produce sul significato, sulle funzioni e sulle modalità di tutela dei diritti medesimi.

Una provocazione finale può aiutare a chiarire il punto. L’AI Act dedica un riferimento importante all’alfabetizzazione in materia di IA (*AI literacy*), avendo intuito che la diffusione e la pervasività di tali tecnologie richieda oggi consapevolezze, competenze e capacità nuove. Nel solco dell’effettività, tuttavia, altrettanto importante potrebbe apparire la promozione di una *rights literacy* che restituisca la complessità e la profondità della riflessione giuridica in tutti gli spazi all’interno dei quali si è chiamati a garantire tutela ai diritti.