Centro Studi sul Pensiero Contemporaneo

Lessico di Etica Pubblica

Lexicon of Public Ethics

Anno 15, numero 1, 2024

Rivista semestrale di proprietà del CeSPeC, registrata presso il Tribunale di Cuneo, n. 621, in data 26/3/2010

Citabile come: «Lessico di etica pubblica», 1 (2024). ISSN 2039-2206
Cite this journal as: «Lexicon of public ethics», 1 (2024). ISSN 2039-2206

La rivista pubblica contributi selezionati tramite sistema di blind review e apposite *calls for paper*.
The journal publishes contributions selected through blind review and special calls for paper.

Per sottoporre il proprio testo e per ogni altra informazione, contattare la redazione all'indirizzo: redazione.eticapubblica@gmail.com
To submit your text and for any further information, please contact the editorial team at: redazione.eticapubblica@gmail.com

Trasparenza.
Riflessioni estetiche, implicazioni politiche

Transparency.
Aesthetic Reflections, Political Implications

A cura di | Edited by
Graziano Lingua & Francesco Striano

Trasparenza. Riflessioni estetiche, implicazioni politiche| **Transparency. Aesthetic Reflections, Political Implications**, a cura di |edited by Graziano Lingua & Francesco Striano

Indice - Table of Contents

RICERCHE – RESEARCH

RECENSIONI – REVIEWS

Trasparenza.
Riflessioni estetiche, implicazioni politiche

Graziano Lingua[*], *Francesco Striano*[†]

La nozione di trasparenza, nel suo significato originario nell'ottica fisica, indica la proprietà posseduta da un materiale di poter essere attraversato dalla luce, mostrandoci quindi ciò che si trova al di là di esso. Un materiale, quanto più è trasparente, tanto più efficacemente si nasconde e dà l'impressione di non esserci.

La fortuna dell'uso metaforico di questo concetto risiede in questa sua ambiguità: un oggetto, un'istituzione, una pratica sono "trasparenti" nella misura in cui esistono e esercitano la propria funzione non apparendo, come se non ci fossero. Per questo la trasparenza diventa sinonimo di apertura illimitata e accesso diretto alla realtà. Essa tende ad occultare la natura dei processi di mediazione che sono comunque in gioco e le zone d'ombra che accompagnano ogni processo che si pretende trasparente. E allo stesso tempo però il bisogno di trasparenza nasce proprio per comprendere il senso profondo delle mediazioni e renderle accessibili perché non siano semplicemente una barriera, invisibile ma reale, alla comprensione del mondo.

Sin dagli albori della società digitale, pertanto, "trasparenza" è diventata una parola chiave, nella convinzione che l'accesso a un'enorme massa di informazioni potesse spalancarci le porte di una totalità trasparente, migliorando così la nostra conoscenza del mondo, degli eventi, dei processi deliberativi e decisionali. Essa si è trasformata in un vero e proprio imperativo che ha fatto della disintermediazione l'antidoto alle opacità del potere e la condizione essenziale per la fiducia nei confronti delle istituzioni che sta alla base della partecipazione democratica. Tale imperativo non si è manifestato unicamente a livello politico nel rapporto verticale tra cittadini e governanti, ma si è esteso anche a livello orizzontale nelle relazioni sociali fino a implicare la vita personale. In ogni ambito la trasparenza sembra oggi elevarsi a valore assoluto e a principio indiscutibile, facendo dimenticare le ambiguità e i rischi di questo concetto.

Questo entusiasmo per la visibilità totale elude però un dato di fatto: la trasparenza non è l'assenza di mediazione e quella che definiamo come "disintermediazione" si traduce inevitabilmente in nuove forme di mediazione che magari diventano anche meno "trasparenti" perché sono più subdole e difficili da

[*] Professore ordinario, Università di Torino, e-mail: graziano.lingua@unito.it.
[†] Assegnista di ricerca, Università di Torino, e-mail: francesco.striano@unito.it.

individuare. Un medium "trasparente" permette di vedere una porzione di realtà, ma ne nasconde un'altra. Ogni relazione sociale che si pretende trasparente è in realtà guidata da scelte individuali e collettive che strutturano specifici "regimi di visibilità". Da questa ambiguità e dal suo occultamento derivano alcuni problemi che una riflessione critica sulla cultura e sulla società deve oggi affrontare.

1. Storia e significati della trasparenza

L'uso metaforico della trasparenza e il suo riferimento all'apertura e alla visibilità come principi indiscutibili della vita contemporanea non deve far dimenticare che originariamente il concetto viene elaborato in ambito politico con un più esplicito riferimento alla parola che non allo sguardo, alla discussione piuttosto che alla visibilità. La nozione moderna di di trasparenza si costruisce infatti durante l'Illuminismo (anche, se come ricordano Lingua e Alloa, nella letteratura del periodo il termine è praticamente assente) e matura intorno all'esigenza che gli atti di governo siano pubblici e che sia possibile accedere alle informazioni e discuterne in modo libero. Anche dopo l'epoca dei Lumi la tradizione liberal-democratica ha mantenuto questo primato della parola come strumento per combattere i segreti del potere e per creare una sfera pubblica capace di trasformare i cittadini in attori della vita politica. Poter dibattere delle proprie idee e poterle esprimere a tutti senza vincoli di sorta non serve unicamente a emanciparsi da poteri assoluti, ma rappresenta anche il presupposto per sentirsi coinvolti in quanto soggetti che contribuiscono ad una costruzione cooperativa della cosa pubblica.

Quella dimensione verbale della trasparenza è stata però progressivamente messa in secondo piano per lasciare spazio allo sguardo e alla richiesta di una *totale visibilità*. A favore di quest'ultimo aspetto ha sicuramente giocato un diverso immaginario politico che individuava nelle "pareti di vetro" l'emblema architettonico di una utopia sociale di completa accessibilità del potere e di pieno controllo di chi governa, ma ciò che ha permesso il definitivo slittamento della trasparenza dal verbale al visuale è stato senza dubbio l'ingresso dei media visuali di massa. La nascita della fotografia e del cinema, e la presenza della televisione nelle case di ogni famiglia hanno trasformato la sfera pubblica da luogo della sola parola a spazio in cui vanno in scena gli attori politici e ogni loro pratica può essere messa sotto i riflettori.

Con la svolta digitale si è fatto un passo ulteriore in questa direzione grazie alle promesse di piena accessibilità che hanno accompagnato la nascita e lo sviluppo di internet. Online la visibilità è divenuta la forma per eccellenza della trasparenza tanto da debordare ben oltre l'esperienza politica e invadere tutti i campi dell'esistenza, fino a toccare le esperienze più intime e personali, anche a costo di sacrificare la privacy. Insomma, l'apertura e la visibilità si sono trasformati in un presupposto di moralità e di affidabilità e la vera e propria ossessione per l'esposizione ha finito per nasconderne i problemi e gli effetti indesiderati. L'overdose di informazioni visive ha generato un rumore di fondo in cui le questioni rilevanti rischiano però di perdersi sacrificate sull'altare di una trasparenza che si vuole totale senza in realtà poterlo essere.

2. Trasparenza e sfera pubblica

Malgrado l'espansione in tutti i campi dell'esistenza individuale e collettiva, la politica e l'economia restano due degli ambiti privilegiati in cui si può misurare il significato del dibattito odierno sulla trasparenza. Come evidenziato da Zani, la trasparenza è stata intesa come cruciale per il corretto funzionamento dei mercati nel pensiero liberale e, in particolare, nel contesto della scuola austriaca: nel suo articolo l'autore esplora le diverse interpretazioni di questo concetto mostrando come, da un lato, Mises la invochi per limitare l'intervento dello Stato, mentre Hayek la veda come una mera illusione, data la natura opaca della società umana. Eppure entrambi gli autori usano la coppia concettuale trasparenza/opacità per giustificare l'inibizione dell'intervento pubblico nell'economia.

Il dibattito sulla trasparenza economica si ripresenta poi nelle analisi delle crisi finanziarie, come quella del 2007-2008, dove l'asimmetria informativa ha giocato un ruolo cruciale. A questo riguardo, l'articolo di Miotti sottolinea la necessità di strategie di trasparenza intesa come "disclosure" per garantire una comprensione omogenea dei prodotti finanziari da parte degli investitori.

La trasparenza, comunque, non è da intendersi nella sfera pubblica solo come un requisito economico, ma anche come pilastro della pubblica amministrazione. L'articolo di Pedretti ripercorre la storia del concetto di trasparenza nella pubblica amministrazione, a partire da una riflessione sulla celebre metafora della "casa di vetro" ripresa da Filippo Turati. L'autore sottolinea che la trasparenza non può limitarsi alla pubblicazione degli atti, ma richiede che essi siano comprensibili e accessibili ai cittadini, garantendo il controllo diffuso sull'esercizio dei pubblici poteri. L'articolo ricorda anche le pietre miliari nella storia del diritto all'accesso alle informazioni, come il *Tryckfrihestörordningen* svedese del 1766 e il FOIA statunitense del 1966. Si evidenzia, però, come nonostante l'evoluzione normativa, permangano delle criticità nell'effettiva accessibilità. Viene anche introdotto il concetto di "democrazia monitorante" che sottolinea il ruolo *attivo* dei cittadini nel controllare l'operato delle istituzioni (come nel caso del monitoraggio civico sui beni confiscati alle mafie), introducendo dunque una concezione dinamica e multilaterale della trasparenza.

Se questo ruolo attivo nella rivendicazione dal basso di una trasparenza politica è senza dubbio importante, Lingua, Monti e Poirer introducono, però, nel loro articolo un'interessante osservazione critica circa il rischio di derive populistiche della retorica "trasparentiste". Il loro articolo, infatti, esamina la crescente influenza della leadership carismatica nella politica dei movimenti populisti e la spinta per una democrazia diretta digitale come alternativa al ruolo dei parlamenti per criticarne i presupposti. Tuttavia, osservano gli autori, le strategie populiste non soddisfano gli standard di immediatezza e trasparenza su cui basano la loro retorica e non risultano, da un punto di vista politico, in grado di assolvere alle esigenze di una

rappresentazione pluralistica degli interessi dei cittadini e quindi non possono sostituire la funzione democratica delle assemblee elette.

3. Trasparenza e tecnologie emergenti

Quando si parla di tecnologie digitali, robotica, o intelligenza artificiale, la questione della trasparenza assume però ulteriori sfaccettature e la nozione stessa potrebbe non essere riconducibile ai significati di "disclosure" o di accessibilità.

Nel contesto dell'Intelligenza Artificiale (AI), l'articolo di Tangari, ad esempio, introduce la distinzione tra "trasparenza" e "trasparibilità", suggerendo che la comprensione (e quindi la fiducia) non è sempre legata alla piena accessibilità ai dati, ma anche a fattori personali e spesso inconsci. La difficoltà di comprendere i processi decisionali delle AI, spesso descritti come "black box", solleva interrogativi sulla necessità di criteri condivisi per valutare l'interpretabilità dei sistemi. L'articolo sottolinea come la stessa razionalità umana sia complessa da descrivere – e quindi anche da confrontare con i sistemi di intelligenza artificiale –, sollevando dubbi sulle aspettative che si hanno verso questi ultimi.

La difficoltà di ottenere una "trasparenza totale" da parte di questi sistemi viene ribadita nell'articolo di Demichelis, che sottolinea come i modelli esplicativi dell'AI siano sempre delle "mappe" che non corrispondono esattamente al "territorio". La metafora della mappa e del territorio serve a illustrare che, nella spiegazione di un modello di intelligenza artificiale, non si può e non si deve ricercare una corrispondenza 1:1 con il modello originale. Come la mappa rispetto al territorio, anche un modello esplicativo deve contenere meno informazioni del modello originale.

Di Cicco, dal canto suo, si concentra sulla complessità della trasparenza nel campo della robotica sociale. L'autrice distingue tra "trasparenza della robotica sociale" (come campo di ricerca) e "trasparenza dei robot sociali" (come attori sociali), sottolineando ancora una volta come la trasparenza sia un concetto polisemico. L'articolo mette in luce come l'antropomorfismo dei robot sia alla base della loro socialità percepita, ma possa essere anche fonte di inganno. Di conseguenza, un robot sociale genuinamente trasparente potrebbe non essere realizzabile, perché la trasparenza rischia di vanificare il ruolo sociale dei robot. L'articolo suggerisce quindi che la trasparenza non è sufficiente per garantire una robotica sociale responsabile

Se, tuttavia, la trasparenza totale non è mai ottenibile, Giacomini e Aprilis ci mettono in guardia dal dismetterla troppo sbrigativamente, ricordandoci gli aspetti nocivi dell'opacità algoritmica, che può portare alla perpetuazione e amplificazione di disuguaglianze esistenti. L'articolo richiama molti degli esempi classici in cui i sistemi di intelligenza artificiale hanno mostrato bias nei confronti di donne, minoranze etniche e persone con disabilità. Tali bias non sono "glitch" del sistema, ma derivano da dalle basi di dati con cui è sono alimentate le IA nonché dai pregiudizi dei programmatori e sono sfruttati dalle aziende per trarne profitto. Da queste considerazioni emerge l'importanza di rendere l'AI "spiegabile" (XAI) per verificare

che non promuova decisioni discriminatorie.

4. Etica e trasparenza

La tensione tra trasparenza e opacità emerge anche nel contesto dello sport. Matteo Cresti, nel suo articolo, discute il valore morale della trasparenza riguardo all'uso di sostanze dopanti nel bodybuilding, sottolineando come l'apertura e la divulgazione di tali pratiche permetta agli atleti di ricalibrare le proprie aspettative. L'articolo descrive come la trasparenza in questo ambito si sia evoluta nel tempo, da una fase di aperta promozione, al rifiuto, per poi giungere ad una fase di nuova ammissione da parte di alcuni atleti.

Al tema del doping si collega anche quello del biohacking, analizzato nel saggio di Baron e Graffeo come una forma di tecno-ascetismo. L'articolo sottolinea come la trasparenza, in questo contesto, si concretizzi nella creazione di "spazi di visibilità" in cui le informazioni sul corpo e le sue funzioni interne vengono rese trasparenti, organizzate e condivise. Viene evidenziato come il biohacking possa essere visto come una risposta alla proliferazione di disinformazione e una forma di "democratizzazione" della scienza

Il tema della trasparenza si interseca con quello dell'etica e della responsabilità non soltanto in ambito sportivo o bioetico e biopolitico, ma anche nel già citato rapporto tra umano e tecnologia. L'articolo di Striano, a questo proposito, esplora la complessità del rapporto tra trasparenza e fiducia, soprattutto in relazione alla tecnologia. L'autore mette in discussione l'idea che la trasparenza sia un fattore sempre moralmente connotato positivamente, suggerendo che un'eccessiva enfasi sulla trasparenza possa paradossalmente minare la fiducia. Striano sostiene, infine, che sia l'*onestà* la virtù tecno-morale da coltivare nella sfera tecno-sociale, evidenziando come la trasparenza da sola non sia sempre sufficiente a garantire la fiducia nei confronti delle tecnologie o nei rapporti mediati da esse.

Ancora sul versante etico, il saggio di Alloa e Lingua analizza come l'ossessione per la trasparenza possa portare a un'ideologia della neutralità che nasconde le dinamiche di potere e gli effetti distorsivi che tale ossessione produce. L'articolo evidenzia come la trasparenza venga spesso identificata con l'immediatezza – idea alimentata dai media digitali – ignorando il ruolo della mediazione.

Allo stesso modo, la trasparenza individuale può diventare una forma di auto-assoggettamento se non si è consapevoli delle dinamiche di potere che si nascondono dietro la condivisione di dati online. E anche qualora la trasparenza venisse intesa all'esatto opposto dell'ipertrofia dell'identità personale, e cioè come un'invisibilizzazione – come può avvenire in alcuni contesti di rete – essa potrebbe essere ugualmente eticamente indesiderabile, come argomenta Davies, evidenziando come, alla fine, la perdita di beni morali come l'identità, l'integrità e l'autostima possa essere più dannosa dei vantaggi materiali che si potrebbero ottenere con l'invisibilità.

5. Narrative differenti

Il fatto di superare la facile identificazione della trasparenza con l'immediatezza e l'apertura totale è un tratto comune che caratterizza molti dei saggi raccolti in questo numero. Peraltro come sottolineano Lingua e Alloa nell'articolo che apre la raccolta questa identificazione non ha soltanto qualcosa di controfattuale rispetto alle concrete pratiche della trasparenza, ma alligna un elemento ideologico perché rende difficile discutere a fondo di un principio che sembra imporsi come totalmente evidente e comunque sempre positivo.

Lo sforzo che compiono gli autori, ciascuno nel proprio ambito di competenza, di tematizzare invece le ambiguità e gli effetti distorsivi che la trasparenza può generare sono un contributo a sfatare quell'aura di intoccabilità che avvolge questo concetto. La consapevolezza è un passo importante per contribuire a delineare i limiti e a chiarire gli ambiti all'interno dei quali la storia di lunga durata di questo concetto può oggi ancora avere un senso. Emerge così il ruolo che altri temi possono avere nel lavoro di ripensamento critico, come la funzione centrale della fiducia sociale, l'importanza dell'onestà come alternativa a un "trasparentismo" generalizzato o i rischi sempre più significativi che saldano la trasparenza alla sorveglianza e al controllo generalizzato.

D'altro canto, limitandoci anche solo a quest'ultimo tema che sembra oggi l'effetto collaterale più negativo delle dimensioni individuali e collettive della trasparenza, occorre segnalate che la sorveglianza non esprime soltanto i poteri di controllo sempre più diffusi, ma risponde anche a esigenze di protezione, di sicurezza e di cura per gli altri, riproponendo le stesse ambiguità e lo stesso bisogno di una tematizzazione critica che richiede la trasparenza. Si pensi per esempio alle pratiche di contro-sorveglianza con cui i cittadini possono "sorvegliare i sorveglianti", utilizzando i propri dispositivi digitali per documentare per esempio le atrocità della polizia o altri abusi di potere, pratiche che ancora una volta mostrano come la ricerca della trasparenza continui nonostante tutto a alimentare al proprio interno sincere istanze di partecipazione e di resistenza anti-autoritaria.

Le piste di ricerca percorse in questo numero rappresentano quindi uno sforzo di aprire un varco nell'ordine dominante del discorso per fare spazio a narrazioni differenti che ci facciano intravedere percorsi alternativi alla mera sottomissione all'ideologia della trasparenza totale e alle distorsioni che essa può produrre, contribuendo così a una diversa cultura della trasparenza.

Abstracts

QUESTIONI – INQUIRES

Graziano Lingua & Emmanuel Alloa, *Trasparenza. Una metafora indiscutibile?* | *Transparency. An unquestionable metaphor?*

Italiano
La metafora della trasparenza è oggi utilizzata in tutti gli ambiti della vita collettiva come principio in grado di assicurare la moralizzazione dei rapporti sociali e dei comportamenti individuali. Questa ampiezza semantica nasconde però una serie di ambiguità e rischia di creare un consenso non tematizzato. L'obiettivo del saggio non è soltanto quello di offrire una griglia analitica sui molti significati che ha oggi la nozione di trasparenza, ma anche di proporre una critica all'ideologia che la avvolge. Per fare questo gli autori analizzano alcuni nuclei teorici, come la disintermediazione, l'ossessione all'esposizione online e il nesso tra trasparenza e sorveglianza per indicare dei percorsi che valorizzino la portata operativa della nozione, ma anche le resistenze possibile a una sua traduzione ideologica.

English
The metaphor of transparency is now used in all areas of collective life as a principle capable of ensuring the moralisation of social relations and individual behaviour. However, this semantic breadth conceals a number of ambiguities and risks creating an unthematic consensus. The aim of this essay is not only to provide an analytical framework for the many meanings that the concept of transparency has today, but also to propose a critique of the ideology that surrounds it. To this end, the authors analyse certain theoretical kernels, such as disintermediation, the obsession with online exposure and the link between transparency and surveillance, in order to point out ways of expanding the operational scope of the term, but also possible resistance to its ideological translation.

Gabriele Giacomini & Chiara Aprilis, *Discriminazione digitale. La sfida dei pregiudizi e della trasparenza nell'IA* | *Digital Discrimination. The Challenge of Bias and Transparency in AI*

Italiano
Il saggio esplora l'impatto crescente dell'intelligenza artificiale sui sistemi decisionali e le sue implicazioni etiche, concentrandosi sui pregiudizi algoritmici che possono portare a discriminazioni basate su genere, etnia e altri fattori. Attraverso esempi concreti, si discute di come i pregiudizi possano manifestarsi e si sottolinea l'importanza di un approccio responsabile alla governance dell'IA. Ciò implica la promozione di una riflessione sia accademica che pubblica sull'adozione di principi etici e procedure il più possibile trasparenti e inclusive.

English
This paper explores the growing impact of artificial intelligence on decision-making systems and its ethical implications, focusing on algorithmic biases that can lead to discrimination based on gender, ethnicity, and other factors. Through concrete examples, it discusses how biases may manifest and emphasises the importance of a responsible approach to AI governance. This involves promoting both academic and public reflection on the adoption of ethical principles and procedures that are as transparent and inclusive as possible.

Emanuela Tangari, *AI: inevitabile o evitabile, questo (non) è il problema. Ciò che precede la trasparenza algoritmica* | *AI: inevitable or avoidable, that is (not) the question. What precedes algorithmic transparency*

Italiano
L'articolo esplora la relazione tra Intelligenza Artificiale (IA) e fiducia, ponendo l'accento sulla trasparenza e la "trasparibilità" come elementi chiave per l'analisi di un utilizzo etico e responsabile, di cui il contesto medico si pone come caso d'uso privilegiato. Attraverso riferimenti a teorie filosofiche e psicologiche, si analizzano le sfide e le implicazioni delle decisioni autonome delle IA, mettendo in luce il loro impatto sul ragionamento umano. Viene preso in esame il progetto europeo MES-CoBraD per evidenziare i benefici e i limiti dell'applicazione dell'IA in medicina. Il tema centrale rimane la necessità di una trasparenza che superi la mera comprensione tecnica, per abbracciare una comprensione relazionale capace di sostenere una fiducia autentica e un utilizzo della ragione *tout court* nelle decisioni.

English
The article explores the relationship between Artificial Intelligence (AI) and trust, emphasizing transparency and "traceability" as key elements in analyzing the ethical and responsible use of AI, with the medical field serving as a prime use case. Drawing

on philosophical and psychological theories, it examines the challenges and implications of AI-driven autonomous decisions, highlighting their impact on human reasoning. The European MES-CoBraD project is analyzed to showcase the benefits and limitations of AI applications in medicine. The central theme remains the necessity for transparency that goes beyond mere technical understanding, embracing a relational comprehension capable of fostering genuine trust and the application of reason in decision-making.

Giovanna Di Cicco, *L'impossibilità di robot sociali trasparenti* | *The Impossibility of Transparent Social Robots*

Italiano

La trasparenza è emersa come uno dei concetti più rilevanti nel dibattito etico che circonda diversi ambiti, tra cui la robotica sociale. Questo articolo esplora il modo in cui la trasparenza si applica ai robot sociali e se possa essere uno strumento efficace per proteggere gli interessi degli utenti da potenziali inganni e dinamiche ambigue implicate nelle interazioni tra esseri umani e robot. L'articolo traccia una distinzione preliminare tra la trasparenza intesa come proprietà della robotica sociale e la trasparenza intesa come attributo dei robot sociali, evidenziandone i diversi significati e implicazioni. La discussione si concentra poi sulla trasparenza dei robot sociali e viene fatta un'ulteriore distinzione tra *trasparenza sui robot sociali* e *trasparenza attraverso i robot sociali*. Partendo dalla descrizione dei tre tipi di inganno proposti da John Danaher, l'inganno di stato interno, messo in atto da robot sociali che mostrano facoltà e stati emotivi che in realtà non hanno, viene identificato come la forma più costitutiva di inganno coinvolta nelle interazioni con i robot sociali. Questo aspetto viene poi considerato alla luce dell'antropomorfismo, per esaminare la progettazione di robot trasparenti, che dovrebbero attenuare le risposte antropomorfiche come possibile rimedio per proteggere gli interessi degli individui ed evitare l'inganno. Tuttavia, poiché l'antropomorfismo sembra essere il fondamento stesso della socialità percepita dai robot, è impossibile rinunciare al loro comportamento ingannevole senza rinunciare anche al loro ruolo sociale. Ciò porta, infine, a sostenere che un robot sociale veramente trasparente non è realizzabile e che la trasparenza non è sufficiente a garantire una robotica sociale responsabile.

English

Transparency has emerged as one of the most relevant concepts in the ethical debate surrounding several fields, and social robotics is one of them. This paper explores how transparency relates to social robots and whether it could be an effective tool to protect users' interests from potential deception and misleading dynamics involved in human-robot interactions. The paper outlines a preliminary distinction between transparency understood as a property of social robotics and transparency understood as an attribute of social robots, highlighting their different meanings and implications.

The discussion, then, focuses on the transparency of social robots, where a further distinction is drawn between *transparency on social robots* and *transparency through social robots*. Starting from the description of three types of deception proposed by John Danaher, internal state deception, enacted by social robots that display faculties and emotional states they do not really have, is identified as the most constitutive form of deception involved in interactions with social robots. This is then considered in the light of anthropomorphism, to examine the design of transparent robots, which should mitigate the anthropomorphic responses as a possible remedy to protect the interests of individuals and avoid deception. However, since anthropomorphism appears to be the very foundation of robots' perceived sociality, it is impossible to forego their deceptive behaviour without also foregoing their social role. This leads, finally, to argue that a genuinely transparent social robot is not achievable, and that transparency is not enough to ensure a responsible social robotics.

Rémy Demichelis, *« Qu'est-ce que tu ne comprends pas ? » Jeux de langage et algorithmes boîte noire* | *"What don't you understand?" Language games and black box algorithms*

Français
L'enjeu de cet article est de déterminer ce qui pose problème dans notre compréhension des algorithmes dits « boîte noire », une problématique propre à la jeune discipline de l'*Explainable Artificial Intelligence* (XAI). Car, s'il est aisé de comprendre quelque chose que quelqu'un nous explique, c'est plus délicat lorsque personne n'arrive à saisir le problème. Cependant, notre propos consiste à souligner : (1) qu'il convient de parler d'*interprétabilité* plutôt que d'*explicabilité* lorsque nous cherchons à comprendre les modèles, principalement parce que nous n'avons jamais un accès complet et sans ambiguïté à l'information ; (2) que la machine fait face au problème de l'inscrutabilité de la référence, de la même manière que le linguiste imaginé par Willard Van Orman Quine ne peut pas déterminer précisément ce que désigne le terme « *gavagai* » dans une situation de traduction radicale ; (3) qu'il n'y a pas de règle pour l'application de la langue, si ce n'est des « *language games* », comme la linguistique de Ludwig Wittgenstein nous l'enseigne. Il en découle que l'espoir d'arriver à une explicabilité des algorithmes, et donc à la transparence attendue, est sans doute vain : nous ne pouvons nous contenter que d'interprétations qui ne mentionneront jamais la règle de la règle.

English
The aim of this article is to understand the problem of "black box" algorithms, an issue inherent to the nascent field of Explainable Artificial Intelligence (XAI). While it is relatively easy to understand something someone explained to us, it becomes more complicated when no one can fully grasp the issue. Our purpose is however to highlight: (1) that we should speak of *interpretability* rather than *explainability* when we seek to understand models, mainly because we never have complete and unambiguous

access to information; (2) that the machines face the problem of the inscrutability of reference, in the same way that the linguist imagined by Willard Van Orman Quine cannot precisely determine what the term "gavagai" refers to in a situation of radical translation; (3) that there is no rule for the application of language, except for "language games", as Ludwig Wittgenstein's linguistics teaches us. The hope of achieving complete explicability and transparency of algorithms is undoubtedly in vain: we can only rely on partial and broad interpretations that will never fully explain the underlying rules.

Francesco Striano, *Il vizio della trasparenza. Etica delle virtù e fiducia nella tecnologia* | *The Vice of Transparency. A Virtue Ethics Account of Trust in Technology*

Italiano
Questo articolo esplora il rapporto tra fiducia, trasparenza e tecnologia da una prospettiva di etica delle virtù. Mette in discussione l'assunto che la trasparenza sia essenziale per la fiducia, distinguendo tra fiducia, affidamento e confidenza. La trasparenza viene poi esaminata sia come disponibilità informativa sia come processo sociale di negoziazione. L'articolo sostiene che la trasparenza nel primo senso può portare a un sovraccarico di informazioni e a dinamiche di controllo, sostenendo invece un rapporto equilibrato e virtuoso con la tecnologia che enfatizzi le capacità interpretative dell'utente. Propone che la fiducia nella tecnologia dipenda sia dagli atteggiamenti individuali sia dall'affidabilità degli oggetti. Infine, l'articolo critica il "culto della trasparenza" contemporaneo, proponendo che l'onestà, come virtù tecno-morale, sostituisca la trasparenza quale obiettivo progettuale. Le tecnologie oneste medierebbero e negozierebbero l'accesso degli utenti alle informazioni, promuovendo una fiducia autentica e sostenendo la fioritura umana.

English
This article explores the relationship between trust, transparency, and technology from a virtue ethics perspective. It challenges the assumption that transparency is essential for trust, distinguishing between trust, reliance, and confidence. Transparency is then examined as both informational openness and a social process involving negotiation. The article argues that transparency in the first sense can lead to information overload and control dynamics, advocating instead for a balanced, virtuous relationship with technology that emphasizes user interpretative skills. It proposes that trust in technology depends both on individual attitudes and objectual reliability. Finally, the article critiques the contemporary "cult of transparency," proposing that honesty, as a techno-moral virtue, should replace transparency as the design goal. Honest technologies would mediate and negotiate user access to information, fostering authentic trust and supporting human flourishing.

Giustina Benedetta Baron & Accursio Graffeo, Go Hack Yourself! *La trasparenza attraverso la lente del biohacking* | *Go Hack Yourself! Transparency Through the Lens of Biohacking*

Italiano
L'antropologia e gli studi sociali hanno ampiamente studiato le culture del self-tracking, ma i potenziali risultati del "framework del biohacking" rimangono relativamente poco esplorati. Il biohacking incarna una forma distintiva di tecno-ascetismo moderno con le sue norme uniche di autoregolazione del corpo. Come verrà chiarito, questo paradigma stabilisce nuovi "spazi di visibilità" in cui le informazioni relative al corpo e alle sue funzioni interne sono rese trasparenti, organizzate e condivise. Tuttavia, l'intricata politica che circonda la scienza aperta trascende una dicotomia semplicistica tra trasparenza e chiusura. È necessaria un'esplorazione più approfondita delle attuali trasformazioni non solo all'interno della ricerca scientifica, ma anche dei quadri epistemologici ad essa associati. Partendo da queste basi, questo studio cerca di contestualizzare gli approcci antropologici contemporanei al corpo all'interno di un panorama più ampio, esplorando il loro allineamento con modelli distinti di elaborazione delle informazioni e culture sanitarie alternative che possono influenzare le risposte tipologiche al paradigma dominante stabilito dai discorsi sul biohacking che enfatizzano la trasparenza attraverso la raccolta dei dati.

English
The realm of anthropology and social studies has extensively investigated self-tracking cultures, yet the potential outcomes of the "biohacking framework" remain relatively underexplored. Biohacking embodies a distinctive form of modern techno-asceticism with its unique norms for self-regulation of the body. As will be elucidated, this paradigm establishes novel "spaces of visibility" where information regarding the body and its internal functions is rendered transparent, organized, and shared. Nonetheless, the intricate politics surrounding open science transcend a simplistic dichotomy between transparency and closure. It necessitates a more profound exploration of current transformations not only within scientific research but also concerning its associated epistemological frameworks. Building upon these foundations, this study seeks to contextualize contemporary anthropological approaches to the body within a broader landscape, exploring their alignment with distinct models of information processing and alternative health cultures that may influence typological responses to the dominant paradigm set forth by biohacking discourses emphasizing transparency through data collection.

Matteo Cresti, *Il valore morale della trasparenza nell'uso delle* Performance Enhancing Drugs*. Il caso del bodybuilding* | *The Moral Value of Transparency in the Use of Performance Enhancing Drugs. The Case of Bodybuilding*

Italiano
L'articolo ha l'obiettivo di sostenere il valore morale positivo della trasparenza riguardo all'assunzione di Performance Enhancing Drugs (PED) nel bodybuilding. Per prima cosa darò una definizione di trasparenza adeguata all'ambito sportivo. In secondo luogo descriverò l'uso di PED nel bodybuilding, in particolare di steroidi anabolizzanti, mostrando come negli ultimi anni si possa registrare un fenomeno di rivelazione dell'uso di PED. Proporrò poi il mio argomento in difesa della trasparenza sull'assunzione di PED basato su considerazioni consequenzialiste. La tesi è che i bodybuilder che rivelano di fare uso di PED stiano compiendo un'azione moralmente positiva, in quanto consentono a chi si ispira a loro come modelli di ricalibrare le proprie aspettative e di fare scelte più informate. Infine risponderò all'obiezione che questa pratica possa incentivare l'uso di PED.

English
The paper aims to support the positive moral value of transparency regarding the intake of Performance Enhancing Drugs (PEDs) in bodybuilding. First, I will adequately define transparency for the sports sector. Secondly, I will describe the use of PEDs in bodybuilding, in particular of anabolic steroids, showing how, in recent years, there has been a phenomenon of disclosure of the use of PEDs. I will then propose my argument in defense of transparency on PEDs intake based on consequentialist considerations. The thesis is that bodybuilders who reveal that they use PEDs are doing a morally positive action, as they allow those who look up to them as role models to recalibrate their expectations and make more informed choices. Finally, I will respond to the objection that this practice could encourage the use of PEDs.

Richard Davies, *Ora mi vedi, ora non mi vedi più. La sorte di Gige nella Repubblica di Platone* | *Now you see me, now you don't. The predicament of Gyges in Plato's Republic*

Italiano
In questo saggio esaminiamo il caso di un oggetto trasparente, inteso nella sua accezione fisica di base, cioè tale che la luce lo attraversa in modo da renderlo invisibile. Il caso centrale che consideriamo è quello di Gige, raccontato all'inizio del Libro II della *Repubblica* di Platone. Trattiamo questa narrazione come se rendesse evidente un caso estremo di impunità e le sue conseguenze, e cerchiamo di tenere conto di alcuni aspetti del *topos* degli agenti invisibili che ha visto una rinascita nell'ultimo secolo e mezzo. Dopo un breve sguardo a come gli esperimenti di pensiero figurano

nell'argomentazione filosofica, notiamo alcune varianti nelle storie associate al nome di Gige. I due punti principali che ci proponiamo di evidenziare sono, in primo luogo, che l'opportunità di non essere visibile a piacimento che l'anello di Gige conferisce è in contrasto con la sua capacità di essere un agente efficace, perché sarà cieco, e, in secondo luogo, che gli svantaggi di tale opportunità possono, nel complesso, superare i vantaggi, perché perde il rispetto per se stesso e per coloro che lo circondano.

English
In this essay, we look at a case of a transparent object taken in its basic, physical sense of being such that light passes through it so as to make it invisible. The central case we consider is that of Gyges, as recounted at the outset of Book II of Plato's *Republic*. We treat this narrative as making vivid an extreme case of impunity and its consequences, and we try to take account of some aspects of the *topos* of invisible agents that has seen a revival in the last century and a half. After a brief look at how thought experiments figure in philosophical argumentation, we note some of the variants in the stories associated with the name of Gyges. The two main points we aim to bring out are, first, that the opportunity not to be visible at will that Gyges' ring confers is at odds with his being an effective agent because he will be blind, and, second, that the disadvantages of such an opportunity may, overall, outweigh the advantages because he loses respect for himself and those around him.

RICERCHE – RESEARCHES

Paolo Monti, Graziano Lingua & Philippe Poirier, *Rappresentanza democratica e processo decisionale al tempo della disintermediazione digitale. Una critica all'erosione populista del ruolo dei Parlamenti* | *Democratic Representation and Decision-making at the Time of Digital Disintermediation: A Critique of the Populist Erosion of the Role of Parliaments*

Italiano
Diversi autori hanno analizzato l'ascesa dei movimenti populisti in tutto il mondo come un fenomeno che deve essere inquadrato nel contesto di una trasformazione generale della democrazia rappresentativa in una forma di democrazia del pubblico, in cui il valore dell'intermediazione è sempre più contestato a tutti i livelli della vita sociale. Nell'esaminare questo cambiamento in corso, illustriamo innanzitutto alcune implicazioni generali che i fenomeni sociali di disintermediazione hanno per la pratica della democrazia rappresentativa, assottigliando e rimodellando i confini tra la sfera pubblica formale e quella informale. In particolare, esaminiamo la crescente influenza della leadership carismatica nella politica dei partiti e la spinta alla democrazia diretta digitale come alternativa al ruolo delle assemblee elettive, per mostrare come un ideale

normativo di rappresentanza politica come specchio in tempo reale dell'opinione pubblica sia alla base di entrambe queste strategie populiste. Valutiamo poi criticamente queste implicazioni pratiche e teoriche della disintermediazione. Da un punto di vista pratico, scopriamo che le leadership carismatiche e le strategie populiste di democrazia digitale diretta non soddisfano gli standard di immediatezza e trasparenza su cui si basano e non possono sostituire la funzione democratica pluralistica delle assemblee elettive. Da un punto di vista teorico, sosteniamo che la premessa concettuale su cui si basano è fondamentalmente errata: la rappresentanza politica è un processo che implica sempre un grado rilevante di interpretazione e intermediazione, e pertanto le affermazioni dei rappresentanti non possono essere interpretate come riflessi speculari dei rappresentati. Concludiamo suggerendo che i parlamenti dovrebbero invece adottare pratiche innovative come le audizioni pubbliche e la democrazia diretta avviata dai cittadini, che ricentrano la funzione rappresentativa dell'assemblea sull'ascolto attivo dei rappresentanti e sulla partecipazione dei rappresentati.

English

Several authors have analyzed the rise of populist movements around the world as a phenomenon that must be seen in the context of a general transformation of representative democracy into a form of audience democracy, in which the value of intermediation is increasingly contested at all levels of social life. In examining this ongoing shift, we first illustrate some general implications that social phenomena of disintermediation have for the practice of representative democracy by thinning and reshaping the boundaries between the formal and informal public spheres. Specifically, we examine the growing influence of charismatic leadership in party politics and the push for digital direct democracy as an alternative to the role of elected assemblies, to show how a normative ideal of political representation as a real-time mirroring of public opinion underpins both of these populist strategies. We then critically assess these practical and theoretical implications of disintermediation. From a practical perspective, we find that charismatic leaderships and direct digital democracy populist strategies do not meet the standards of immediacy and transparency on which they are based, and cannot replace the pluralistic democratic function of elected assemblies. From a theoretical perspective, we argue that the conceptual premise on which they rely is fundamentally flawed: political representation is a process that always involves a relevant degree of interpretation and intermediation, and therefore representative claims cannot be construed as mirror reflections of the represented. We conclude by suggesting that parliaments should instead adopt innovative practices such as public hearings and citizen-initiated direct democracy, which refocus the representative function of the assembly on the active listening of the representatives and the participation of the represented.

Nicola Pedretti, *Trasparenza e democrazia monitorante. La trasparenza integrale come occasione di partecipazione dei cittadini* | *Transparency and monitoring democracy. Full transparency as an opportunity for citizen participation*

Italiano
La garanzia di una reale trasparenza della pubblica amministrazione rappresenta indubbiamente un'occasione di partecipazione della cittadinanza alla gestione della *res publica*. In quest'ottica, va osservato come la normativa italiana abbia introdotto gradualmente strumenti di accesso civico che hanno reso maggiormente fruibili dati e informazioni ai cittadini. In tale ambito, verrà esaminato il caso studio del rapporto Rimandati che applica il community-based monitoring al delicato settore dei beni confiscati.

English
The guarantee of real transparency of the public administration undoubtedly represents an opportunity for citizen participation in the management of the *res publica*. From this perspective, it should be noted that Italian legislation has gradually introduced civic access tools that have made data and information more accessible to citizens. In this context, the case study of the Rimandati report will be examined, which applies community-based monitoring to the delicate sector of assets confiscated.

Giulia Miotti, *La trasparenza nei mercati finanziari: approccio classico e nuovi paradigmi* | *Transparency in Financial Markets: Classical Approach and New Paradigms*

Italiano
Il sistema finanziario è un sistema sociale molto complesso con una forte ramificazione all'interno di tutti gli altri sistemi sociali dei Paesi avanzati: ha un impatto fortissimo sull'economia reale e i tempi degli scambi finanziari hanno ormai determinato un'accelerazione anche nei tempi della vita sociale anche fuori dai mercati. Questo sistema è un sistema sociale anche perché basa il proprio funzionamento sulla ricerca e lo scambio di informazioni; eppure, sebbene l'informazione sia un concetto cardine all'interno della pratica e delle teorie dei mercati finanziari, questa sembra slegata dal concetto di trasparenza intesa come apertura, comunicazione e accountability. Questa mancata corrispondenza produce degli effetti rilevanti all'interno dei mercati e nella possibilità di una scelta informata ed equa degli agenti del mercato. Vedremo due possibili risposte a questa disfunzione interna ai mercati, una di ordine politico e l'altra di governance.

English
The financial system is a highly complex social system, deeply entangled with all other social systems in developed Countries. It exerts a disruptive impact on the real economy sector and the speed of financial exchanges seems to have determined a similar acceleration also in the speed and nature of social life itself. Another reason why the financial system can be considered a social system lies in the fact that the financial system ground its functioning in the research and exchange of information. In this context, information represents a pivotal concept around which financial practice and financial theories alike turn. Notwithstanding this, the notion of information seems detached from the notion of transparency meant as openness, communication and accountability. The lack of such correspondence engenders critical effects on financial markets, especially when it comes to the possibility of making fair and informed choices by market agents. We shall describe and discuss two possible alternative scenarios for such market dysfunction; the first one is of a political kind, the second one of a governance-oriented one.

Gian Vito Zani, *Delle trasparenze economiche* | *On economic transparency*

Italiano
Il presente articolo esplora il concetto di trasparenza nell'ambito economico, concentrandosi sulle prospettive offerte dalla Scuola Austriaca di economia, qui rappresentata da Ludwig von Mises e Friedrich von Hayek. L'analisi si articola in tre parti: la prima introduce il ruolo centrale della trasparenza nel discorso economico e nei suoi dibattiti sulle crisi. La seconda esamina la teoria di Mises, per cui la trasparenza del mercato è essenziale in quanto qualsiasi interferenza statale rappresenta una distorsione delle informazioni e delle dinamiche economiche. La terza si concentra sull'approccio di Hayek, per il quale, diversamente da Mises, il mercato opera in un contesto di opacità intrinseca, dove la conoscenza è dispersa e spesso tacita. Questi autori offrono interpretazioni opposte della trasparenza per giungere al medesimo obiettivo, cioè, rendere il mercato un'istituzione incontestabile. Il lavoro evidenzia come attraverso il binomio trasparenza/opacità, si siano sviluppati strumenti concettuali per legittimare il primato del mercato, negando al contempo validità alle possibili alternative.

English
This article explores the concept of transparency in economics, focusing on the perspectives offered by the Austrian School of Economics, as represented by Ludwig von Mises and Friedrich von Hayek. The analysis is divided into three parts: the first introduces the central role of transparency in economic discourse and its debates on crises. The second examines Mises's theory, which posits that market transparency is essential, as any state interference distorts information and economic dynamics. The third focuses on Hayek's approach, which, in contrast to Mises, views the market as

operating in a context of intrinsic opacity, where knowledge is dispersed and often tacit. These authors provide opposing interpretations of transparency to reach the same goal: making the market an uncontestable institution. The study highlights how the transparency/opacity dichotomy has been used to develop conceptual tools that legitimize the primacy of the market while simultaneously dismissing the validity of alternative models.

Trasparenza.
Una metafora indiscutibile?[a]

Graziano Lingua[*] e Emmanuel Alloa[†]

Abstract

La metafora della trasparenza è oggi utilizzata in tutti gli ambiti della vita collettiva come principio in grado di assicurare la moralizzazione dei rapporti sociali e dei comportamenti individuali. Questa ampiezza semantica nasconde però una serie di ambiguità e rischia di creare un consenso non tematizzato. L'obiettivo del saggio non è soltanto quello di offrire una griglia analitica sui molti significati che ha oggi la nozione di trasparenza, ma anche di proporre una critica all'ideologia che la avvolge. Per fare questo gli autori analizzano alcuni nuclei teorici, come la disintermediazione, l'ossessione all'esposizione online e il nesso tra trasparenza e sorveglianza per indicare dei percorsi che valorizzino la portata operativa della nozione, ma anche le resistenze possibile a una sua traduzione ideologica.

Parole chiave: Trasparenza, disintermediazione, sorveglianza, datificazione

The metaphor of transparency is now used in all areas of collective life as a principle capable of ensuring the moralisation of social relations and individual behaviour. However, this semantic breadth conceals a number of ambiguities and risks creating an unthematic consensus. The aim of this essay is not only to provide an analytical framework for the many meanings that the concept of transparency has today, but also to propose a critique of the ideology that surrounds it. To this end, the authors analyse certain theoretical kernels, such as disintermediation, the obsession with online exposure and the link between transparency and surveillance, in order to point

[a] Questo saggio è un prodotto delle ricerche e delle collaborazioni condotte all'interno del progetto PRIN 2022 "Social Transformations & the Crisis of Expertise" (2022JR8Z8P) finanziato dall'Unione Europa – Next Generation EU. Saggio ricevuto in data 31/05/2024 e pubblicato in data 22/01/2024.

[*] Professore ordinario, Università di Torino, email: graziano.lingua@unito.it.

[†] Professore ordinario, Université de Fribourg, email: emmanuel.alloa@unifr.ch.

out ways of expanding the operational scope of the term, but also possible resistance to its ideological translation.

Keywords: Transparency, disintermediation, surveillance, datafication

1. Su un concetto "magico", ma ambiguo

Il consenso che accompagna l'idea di trasparenza è oggi sintomo di una vera e propria ossessione collettiva. Parola d'ordine tra le più abusate, è sulla bocca di politici, di amministratori delegati, di *opinion-makers* di ogni sorta, finanche di leader religiosi, quasi non ci fosse principio più evidente e immediato. Chi si permette di criticarla viene etichettato come un "reazionario" o comunque come qualcuno che ha privilegi da difendere e zone franche da sottrarre al confronto e al controllo pubblico. Tutti conoscono il mantra che accompagna questa ossessione: "Se non si ha niente da nascondere, non si ha nulla da temere". Si dà per scontato che l'apertura e la piena accessibilità delle informazioni conducano automaticamente al progresso sociale e a una moralizzazione individuale e collettiva. Quindi tutto deve essere visibile e sotto i riflettori, perché ciò che è segreto nasconde sicuramente manipolazione e inganno. Insomma: «La luce del sole è il miglior disinfettante», come ebbe a dire già negli anni '10 dello scorso secolo il giudice della Suprema Corte americana Louis Brandeis[1].

Dal punto di vista storico, la nozione di trasparenza viene per lo più collegata all'ambito politico del rapporto con il potere, ed è su questo versante che il termine si è costruito nella modernità, attraverso una complessa serie di stratificazioni concettuali che ne hanno determinato il progressivo successo. L'idea che ne sta alla base è semplice: meno opacità e segreti ci sono nelle procedure e nei sistemi politici, più grandi sono la libertà e la possibilità di partecipazione democratica. Questo legame diretto con le dinamiche di controllo democratico del potere fa della trasparenza una delle eredità più vive della tradizione illuminista. Essa rappresenta oggi una specie di sintesi residuale del progetto di emancipazione lanciato nel XVIII secolo, che aveva i propri cavalli di battaglia nei valori del rischiaramento razionale, dell'autonomia individuale e dell'universalità delle norme. Mentre molti di questi valori hanno perso di mordente nel Novecento, scontrandosi con la crisi della ragione e il ritorno dei totalitarismi, così non è stato per la trasparenza, che ha continuato la sua marcia trionfale. Anzi, essa è stata assunta e fatta propria esplicitamente come valore politico anche da regimi tutt'altro che democratici[2].

[1] L. Brandeis, *What Publicity Can Do,* in «Harper's Weekly», 20 dicembre 1913, p. 10.

[2] Va notato, per esempio, che al valore della trasparenza fa riferimento Benito Mussolini quando definisce il fascismo come "una casa di vetro in cui tutti possono guardare", per non parlare del ruolo che ha avuto l'architettura del vetro nella Germania Nazista e nel primo periodo del sistema sovietico, prima del trionfo del realismo socialista. Per approfondire questi aspetti rimandiamo a E. Alloa, *Une transparence révolutionnaire. Le rêve d'une société perméable*, in C. Beaufort, B. Rougé (a cura di), *Transparence/Transparaître*, Presses universitaires de Rennes, Rennes 2023, pp. 39-63.

Anche se nell'epoca dei Lumi il termine, nella sua accezione attuale, è praticamente assente[3], le tesi illuministiche sul potere emancipativo della ragione e sulla centralità della "pubblicità" del governo hanno contribuito in modo decisivo alla formazione della nozione. La sensibilità all'uso pubblico della ragione ha rappresentato un duraturo vettore concettuale che ha attraversato gran parte della tradizione politica liberal-democratica occidentale, fino alle proposte più recenti di filosofi come John Rawls e Jürgen Habermas[4]. Il tratto fondamentale questa tradizione di pensiero politico è stata l'importanza di coinvolgere i cittadini nella vita pubblica in quanto soggetti che non sono soltanto sottoposti alle leggi, ma che devono poter partecipare alla loro creazione.

Accanto a questo profilo politico, da alcuni decenni si è imposta però una progressiva dilatazione semantica e concettuale del termine, che tocca contesti sempre più ampi della vita collettiva. Dall'ambito delle procedure pubbliche, la richiesta di trasparenza ha invaso per esempio il settore aziendale divenendo un fulcro della logica organizzativa, come testimonia una ricerca condotta nel 2011 tra i principali CEO di compagnie internazionali, secondo cui "autenticità" e "trasparenza" sono le due parole chiave più utilizzate nelle loro organizzazioni[5]. Allo stesso modo la trasparenza si è trasformata in un imperativo che legittima la fiducia reciproca tra gli attori sociali, fino a entrare nell'organizzazione degli spazi della vita quotidiana. Ci appelliamo alla trasparenza senza soluzione di continuità per connotare pratiche di ogni tipo e finiamo per considerarla un presupposto fondamentale anche per sentirci a nostro agio negli ambienti che frequentiamo: uffici *open space*, edifici aziendali in vetro, cucine di ristoranti dove il cibo viene preparato in bella vista, tutto deve tendere alla totale visibilità e alla massima esposizione.

Ancora più rilevante però è l'espansione della richiesta di trasparenza nei confronti dei singoli, delle loro condotte e della loro vita personale. Anche gli aspetti più intimi e privati subiscono la pressione a essere resi totalmente accessibili con il presupposto che soltanto se si rende tutto visibile a tutti ci si può sentire "autentici". Dave Eggers, nel suo romanzo *Il Cerchio*[6], ambientato in una azienda hi-tech

[3] Sulle differenze tra la concezione attuale della trasparenza e la nozione illuministica di "pubblicità", si vedano: S. Baume, *Publicity and Transparency: The Itinerary of a Subtle Distinction,* in E. Alloa, D. Thomä (a cura di)*, Transparency, Society and Subjectivity. Critical Perspectives,* Palgrave Macmillan, Londra 2022, pp. 203-224, J. Pitseys, *Transparency, Publicity, Secrecy and Mendacity. Four Shades of Political Visibility,* in E. Alloa (a cura di), *This Obscure Thing Called Transparency: Politics and Aesthetics of a Contemporary Metaphor*, Leuven University Press, Leuven 2022, pp. 146-152.

[4] Si pensi solo a titolo di esempio al dibattito sulla ragione pubblica che ha coinvolto i due filosofi nel 1995 in «The Journal of Philosophy», vol. 92. Per un'analisi di questa discussione si veda: J. Gordon Finlayson, *The Habermas-Rawls Debate*, Columbia University Press, New York 2019.

[5] Cfr. Guo Wei, *The power of plain language: Executives' rhetoric and stock market reaction* in «Academy of Management Proceedings», XII, n. 1, 2012, p. 18185.

[6] D. Eggers, *Il Cerchio*, trad. it. di V. Mantovani, Mondadori, Milano 2014. Non citiamo a caso questo romanzo che David Lyon considera il corrispondente contemporaneo di *1984* di George Orwell. Se infatti quest'ultimo «ci offriva i concetti con cui valutare la sorveglianza di Stato del Novecento*, Il Cerchio* è un candidato adatto per esaminare la cultura della sorveglianza del Ventunesimo secolo» (D.

contemporanea, esprime in modo incisivo questo bisogno di esposizione totale con uno slogan che fa eco niente meno che a Proudhon: "la privacy è un furto"[7]. Tenere qualche cosa per sé corrisponde a sottrarlo alla collettività, nascondere elementi della propria vita individuale significa privare gli altri di informazioni che potrebbero essere loro utili o addirittura necessarie.

Il romanzo di Eggers, in termini che oscillano tra la realtà e la distopia, fa dell'esposizione totale online la ragione di vita della protagonista Mae, offrendoci un'espressione letteraria efficace dell'immaginario che accompagna la dilatazione della trasparenza alla vita dei singoli. È affidabile soltanto chi si consegna totalmente alla visibilità pubblica, perché chi ha segreti nasconde qualcosa di sbagliato e non può quindi meritare la fiducia degli altri. Eggers offre così una serie di spunti significativi per comprendere le pratiche di esposizione che caratterizzano gli ambienti digitali e i social network in particolare. Alla totale accessibilità in rete *il Cerchio* attribuisce poi una chiara valenza etica: di fronte allo sguardo pubblico il comportamento individuale è spinto automaticamente a moralizzarsi, e questo non vale soltanto per chi ha cariche e responsabilità pubbliche, ma per ogni persona, in ogni situazione della nostra vita.

Nell'ossessione espositiva sui social network c'è però un elemento che va oltre quella retorica morale di cui parla *Il Cerchio* e coinvolge una disposizione antropologica più originaria, non dettata questa volta dall'esigenza eteronoma di adeguamento a pressioni sociali, ma da una scelta personale. A un primo livello questa scelta ha a che fare con il prestigio che deriva dalla visibilità e dai feedback che otteniamo. Pubblichiamo momenti importanti delle nostre vite private perché così facendo dimostriamo che sono realmente accaduti e monetizziamo questa presenza sul Web in termini di ritorno di immagine. Ma vedremo in seguito che questo bisogno di esposizione ha una radice antropologica più profonda che coinvolge la ricerca di riconoscimento e le dinamiche di appagamento del desiderio.

Sta di fatto che l'esposizione online rappresenta un'espressione eloquente dell'espansione semantica della trasparenza e mostra come intorno a essa si sia creata una vera e propria aura sacrale che l'ha trasformata in un "concetto magico", secondo l'espressione che i politologi americani Christopher Pollitt e Peter Hupe utilizzano per qualificare le nozioni politiche con «un alto grado di astrazione, una carica normativa fortemente positiva, un'apparente capacità di sciogliere i precedenti dilemmi e una mobilità tra domini»[8]. Come altri termini oggi abusati nel discorso pubblico – si pensi a "governance", "reti", "partecipazione", "innovazione", solo per citarne alcuni –,

---

Lyon, *La cultura della sorveglianza. Come la società del controllo ci ha reso tutti controllori*, Luiss University Press, Roma 2020, p. 156).

[7] D. Eggers, *Il Cerchio*, cit., p. 275. Per una analisi più approfondita del romanzo si veda M. Carbone, G. Lingua, *Antropologia degli schermi. Mostrare e nascondere, esporre e proteggere*, Luiss University Press, Roma 2024. Nel presente saggio riformuliamo anche alcuni temi già presenti nel Cap. IV di quest'opera. Ringraziamo Mauro Carbone per aver permesso di riprendere e rielaborare il materiale in oggetto.

[8] C. Christopher, P. Hupe*, Talking About Government. The Role of Magic Concepts*, in «Public Management Review», XIII, 2011*,* p. 641, traduzione nostra. Su questo si veda E. Alloa, *Transparency: A Magic Concept of Modernity*, in E. Alloa, D. Thomä (a cura di)*, Transparency, Society and Subjectivity*, cit., pp. 21-25.

anche la trasparenza ha questa capacità di imporsi come intrinsecamente buona e appropriata per risolvere, attraverso una presunta eliminazione delle opacità, questioni che diversamente apparirebbero a prima vista indecidibili.

Tutti questi caratteri che sembrano fare della trasparenza un dato indiscutibile rendono però estremamente vaga questa nozione e il fatto che la possiamo utilizzare in situazioni così diverse denota una forte instabilità semantica. Che cos'è allora veramente in gioco quando parliamo di trasparenza e quali sono i confini che possono rendere questo concetto non solo comprensibile, ma anche utilizzabile?

2. La natura della trasparenza e i suoi diversi significati

Per far fronte a questa indecidibilità è utile quindi partire da alcune distinzioni analitiche che possono aiutarci a orientare il discorso. Intanto è bene ricordare che gli usi del termine "trasparenza" a cui ci siamo riferiti fin qui sono metaforici, poiché il contesto semantico originario della parola è legato alla fisica ottica e indica una specifica qualità di alcuni materiali in grado di permettere il passaggio della luce[9]. Questo significato originario rimanda quindi a uno stato fisico e si differenzia dall'uso in ambito sociale e nel linguaggio comune, dove invece la qualità della trasparenza si riferisce piuttosto a una condizione dinamica, mai definitivamente stabilita, bensì costantemente da raggiungere e mantenere.

L'uso metaforico si differenzia poi secondo diversi assi concettuali. Ne possiamo ricordare almeno quattro che rappresentano due coppie tematiche fondamentali, anche se questa quadripartizione analitica non esaurisce certo tutti i significati possibili del termine, ma si rivelerà funzionale all'analisi critica di alcuni suoi usi particolarmente rilevanti nel dibattito pubblico odierno. Sulla scia di David Heald ed altri, proponiamo di distinguere tra una *trasparenza verticale*, che descrive una qualità propria del rapporto politico tra la sfera sociale e il potere, con le sue istituzioni, strutture e procedure, e una *trasparenza orizzontale*, che raccoglie invece più direttamente le modalità della comunicazione intersoggettiva e la qualità delle relazioni sociali[10]. L'asse verticale coinvolge la critica a quelli che già Tacito chiamava gli *arcana imperii*, ovvero i segreti con cui il governo può gestire il potere senza rendere conto ai cittadini e punta quindi all'accessibilità delle informazioni in mano a chi governa per conoscere le ragioni delle decisioni prese, poter partecipare alle stesse ed essere in grado di esercitare un controllo. All'asse orizzontale possiamo invece riferire la richiesta di rimuovere gli ostacoli alla diffusione universale della conoscenza, così da

[9] Cfr. A. Alloa, *Attraverso l'immagine*, trad. it. di A. De Cesaris, Meltemi, Sesto San Giovanni 2025, pp. 185-267.

[10] Cfr. David Heald, *Varieties of Transparency* in Ch. Hood, D. Heald, *Transparency: The Key to Better Governance?,* Oxford University Press, Oxford, 2006, pp. 27-29. Per un'analisi più diffusa di questa distinzione che ha nel nostro caso un significato differente da quello attribuitogli da Heald si veda G. Lingua, *Transparence numérique et frontières de la désintermédiation politique*, in J. Bodini, M. Carbone, G. Lingua, G. Serrano (a cura di), *L'avenir des écrans*, Éditions Mimésis, Parigi 2020, pp. 193-205.

consentire una partecipazione la più ampia possibile alla vita sociale e culturale in tutte le loro forme.

Sia chiaro: questa è una mera distinzione "idealtipica" perché le due direzioni della trasparenza si sono storicamente mescolate tra loro, convergendo però sull'idea, di evidente origine illuministica, che il rischiaramento prodotto dalla conoscenza produca partecipazione, condivisione del sapere e, in ultimo, emancipazione dei soggetti. Va notato poi che la svolta digitale degli ultimi decenni ha finito per sovrapporre completamente questi due assi facendoli convergere nella convinzione che l'accesso potenzialmente universale ai dati offerti dalle piattaforme del web e dai social network consenta automaticamente una maggiore partecipazione alle decisioni pubbliche, quindi un controllo del potere da parte dei cittadini e una crescita della qualità democratica della sfera pubblica.

Proprio gli ambienti digitali evidenziano però come questi due assi siano oggi attraversati da una seconda coppia tematica che sta diventando sempre più rilevante. Se storicamente la metafora che ci interessa è stata abitualmente associata, come abbiamo visto, a pratiche collettive di natura politica e sociale, oggi la utilizziamo sempre più spesso in relazione alla sfera personale. Distinguere una *trasparenza personale* e una *trasparenza collettiva* ci consente quindi di mettere in evidenza come la massiccia esposizione delle vite individuali online non è soltanto una trasposizione a livello privato delle dinamiche della trasparenza collettiva, bensì ne incarna una concezione specifica. La scelta di rendere visibili le proprie esperienze, anche quelle più intime e private, non è infatti l'effetto di una costrizione esterna o di un adeguamento a norme stabilite, ma sembra piuttosto il risultato di quello che Byung-Chul Han definisce «un bisogno auto-prodotto […] di esporsi»[11].

Bernard Harcourt, nel suo libro significativamente intitolato *Exposed*[12], individua la radice di tale bisogno in una dinamica di appagamento profondo del desiderio. Secondo Harcourt, sulle piattaforme social condividiamo le nostre vicende più intime perché carichiamo questi ambienti digitali di un forte investimento pulsionale, tanto che non siamo più in grado di cogliere i pericoli insiti nel rendere pubblica la vita privata e diventiamo perciò «schiavi […] dei nostri desideri di condivisioni, click, amici, e "likes"»[13]. Harcourt, sulla scia delle analisi contenute in *L'anti-Edipo* di Gilles Deleuze e Félix Guattari[14], interpreta questa pulsione che ci incolla ai nostri dispositivi come un sintomo della macchina del capitalismo che «libera […] i flussi del desiderio»[15]. Così facendo, però, egli finisce per ricondurre univocamente la motivazione dell'esposizione alle pressioni eteronome proprie del sistema capitalistico, mentre a nostro parere questo bisogno auto-prodotto di esporsi

[11] B.-C. Han, *La società della trasparenza*, trad. it. di F. Buongiorno, Nottetempo, Milano 2014, p. 78.
[12] B. Harcourt, *Exposed: Desire and Disobedience in the Digital Age*, Harvard University Press, Cambridge (MA) 2015.
[13] Ivi, p. 228, traduzione nostra.
[14] G. Deleuze, F. Guattari, *L'anti-Edipo. Capitalismo e schizofrenia*, trad. it. di A. Fontana, Einaudi, Torino 2002, pp. 3-53.
[15] Ivi, p. 154.

può esprimere anche una richiesta di riconoscimento della propria singolarità e un desiderio di autenticità personale che si inseriscono bene nella generale sensibilità individualistica della tarda modernità. Ha ragione quindi David Lyon a osservare che questa ricerca di trasparenza insita nel bisogno di essere presenti online «minimizza la "disciplina" […] mettendo in primo piano la "performance" del singolo»[16]. Da questo punto di vista la smania di mostrare la propria vita online non deriva unicamente da pressioni ambientali o dall'interesse a inseguire il prestigio sociale, ma anche dall'importanza che ha nel contesto contemporaneo il fatto di farsi riconoscere per quello che si è[17].

Risulta quindi improprio, a nostro parere, ricondurre questa ricerca di visibilità online unicamente a una forma di vuoto narcisismo[18]. In molti casi, anzi, tale esposizione è uno strumento di soggettivazione per generazioni di nativi digitali che hanno ormai trasferito sul web una fetta importante delle loro relazioni sociali, o contribuisce alla costruzione dell'identità di gruppi minoritari che non potrebbero esprimere altrove le proprie idee e trovano così il modo di condividere convinzioni e pratiche che non avrebbero luoghi più tradizionali per manifestarsi[19].

Tuttavia è interessante notare che questo bisogno di esporsi, qualsiasi ne siano le ragioni, esemplifica a livello personale una ambivalenza tipica di tutta la cultura della trasparenza negli ambienti digitali. L'immediatezza che il web sembra offrirci e l'apparente percezione di "presenza" che sperimentiamo davanti agli schermi nasconde in realtà una serie di contraddizioni su cui dovremo ritornare. Per ora ci limitiamo a notare che l'ossessione di esporsi online*, accanto a una spinta sincera all'espressione personale, rischia di alimentare* un "conformismo compulsivo" che, sottolinea sempre Byung-Chul Han, stabilizza il "sistema dominante"[20] e spinge ad adattarsi a esso.

Se a livello individuale tale conformismo può condizionare l'espressione di sé che avviene nel Web, a livello collettivo esso produce paradossi e ambiguità anche più significative. Non si può dimenticare per esempio che alcune pratiche di trasparenza collettiva, come quelle di rendere indiscriminatamente accessibili le procedure politiche o far firmare lunghi disclaimer legali a tutela della privacy, pur venendo

---

16 Cfr. D. Lyon, *La cultura della sorveglianza. Come la società del controllo ci ha resi tutti controllori*, trad. it. di C. Veltri, Luiss University Press, Roma 2020, pp. 28; 121 e ss.

17 Su questo si veda Ch. Taylor*, L'età secolare*, trad. it. di P. Costa, M.C. Sircana, Feltrinelli, Milano 2009, pp. 595 ss. Taylor propone un'ampia disamina dell'immaginario dell'autenticità generatosi in Occidente a partire dagli anni sessanta. Anche se non si occupa dell'impatto degli ambienti digitali, è evidente che l'argomento del "nulla da nascondere" può risultare significativo soltanto in un contesto di ricerca ossessiva dell'autenticità.

18 Rispetto al ruolo del narcisismo nell'individualismo contemporaneo cfr. G. Lipovetsky, *L'era del vuoto. Saggi sull'individualismo contemporaneo*, trad. it. di P. Peroni e G. Caviglione, Luni, Sesto San Giovanni 2013. Per quanto riguarda più specificamente gli ambienti digitali si veda: C.S. Andreassen, S. Pallesen, M.D. Griffiths, *The Relationship Between Addictive Use of Social Media, Narcissism, and Self-esteem: Findings from a Large National Survey*, in "Addictive Behaviors", n. 64, 2017, pp. 287–293.

19 Al riguardo si veda M. Carbone, G. Lingua, *Antropologia degli schermi*, cit., pp. 149-150.

20 B.-C. Han, *La società della trasparenza*, cit., p. 10.

incontro a diritti reali dei cittadini, possono anche creare una cortina fumogena che devia l'attenzione e tiene sottotraccia questioni più sostanziali. È facile infatti sommergere informazioni potenzialmente compromettenti in un diluvio di dati e di clausole che spesso non vengono neanche lette. Insomma, la retorica della piena trasparenza può distogliere lo sguardo dai rischi e dalle responsabilità reali, dando una patina di correttezza a pratiche che non lo sono affatto, tanto che non è privo di fondamento parlare, come fanno alcuni, di un *transparency washing*[21] in linea con le diverse forme di *ethics washing* che sono sempre più diffuse a livello sia pubblico che privato.

3. Dentro le contraddizioni: l'ideologia della trasparenza

La disinvoltura con cui ci appelliamo alla trasparenza in ambiti così diversi tra loro, e la convinzione che essa sia di per sé in grado di spiegare il modo in cui la società deve funzionare[22], non possono farci dimenticare gli effetti distorsivi che anche solo quell'aura di intoccabilità porta con sé. Anzi, proprio il fatto che essa si imponga come qualcosa di indiscutibile è già in realtà parte integrante dell'ideologia che la sostiene. Da questo punto di vista, tematizzare questa presunta indiscutibilità della trasparenza può costituire un primo passo in direzione di una critica della concezione dominante e rappresentare il presupposto su cui costruire un discorso alternativo, maggiormente consapevole delle trappole che si nascondono nella sua celebrazione incondizionata.

Per affrontare quest'ultimo aspetto è però necessaria una premessa: parlare di una critica dell'ideologia qui non significa rincorrere una forma autentica di trasparenza, semplicemente smascherando il suo uso dogmatico o le dinamiche di potere implicate, né demonizzare la sua funzione sociale e politica, dimenticando che essa costituisce un ingrediente importante per la vita dei singoli e per i processi democratici collettivi. Ciò che va messo a tema è piuttosto la pretesa di "neutralità ideologica" che si nasconde dietro l'apparente indiscutibilità e che finisce per capovolgersi invece in una "ideologia della neutralità"[23]. In un contesto post-ideologico la trasparenza si presenta come un valore che non ha connotazioni di parte e può quindi ambire a un'universalità che altri valori, come la libertà o l'equità, non raggiungono perché hanno alle spalle una lunga storia di conflitti culturali e politici. Tuttavia la tendenza della trasparenza a imporsi come "concetto magico" sottratto a ogni contraddittorio contribuisce a ipostatizzarne un profilo che solo apparentemente neutralizza quei conflitti, perché in realtà semplicemente li rimuove.

---

[21] Si veda per esempio M. Zalnieriute, "*Transparency Washing" in the Digital Age: A Corporate Agenda of Procedural Fetishism*, in «Critical Analysis of Law: An International & Interdisciplinary Law Review», VIII, n. 1, 2021, pp. 139-153.

[22] Crf. C. L. Thøger e J. Cornelissen, *Organizational Transparency as Myth and Metaphor*, in «European Journal of Social Theory», XVIII, n. 2, 2015, pp. 132–149.

[23] Questo capovolgimento è segnalato in E. Alloa, Y. Citton*, Tyrannies de la transparence*, in «Multitudes», IV, n. 73, 2018, p. 51.

Non dobbiamo dimenticare che la trasparenza non è una qualità delle pratiche umane di cui sia possibile isolare un senso valido una volta per tutte: essa è piuttosto l'idealizzazione pragmatica di un insieme di dinamiche individuali e collettive che si trasformano costantemente a partire da specifiche attese sociali incorporate nelle narrazioni che le avvolgono. Sono queste narrazioni, allora, che devono essere vagliate criticamente, tanto più in un contesto in cui, con l'avvento del digitale, si è incrementata l'accessibilità alle informazioni e alla visibilità, rendendo così ancora più indiscutibile l'imperativo della trasparenza e dei discorsi che la legittimano. Ecco perché è di grande interesse il recente sviluppo dei cosiddetti *Critical Transparency Studies*[24], a cui va il merito non solo di aver aperto un ampio dibattito sulle ambiguità che caratterizzano le retoriche dominanti, ma anche di aver cominciato a delineare i limiti e a chiarire gli ambiti all'interno dei quali la storia di lunga durata di questo concetto può oggi ancora avere un senso.

In linea con questo approccio critico, nel seguito del nostro discorso isoleremo tre questioni a partire da cui si può comprendere lo scollamento tra le narrazioni che legittimano l'imperativo della trasparenza e le pratiche reali in cui tale imperativo si concretizza. Ci sembra questa la migliore strategia per affrontare l'ideologia della neutralità, perché ci permette di evidenziare quanto con Foucault possiamo chiamare il "non detto" di queste narrazioni, cioè i regimi di dicibilità e di visibilità che indirizzano oggi quel consenso incontrastato a livello pubblico.

Si tratta certo di un primo passo e forse ci si potrebbe aspettare di più. Tuttavia, se intendiamo l'ideologia come l'insieme di narrazioni che contribuiscono a offrire una rappresentazione condivisa di quanto consideriamo come reale e a delineare una cornice assiologica su cui si fondano le attese collettive[25], allora nel momento stesso in cui ne evidenziamo le ambiguità e le contraddizioni creiamo i presupposti perché possano nascere resistenze e contro-narrazioni in grado di rendere operative alternative realistiche.

4. La natura mediale della trasparenza e le illusioni della disintermediazione

Innanzitutto merita una più precisa tematizzazione l'identificazione di trasparenza e immediatezza, per cui la prima si otterrebbe solo attraverso un'attiva negazione di

[24] Dei lavori che stanno emergendo in questo ambito si veda: C. Birchall, *Radical Transparency?*, in «Cultural Studies? Critical Methodologies», XIV, n. 1 2014, pp. 77-88; E. Alloa, D. Thomä (a cura di), *Transparency, Society and Subjectivity. Critical Perspectives*; cit.; C. Birchall, *Radical Secrecy: The Ends of Transparency in Datafied America* Universiy of Minnesota Press, Minneapolis, 2021; J. I. Valdovinos, *Transparence and Critical Theory: The Becoming Transparent of Ideology*, New York, Springer International Publishing, 2022; E. Alloa (a cura di), *This Obscure Thing called Transparency*, cit., e I. Koivisto, *The Transparency Paradox*, Oxford University Press, Oxford, 2022.

[25] Ci riferiamo qui a una concezione non marxiana dell'ideologia, quale è per esempio espressa da Régis Debray: «Ogni cultura si definisce per ciò che essa concorda di considerare reale. Da un secolo scarso chiamiamo "ideologia" questo *consensus* che cementa ogni gruppo organizzato» (R. Debray, *Vita e morte dell'immagine. Una storia dello sguardo in Occidente*, trad. it. di A. Pinotti, Il Castoro, Milano 1999, p. 295).

ogni mediazione, così da consentire un accesso diretto e senza filtri alla realtà. Secondo questa convinzione, sarebbe trasparente ciò che è totalmente aperto, visibile e privo di opacità, mentre ogni mediazione costituirebbe un impedimento, una barriera al flusso di informazioni e alla spontaneità delle relazioni[26].

Va da sé che questa identificazione è oggi favorita dalla comunicazione online e in particolare dalla proliferazione degli schermi digitali, che producono negli utenti la sensazione di avere un rapporto diretto con ciò che percepiscono sui display[27]. Già alla fine del secolo scorso Jay David Bolter e Richard Grusin, nel loro libro *Remediation*[28], avevano evidenziato che la definitiva affermazione delle interfacce grafiche nei computer rispondeva a una "logica dell'immediatezza trasparente"[29] che mirava a far scomparire il più possibile la presenza mediatrice della macchina. È poi divenuto sempre più evidente che il modo stesso in cui è progettata la comunicazione digitale si allinea all'ideologia dell'immediatezza, perché mira a cancellare ogni soluzione di continuità tra gli schermi e il nostro mondo abituale.

Questa logica contrabbanda quindi l'apparente naturalezza delle esperienze schermiche con l'assenza di ogni mediazione, nascondendo attivamente la prestazione mediatrice dei dispositivi informatici, oltretutto organizzati secondo logiche di trasmissione e di elaborazione dei dati tutt'altro che neutrali. Quando nel linguaggio comune diciamo che questi media sono "programmati", stiamo indicando, spesso senza rendercene conto, che essi rispondono a precise scelte condizionate dal modo in cui sono stati progettati i software, dagli algoritmi utilizzati e dal funzionamento materiale dell'hardware. Possiamo illuderci che vedere in diretta un evento sportivo corrisponda con il parteciparvi senza mediazioni, come se fossimo lì di persona, o che rispondere a un sondaggio su una piattaforma politica ci dia la possibilità di contare davvero nelle dinamiche di governo, ma in realtà quello che vediamo o le possibilità di intervento che ci sono date sono vincolate all'apparato tecnologico che incapsula la nostra esperienza in specifici regimi di visibilità e di dicibilità. Come non ricordare al riguardo le pagine dedicate già negli anni '90 dello scorso secolo da Don Ihde al ruolo della "mediazione tecnologica" e alla capacità degli oggetti tecnici di generare un *background* che crea le condizioni di possibilità del nostro rapporto con il mondo[30], o ancora, più di recente, alla distinzione proposta Bruno Latour, secondo cui le tecnologie non sono

[26] Cfr. B.-C. Han, *Nello sciame. Visioni del digitale*, trad. it. di F. Buongiorno, Nottetempo, Roma 2015, p. 29.

[27] Si pensi al riguardo all'importanza che viene attribuita nella comunicazione digitale all'effetto immersivo e alla sensazione di "presenza", cioè alla convinzione di essere "realmente lì", in contatto diretto con un ambiente che è però artificiale e tecnologicamente mediato. Cfr. al riguardo G. Riva et al., *Being There. Concepts, Effects and Measurements of User Presence in Synthetic Environments*, IOS Press, Amsterdam 2003.

[28] J.D Bolter, R. Grusin, *Remediation. Competizione e integrazione tra media vecchi e nuovi*, trad. it. di B. Gennaro, Guerini, Milano 2018.

[29] Ivi, p. 44.

[30] D. Ihde, *Technology and the Lifeworld: From Garden to Earth*, Indiana University Press, Bloomington 1990, pp. 109-112.

dei semplici "intermediari"[31], ma dei veri e propri "mediatori" che contribuiscono a dare forma alla realtà che ci circonda.

A livello sociale e politico, più che l'oblio della mediazione diventa però rilevante la ricerca attiva della disintermediazione. In questo caso il tratto ideologico della trasparenza si manifesta nella convinzione che ogni forma di mediazione vada categoricamente negata perché rappresenta una indebita intromissione nel libero flusso dei dati e un impedimento alla diffusione delle idee. Emblematica è al riguardo la generale crisi degli esperti e il crollo di fiducia nei confronti delle forme istituzionali della comunicazione e del sapere[32]. Poiché il Web permette di accedere a informazioni che in precedenza erano appannaggio di una cerchia ristretta di specialisti, si diffonde la convinzione che tutti possano essere competenti su tutto. Questa accessibilità generalizzata rende possibile a chiunque di produrre notizie spesso non verificate, aumentanto il tasso di *fake news* che circolano in rete o ancora consente di presentarsi come esperti solo perché ci si è occupati di un determinato problema, senza avere però una formazione specialistica. Ne fanno le spese innanzitutto i giornalisti, il cui ruolo diventa anacronistico[33] in un contesto in cui ciascuno può cercarsi indipendentemente le proprie fonti di informazioni in rete. La stessa sorte tocca agli scienziati, di cui si fa sempre più fatica a riconoscere l'autorevolezza, con una complessiva perdita di fiducia nei loro confronti. Non va meglio ai politici di professione, che vengono descritti come una casta di corrotti che non hanno più alcun rapporto con la gente comune che dovrebbero rappresentare.

Quel disprezzo nei confronti della competenza è oggi particolarmente evidente in alcuni movimenti populisti. In essi l'esaltazione della disintermediazione fa il paio con la presa di distanza dalla democrazia rappresentativa[34], con la demonizzazione della stampa "asservita ai poteri forti", con l'adozione di tesi complottiste o apertamente antiscientifiche. Questa retorica non infarcisce soltanto il tecnopopulismo, alimentando la convinzione che la democrazia digitale debba fare a meno delle tecniche tradizionali della mediazione politica per permettere a tutti senza distinzione un attivismo "in presa diretta" sulle decisioni. Essa avvolge anche il modo con cui agiscono e si impongono elettoralmente alcuni leader – si pensi, per fare un esempio emblematico, a Donald Trump e alla recente campagna per le presidenziali americane – la cui unica funzione è quella di rispecchiare le attese immediate dei loro elettori utilizzando un linguaggio semplice e diretto che rifiuta l'articolazione di ragionamenti complessi e fa leva invece su emozioni e paure viscerali, come l'invasione degli immigrati o la distruzione della famiglia tradizionale.

In realtà questa pretesa di totale disintermediazione cela il proprio contrario: l'accesso diretto al governo della cosa pubblica decantato dal tecnopopulismo e la

[31] Cfr. B. Latour, *Riassemblare il sociale*, trad. it. di D. Caristina, Meltemi, Milano 2022.

[32] Sulla crisi degli esperti oltre all'ormai classico T. Nichols, *La conoscenza e i suoi nemici. L'era dell'incompetenza e i rischi per la democrazia*, trad. it. di C. Veltri, Luiss University Press, Roma 2020; si veda anche G. Eyal, *The Crisis of Expertise*, Polity Press, Cambridge 2019.

[33] Si veda al riguardo B.-C. Han, *Nello sciame*, cit., p. 30.

[34] Cfr. N. Urbinati, *Democrazia in diretta. Le nuove sfide alla rappresentanza*, Feltrinelli, Milano 2013.

spontaneità con cui si presentano i principali leader populisti nascondono complessi montaggi mediatici ben lontani dalla presunta immediatezza di cui è infarcita la loro retorica. La trasparenza collettiva, di cui si vorrebbero paladini i movimenti populisti, poggia infatti su meccanismi opachi di manipolazione e dominazione politica dove invece di una disintermediazione entrano in gioco semplicemente altre forme di mediazione, più subdole perché meno evidenti e quindi meno facilmente criticabili. L'assenza di mediazione si capovolge così in una iper-mediazione in cui la pretesa neutralità ideologica si alimenta invece di forme di dominazione ancora più pericolose, perché immunizzate da ogni messa in discussione.

5. Le trappole della trasparenza individuale

Questo capovolgimento si ritrova anche nella seconda questione che ci interessa, ovvero la convinzione che la trasparenza favorisca di per sé il miglioramento della vita morale del singolo e delle collettività e che essa, grazie alle sempre maggiori possibilità di condivisione orizzontale, permetta di portare a compimento l'ideale moderno di autonomia individuale e di emancipazione. Per comprendere in che senso anche questa narrazione progressiva della trasparenza nasconda una serie di insidie, non solo concettuali, può essere significativo ricordare la vicenda di un celebre libro di Gianni Vattimo, guarda caso intitolato *La società trasparente*[35]. Dato alle stampe nel 1989, il volume viene ripubblicato in nuova edizione nel 2000 con l'aggiunta di un capitolo finale che segna un significativo cambio di prospettiva. Nel testo della fine degli anni '80 il filosofo torinese sosteneva con una certa enfasi che i media di massa, l'accesso sempre più allargato alla comunicazione e la crescente trasparenza avrebbero prodotto la "liberazione delle differenze"[36], aprendo a un mondo in cui la pluralizzazione dei valori ci avrebbe finalmente emancipato da verità uniche e poteri soffocanti, consentendo una crescita di consapevolezza e autonomia individuale. Nel capitolo aggiunto per la seconda edizione[37] egli ammette invece che il suo "ottimismo mediatico" era mal riposto. Anziché suscitare un riconoscimento delle differenze e liberare inedite dinamiche di soggettivazione, quella trasparenza, dispiegata in poco più di dieci anni, aveva generato nuove forme di dominio economico e inspiegabili irrigidimenti dogmatici.

Ebbene, la parabola di questo libro di Vattimo ci può aiutare a rileggere le affermazioni di Harcourt e di Han che abbiamo citato in precedenza rispetto al bisogno auto-indotto di esporsi senza riserve nell'epoca dei social network. Anche nel caso dell'esposizione online la trasparenza sembra legittimarsi grazie al bisogno di liberare la propria autenticità e di condividere in modo schietto momenti importanti della propria vita. Certo, se la privacy è un furto, come si legge ne *Il Cerchio*, allora ha

[35] G. Vattimo, *La società trasparente*, Garzanti, Milano 2000, prima ed. 1989.
[36] Ivi, p. 17.
[37] Ivi, pp. 101-121.

ragione Jeff Jarvis a sostenere in *Public Parts*[38] che online sono vincenti coloro che hanno una *publicness* totale, perché solo in questo modo si può sfruttare a fondo il nuovo paradigma relazionale della società digitale. «Ogni volta che non condividi», afferma Jarvis non senza ridondanza retorica, «una relazione perde le sue ali. Questa è una perdita tangibile»[39].

Tuttavia questa compulsione a condividere è davvero espressione di una nuova libertà conquistata grazie al digitale o nasconde tratti profondamente ambigui, che finiscono per negarla? Non è un caso che, come ha segnalato Olivier Aïm[40], negli ultimi due decenni è stato ampiamente citato il *Trattato sulla servitù volontaria* di Etienne de la Boétie, un testo del XVI secolo, solo in apparenza lontano dalla situazione attuale. De la Boétie descrive come nell'uomo vi sia una tendenza, non per forza consapevole, a servire il potere[41] che impedisce e blocca i processi di resistenza e di rivolta. Ebbene, nell'esposizione online sembra emergere la stessa dinamica: consegniamo volontariamente parti della nostra vita illudendoci di esercitare la nostra libertà di esprimerci, ma in realtà stiamo servendo dinamiche di potere e di assoggettamento.

Ma c'è un elemento ulteriore: Aïm nota che il testo di La Boétie avanza l'idea che la natura stessa del potere si fondi «sulla sua capacità di far aderire i dominati alla loro condizione»[42]. Ed è proprio su questo versante che il capovolgimento della libertà tocca il suo apice, perché il modo stesso in cui agiscono le grandi aziende hi-tech nasconde la coercizione, agendo in modo anonimo e apparentemente "inoffensivo". Una delle caratteristiche della cosiddetta "governamentalità algoritmica" è secondo Antoinette Rouvroy e Thomas Berns proprio il fatto che lo sfruttamento dei dati, conseguente alla esposizione online, avviene in un modo neutro e "gentile". Esso non è per nulla invasivo, ma approfitta del fatto che più o meno inconsapevolmente accettiamo liberamente di consegnare informazioni al web, sia per ignoranza, sia, forse, per semplice pigrizia intellettuale. I dati non ci vengono "rubati", ma semplicemente li abbandoniamo sui social network, sulle piattaforme che frequentiamo e nella mail che spediamo; essi sono tracce che lasciamo e non informazioni che vogliamo trasmettere intenzionalmente, ma proprio questo allenta la nostra consapevolezza rispetto agli effetti di quanto stiamo facendo[43].

A questa attitudine non manca di dare un contributo la retorica della trasparenza personale, perché attribuisce a queste pratiche un significato morale ed emancipatorio. Se si fa però un passo a lato rispetto alle narrazioni sull'esposizione totale ci si rende conto che in gioco qui non vi è più un progresso etico a livello individuale e collettivo, e tantomeno una dinamica emancipatrice, bensì qualcos'altro. Nell'attività

[38] Cfr. J. Jarvis, *Public Parts: How Sharing in the Digital Age Improves the Way We Work and Live*, Simon and Schuster, New York 2011.
[39] Ivi, p. 46, traduzione nostra.
[40] Cfr. O. Aïm, *Les théories de la surveillance*, cit., p. 137.
[41] Cfr. M. Carbone, G. Lingua, *Antropologia degli schermi*, cit., p. 106.
[42] A. Rouvroy, T. Berns, *Le nouveau pouvoir statistique. Ou quand le contrôle s'exerce sur un réel normé, docile et sans événement car constitué de corps 'numériques'*, in «Multitudes», I, n. 40, 2010, traduzione nostra.
[43] Cfr. A. Rouvroy, T. Berns, *Gouvernementalité algorithmique et perspectives d'émancipation. Le disparate comme condition d'individuation par la relation?,* in «Réseaux», n. 177, 2013.

di raccolta e profilazione prodotta dal *data mining* ciò che interessa ai padroni del web non è il processo di soggettivazione, di cui la libertà e l'emancipazione sarebbero il risultato, ma l'aggregazione di dati per prevedere in anticipo dei comportamenti e poterli poi sfruttare economicamente. La personalizzazione delle offerte di informazioni e di prodotti, che viene declamata da istituzioni pubbliche e da aziende private come un servizio agli utenti, ha in realtà una natura ambivalente perché è anche usata per catturare la nostra attenzione sul mercato delle preferenze e per condizionare il nostro comportamento futuro.

Resta vero però che tematizzare l'ambiguità della trasparenza personale negli ambienti digitali non significa disconoscere che l'esposizione online per alcuni individui o gruppi può essere, come abbiamo visto in precedenza, una forma di espressione e appropriazione discorsiva delle proprie identità. La presenza sui social, per esempio, può costituire per molte persone una occasione di socializzazione diversamente impossibile e la rete rappresentare l'ambiente adatto per l'esercizio di una riflessività matura grazie alla multimodalità comunicava che consente e alle potenzialità creative che dischiude[44].

Ciononostante si produce uno scollamento tra le forme esplicite di comunicazione online, che, pur nella loro ambivalenza, possono contribuire positivamente a una dimensione costruttiva di trasparenza personale, e un apparato implicito di processi di datificazione che resta opaco e fuori da ogni possibile controllo degli utenti. Tale scollamento è certamente legato alla conformazione ontologica stessa degli oggetti digitali che, come ha ben evidenziato Frieder Nake, hanno una natura duale, costituita da una *surface* accessibile all'utente grazie alle interfacce (per lo più visuali) con cui egli si relaziona alla macchina, e una *subface,* il cui funzionamento è inacessibile ai più, costituita dai diversi processi algoritmici interni al software, nonché dalla struttura materiale dell'hardware. Tenendo a mente questa conformazione ontologica diventa particolarmente significativo quanto Shoshana Zuboff in *Il capitalismo della sorveglianza* definisce come il "problema dei due testi"[45]. Zuboff nota infatti come nella comunicazione digitale ci sia una disgiunzione radicale tra un "primo testo" visibile agli utenti in quanto autori e fruitori, e un secondo "testo ombra" che invece è loro del tutto inaccessibile, tanto che spesso non si è neanche consapevoli della sua esistenza. Da una parte ci sono infatti i post che scriviamo, i video che carichiamo, le storie che pubblichiamo e i like che disseminiamo sui diversi social, dall'altra invece i (meta)dati che vengono estratti per costituire un archivio totalmente al fuori del nostro controllo e che rappresentano una "lettura riservata" unicamente a chi gestisce le piattaforme. Ciò che fa problema però non è soltanto la mancanza di consapevolezza che gli utenti hanno del testo ombra, ma anche il fatto che ad esso non si sfugge nel momento stesso in cui si è connessi con un qualsiasi dispositivo. Insomma senza volerlo alimentiamo costantemente quel secondo testo e «la nostra esperienza viene

[44] Su questo si vedano per esempio le osservazioni di P. Montani, *Emozioni dell'intelligenza. Un percorso nel sensorio digitale*, Meltemi, Milano 2020, pp. 7-12.

[45] Cfr. S. Zuboff, *Il capitalismo della sorveglianza*, trad. it. di P. Bassotti, Luiss University Press, Roma 2019, pp. 197 ss.

costretta a diventare una materia prima da accumulare e analizzare per i fini commerciali di altre persone»[46].

Nel libro di Zuboff, il problema dei due testi si collega alle tesi economiche dell'autrice rispetto al "surplus comportamentale" che si ottiene dai dati degli utenti che vengono aggregati e sfruttati per inferire sui loro comportamenti, sulle loro intenzioni e finanche sulle loro convinzioni politiche, per poterli condizionare. Ma più ancora alimenta il dispositivo complessivo di un nuovo tipo di capitalismo che si nutre di un disegno di sorveglianza globale che l'autrice descrive con la metafora del Grande Altro (*Big Other*). Diversamente dal Grande Fratello di Orwell, il Grande Altro si fonda sul potere di pochi che agiscono come "burattinai" silenziosi attraverso la dinamica anonima della rete e si arricchiscono anche grazie alla nostra disponibilità ad una trasparenza totale. Ebbene, tale potere non ha certamente come obiettivo l'emancipazione degli utenti.

6. Trasparenza collettiva e sorveglianza

Le tesi di Zuboff, che pure hanno attirato molte critiche[47], non esemplificano soltanto le trappole insite nel nesso tra l'esposizione online e l'apparato di sfruttamento economico della datificazione che è divenuto il web, ma ci introducono direttamente all'ultimo tratto ideologico che ci interessa discutere, questa volta relativo alla trasparenza collettiva. Ci riferiamo al nesso che nella società digitale lega in modo sempre più diretto l'imperativo della trasparenza e le nuove forme di sorveglianza generalizzata.

La ricerca di trasparenza, che un tempo rappresentava una prerogativa dei cittadini nei confronti di chi governava e di chi deteneva il potere economico, diventa oggi, come abbiamo intuito nelle sezioni precedenti, una leva di sorveglianza e assoggettamento. A cambiare di natura è la portata politica e sociale della trasparenza collettiva: nata come spinta "dal basso" di partecipazione democratica e di condivisione delle decisioni del governo, essa si trasforma in una pressione "dall'alto" di controllo sociale che continua però ad alimentarsi dell'immaginario originario. L'ideologia si annida proprio qui, nel disallineamento tra la retorica che sostiene la ricerca di trasparenza come condivisione e partecipazione, e la pratica che si attua invece all'interno di un apparato di sorveglianza che rischia di essere sempre più ubiquo. Ad aggravarne il peso si aggiunge poi il fatto che non ci troviamo più di fronte a un Grande Fratello ben identificabile, bensì a un potere anonimo; non facciamo più i conti con la repressione della "polizia del pensiero" o con il sinistro *telescreen* orwelliano, ma con un apparato diffuso che sfrutta milioni di schermi e di dispositivi portatili tutt'altro che minacciosi[48].

---

[46] Ivi, p. 197.

[47] Cfr. al riguardo quanto dice O. Aïm, *Les théories de la surveillance*, cit., pp. 130-133.

[48] Per una analisi precisa di questa dimensione ubiqua e reticolare della trasparenza si veda D. Lyon, *La cultura della sorveglianza*, cit., p. 19 e ss.

Il carattere reticolare e benevolo della sorveglianza ci dice molto della trasformazione in atto. Apparentemente un'immagine eloquente di questa nuova condizione sembrerebbe essere quella del "panottico digitale" proposta da Byung-Chul Han in *La società della trasparenza*[49]. Secondo Han, diversamente da quanto avviene nel modello del panottico di Bentham e nella ripresa fattane da Michel Foucault per illustrare il potere disciplinare[50], la società della trasparenza ci consegnerebbe a una prigione in cui «i suoi stessi abitanti collaborano attivamente alla sua costruzione e al suo mantenimento, esponendosi loro stessi alla vista e denudandosi»[51]. In realtà il legame tra trasparenza totale e sorveglianza, nonché il modo con cui avviene quel coinvolgimento diretto dei sorvegliati, mostrano a nostro parere che il modello panottico va ormai messo da parte[52]. In esso come nella metafora del Grande Fratello si conserva ancora un centro di attrazione delle pratiche di osservazione, mentre oggi è proprio quel centro a essere venuto meno.

Peraltro va notato che già più di vent'anni fa Kevin Haggerty e Richard V. Ericson avevano sottolineato che per comprendere le dinamiche della sorveglianza era necessario pensarle come un "assemblaggio"[53], termine preso a prestito da Deleuze e Guattari per esprimere la pluralità di elementi in gioco che non permettono più di ricondurre i meccanismi di controllo a un potere chiaramente identificabile. Questa evoluzione era allora dovuta innanzitutto allo sviluppo degli ambienti digitali, che permettevano di intensificare le tecniche di registrazione, ed è oggi ulteriormente implementata grazie alla computazione ubiquitaria che integra i computer nelle pratiche umane e all'internet delle cose che mette in connessione gli oggetti tra loro e con gli utenti. Con la svolta digitale la sorveglianza diventa quindi molto più fluida, rizomatica e onnipresente[54].

A partire da queste premesse, invece dei modelli panottici, ci sembra più congruente con le pratiche di trasparenza collettiva quanto Davide Lyon qualifica come una "cultura della sorveglianza"[55], cioè una trasformazione complessiva degli immaginari e dei modi concreti con cui i singoli e le collettività integrano le pratiche di sorveglianza nella loro vita. Invece di essere solo appannaggio dello stato e dei poteri economici, la sorveglianza diventa un habitus quotidiano a cui è difficile non prendere parte. Dagli apparati di videosorveglianza sempre più diffusi ai dati che consegniamo alle piattaforme, dalle tessere di fidelizzazione dei supermercati alle informazioni

[49] Cfr. B.-C. Han, *La società della trasparenza*, cit.
[50] Cfr. M. Foucault, *Sorvegliare e punire*, trad. it. di A. Tarchetti, Einaudi, Torino 1993.
[51] B.-C. Han, *La società della trasparenza*, cit. p. 78.
[52] David Lyon nel libro intervista a Zygmunt Bauman dal titolo *Sesto potere. La sorveglianza nella modernità liquida* (trad. it di M. Cupellaro, Laterza, Roma-Bari 2022) sostiene con forza questa presa di distanza dal modello panottico: «L'utilità sia storica che logica dell'immaginario panottico oggi sembra essersi esaurita» (ivi, p. 39).
[53] Il loro saggio che avrà una funzione seminale per i *surveillance studies* si intitola appunto *The Surveillance Assemblage*, in «The British Journal of Sociology», 51/4, 2000.
[54] Su questi aspetti si veda O. Aïm, *La surveillance comme performance d'écran*, in «Études digitales», n. 12, 2023.
[55] Cfr. D. Lyon, *La cultura della sorveglianza*, cit.

biometriche che disseminiamo nelle pratiche quotidiane, siamo progressivamente entrati in una forma di sorveglianza «che minimizza la 'disciplina' e il 'controllo' mettendo in primo piano la "performace"»[56]. Non è più quindi qualcosa che interviene "dall'esterno", ma si inserisce nella vita sociale "dall'interno" perché si consuma la distinzione tra chi è sorvegliato e chi sorveglia, in quanto tutti «partecipano attivamente alla propria sorveglianza e a quella degli altri»[57].

L'ubiquità e il ruolo sempre più attivo delle persone implicate portano però in primo piano le ambiguità che abitano l'intreccio tra trasparenza e sorveglianza. Da una parte la cessione costante di dati si legittima perché conviene per mantenere alcuni livelli di comfort della nostra vita in quanto ci permette di fruire di servizi personalizzati e di risparmiare tempo. Dall'altra questa comodità si paga con la costante profilazione, che non è soltanto legata allo sfruttamento economico descritto da Zuboff, ma anche a un meccanismo di classificazione sociale, silenzioso e tuttavia pervasivo. La stessa strategia predittiva che sta alla base dei consigli del marketing personalizzato viene utilizzata per costruire indici, schedature e quindi categorie in cui incasellare singoli individui o gruppi che mostrano una certa omogeneità adatta alle strategie del mercato delle preferenze. Se questo uso "performativo" della statistica risale perlomeno al XIX secolo[58], l'avvento dell'informatica e in ultimo l'introduzione massiccia dell'Intelligenza Artificiale consentono oggi di creare enormi basi di dati sfruttabili per classificare l'affidabilità finanziaria nella concessione di prestiti, gli stili di vita nella stipula delle assicurazioni e la propensione alla recidiva in ambito giudiziario, solo per citare gli esempi più significativi.

Quali possono essere le frontiere dell'impatto discriminante della classificazione sociale (*social sorting*) lo si può vedere nel Sistema di Credito Sociale lanciato in Cina nel 2014 e costantemente implementato in seguito. L'obiettivo in questo caso va ben al di là della classificazione dell'affidabilità in specifici settori, ma coinvolge complessivamente la valutazione dell'integrità sociale e civica dei cittadini[59]. Esso si fonda su un sistema di restrizioni e incentivi che dipendono dal punteggio che si ottiene nella valutazione del proprio comportamento sociale (dall'affidabilità economica al rispetto delle norme basilari della convivenza, come le regole della circolazione o l'abbandono della spazzatura in luoghi non autorizzati) con l'obiettivo di distinguere tra "buoni" e "cattivi" cittadini. Quanto sia problematico l'impatto di questa iniziativa del governo cinese lo si vede dalle restrizioni cui sono sottoposti coloro che hanno un credito sociale basso: limitazione negli spostamenti, esclusione dei figli da scuole di alto livello, impossibilità di prenotare determinati hotel e così via.

---

[56] Ivi, p. 28.

[57] Ivi, p. 24.

[58] Sull'uso della statistica e sul desiderio di classificare le persone si veda B. Harcourt, *Against Prediction. Profiling, Policing and Punishing in an Actuarial Age*, University of Chicago Press, Chicago 2007.

[59] Sul funzionamento del Sistema di Credito Sociale si veda V. Brussee, *Social Credit: The Warring States of China's Emerging Data Empire*, Palgrave MacMillan, London 2023; G. Centraco, G.M.D. Dore, *Il Sistema di Credito Sociale cinese: tecnologia come strumento di sorveglianza e persuasione*, «OrizzonteCina», XV, n. 1, 2024, pp. 61-78.

Questo scenario cinese desta evidenti preoccupazioni perché concretizza un immaginario distopico che conosciamo anche in occidente. Basti pensare all'episodio della serie *Black Mirror* "Caduta libera", in cui la protagonista Lacie Pound vive in un mondo in cui tutto è condizionato dai rating dei canali social - non solo ogni relazione, ma anche servizi essenziali come l'assistenza medica[60]. La finzione cinematografica ci fa intravedere gli esiti sociali estremi a cui può condurre l'esposizione online e l'auto-imposizione di una trasparenza personale che contribuisce ad alimentare pratiche discriminatorie di datificazione e classificazione. Senza arrivare alle figure distopiche messe in scena da *Black Mirror*, un autore come Oscar Gandy già nel 1993 in *Panoptic Sort*[61] aveva messo in evidenza i nessi tra sorveglianza e classificazione, e ne aveva studiato gli intrecci con i processi di razionalizzazione sociale propri della seconda modernità. Su questa matrice i social media e le tecnologie dei big data hanno notevolmente incrementato le capacità di schedatura[62] e quindi anche gli effetti di discriminazione che ne derivano dall'incrocio tra l'ossessione per l'esposizione del privato e l'enorme capacità di cattura che gli ambienti digitali hanno raggiunto.

7. Conclusione

Osservata attraverso il prisma della sorveglianza, l'ossessione per la trasparenza sembra aprire scenari distopici. Mettere in evidenza le derive cui va incontro il capitalismo della sorveglianza delle grandi aziende hi-tech o le incognite legate all'espansione del *social sorting* non significa però disconoscere gli aspetti positivi della cultura della partecipazione e della condivisione che alimentano le richieste di trasparenza e ne legittimano l'importanza. Quello che ci interessava evidenziare era piuttosto la natura ancipite e quindi ambigua che sta alla base delle narrazioni dominanti, nonché il fatto che questa natura fatichi a emergere a causa dell'aura di indiscutibilità che la circonda. Nel momento stesso in cui la trasparenza diventa un imperativo sociale dato per scontato restano nell'ombra i rischi a essa connessi e non si possono apprezzare le reali potenzialità operative che dimostrano alcune pratiche individuali e collettive capaci di collocarsi al di fuori dell'ideologia della trasparenza.

Già Edouard Glissant, nei suoi *Discours antillais*, sosteneva l'idea di un "diritto all'opacità", in contrapposizione alla vecchia ossessione dell'Occidente di scoprire "ciò che sta al fondo delle nature" e al suo incessante "desiderio di trasparenza"[63]. Per il grande teorico della creolizzazione, il sogno di una trasparenza totale faceva ancora parte di una fantasia coloniale che ambiva a illuminare il non-europeo, dove "illuminazione" implica sia portare la luce negli angoli presumibilmente bui del

[60] Su questo si veda O. Aïm, *Les théories de la surveillance*, cit., p. 120.
[61] Cfr. O.H. Gandy, *The panoptic sort: A political economy of personal information*, Oxford University Press, 2021 (prima ed. 1993).
[62] Nella postfazione della seconda edizione del suo libro Gandy attualizza in modo incisivo l'interpretazione che aveva proposto negli anni '90 dello scorso secolo, misurando direttamente le sue tesi sul *panoptic sort* con la svolta digitale. Cfr. Ivi, pp. 264-284.
[63] Cfr. E. Glissant, *Discours antillais*, Gallimard, Parigi, 1981.

mondo, sia farne sparire le ambiguità. Contro questo ideale, che segna l'unica via presumibile verso la conoscenza e l'emancipazione, Glissant difese dunque un diritto all'opacità che è stato a volte frainteso. Anni dopo, nella sua *Poetica della relazione,* Glissant dovette infatti precisare che quello che intendeva era un «diritto all'opacità che non sia il confinamento in un'autarchia impenetrabile», ma che permetta interpenetrazioni laterali, riconnessioni e relazioni non soggette a una norma unificata di subordinazione[64].

Se, in un'epoca sotto il segno dei *big data*, la critica della cattura biometrica e il rifiuto dell'identificazione possono essere obiettivi politici ragionevoli da perseguire, voler "diventare impercettibili" non si situa necessariamente nel campo progressista: basta vedere come anche la nebulosa complottista QAnon rivendica ormai strategie di invisibilità e di opacità. Militare in favore dell'inscrutabilità è appunto una delle rivendicazioni principali del libertarismo individualista, che difende la privacy come diritto a essere lasciati in pace e a non dover incontrare forme di vita dissimili. Ecco perché, invece di mobilitarci per un "diritto generale a scomparire", che si avvicina a certe essenzializzazioni problematiche della *privacy*, serve oggi difendere un "diritto settoriale a non essere identificati"[65]. Rivendicare questo diritto significa semplicemente pretendere che alcuni ambiti della nostra vita non siano sotto la minaccia di essere immediatamente giudicati e valutati, con il rischio di compromettere possibilità future. Non va dimenticato che il Sistema di Credito Sociale cinese non è poi così distante da quello che sta già succedendo nei paesi occidentali, dove banche, compagnie assicurative o datori di lavoro commissionano agenzie dedicate per somministrare test automatici o studiare i post sui social dei potenziali candidati, al fine di stabilire il loro profilo comportamentale[66].

Alla luce degli inquietanti usi futuri dei dati personali, e della capacità sempre maggiore di interconnessione automatizzata delle tracce digitali, più che mai sarà importante creare e difendere ambiti dove gli individui possano sperimentare delle identificazioni e delle espressioni alternative, senza essere subito esposti al tribunale algoritmico. È questa decorrelazione di azioni e identità, ma anche l'esplorazione provvisoria di altri modi di concepire la propria esistenza, che potrebbe aiutare oggi a evitare la minaccia di una normalizzazione preventiva dei comportamenti.

Al di là di una sterile opposizione tra messa in scena e autenticità, al di là dell'opposizione tra trasparenza e difesa di una qualsivoglia *privacy* inviolabile, si tratta di ripensare le democrazie come spazi di sperimentazioni trasformative in cui le narrazioni dominanti lascino spazio a condotte differenti. Contro il "trasparentismo",

---

64 E. Glissant, *Poetica della relazione,* trad. E. Restori, Quodlibet, Macerata 2007 (cf. in particolare i capitoli *Trasparenza e opacità* e *Per l'opacità*).

65 Rimandiamo alla presentazione dettagliata di questi argomenti intorno alla *privacy* e all'idea di una 'disattenzione civile', ispirata a Raymond Geuss, nell'articolo E. Alloa, *Transparency, Privacy commons and Civil inattention*, in S. Berger et al., *Cultures of Transparency. Between Promise and Peril*, Routledge, Londra-New York 2021, pp. 171-192.

66 C. O'Neill, *Weapons of Math Destruction. How Big Data Increases Inequality and Threatens Democracy,* Random House, Londra, 2016, pp. 105-178.

che produce nuove forme di autocensura per anticipazione, vanno difesi spazi di sperimentazione creativa, dove gli individui possano giocare con altri ruoli e identità, senza essere subito tracciati e giudicati dai sistemi di *big data*, e senza ipotecare subito il loro futuro. Più che mai, diventa fondamentale poter evadere le attribuzioni e immaginare altri modi d'essere e di concepirsi. La sfida democratica consisterà quindi nel difendere spazi di decorrelazione, dove poter collaudare maniere alternative di intendere l'esistenza, sia sul piano individuale che collettivo.

Digital Discrimination. The Challenge of Bias and Transparency in AI[a]

Gabriele Giacomini[*], *Chiara Aprilis*[†]

Abstract

Il saggio esplora l'impatto crescente dell'intelligenza artificiale sui sistemi decisionali e le sue implicazioni etiche, concentrandosi sui pregiudizi algoritmici che possono portare a discriminazioni basate su genere, etnia e altri fattori. Attraverso esempi concreti, si discute di come i pregiudizi possano manifestarsi e si sottolinea l'importanza di un approccio responsabile alla governance dell'IA. Ciò implica la promozione di una riflessione sia accademica che pubblica sull'adozione di principi etici e procedure il più possibile trasparenti e inclusive.

Parole chiave: Algoritmi, Intelligenza Artificiale, Bias, Discriminazione, AI Governance

This paper explores the growing impact of artificial intelligence on decision-making systems and its ethical implications, focusing on algorithmic biases that can lead to discrimination based on gender, ethnicity, and other factors. Through concrete examples, it discusses how biases may manifest and emphasises the importance of a responsible approach to AI governance. This involves promoting both academic and public reflection on the adoption of ethical principles and procedures that are as transparent and inclusive as possible.

Keywords: Algorithms, Artificial Intelligence, Bias, Discrimination, AI Governance

[a] The conception of the article is by both authors. However, Aprilis focused on section 2, Giacomini on sections 1 and 3. Saggio ricevuto in data 11/03/2024 e pubblicato in data 22/01/2025.

[*] Ricercatore, Università degli Studi di Udine, email: gabriele.giacomini@uniud.it.

[†] Dottoressa in Scienze filosofiche, email: chiara.aprilis@gmail.com.

1. Introduction

The increasing integration of artificial intelligence (AI) into decision-making systems, both corporate and public, raises crucial questions about algorithmic biases, with direct implications for fairness and equality in our societies. As a matter of fact, the consequences of discrimination associated with the use of AI can be considerable, leading to alter people's job opportunities, the information environment, access to social services, even, in the most extreme cases, to establish the life and death of people.

Defining "algorithmic bias" is not a straightforward matter, as evidenced by the still ongoing efforts within this respect.[1] The nature of algorithmic bias, that appear to be embedded in machine learning systems, is mainly classifiable as either statistical or societal[2]. The first concerns technical issues related to the quality or quantity of training data; they can usually be detected in the steps of Machine Learning Pipeline. The second concerns social or cultural biases reflected in the data, which are more difficult to correct as they require in-depth ethical, social and political analysis. In both cases, the risk is that of penalisation of marginalized social categories on the basis of gender, religion, race or ethnicity, sexual orientation. These social groups are underrepresented in the high-tech sector. In other words:

> Bias occurs when certain group, such as racial or ethnic minority, rural and socioeconomically disadvantaged populations, are missing from the data. The potential for bias is also perpetuated by the lack of women and racial ethnic minorities in the technology fields because their ideas, perceptions, and values may not be represented the development, training and deployment of algorithmic tools[3]

The unfair outcome can take an allocative form – the bias result both from the withholding of some opportunity or resource, and the unfair distribution of goods across groups – or a representational one –that is, the systematic representation of some group in a negative light, or in a lack of positive representation.[4] Several authors have given other kinds of distinctions among type of bias[5].

This contribution aims to explore the manifestations of bias in AI, highlighting how they can influence decisions ranging from personnel selection to public administration, from moderation of online content to the use of automatic weapons. Given society's growing dependence on AI and the difficulties of decoupling from it without facing serious socioeconomic consequences, the importance and urgency of

---

[1] R.S. Baker, A. Hawn, *Algorithmic Bias in Education*, in «International Journal of Artificial Intelligence in Education» 32,4, 2022, pp. 1052-1092.

[2] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, K. Lum *Algorithmic Fairness: Choices, Assumptions, and Definitions*, in «Annual Review of Statistics and Its Application», 8, 2021, pp. 141-163: https://doi.org/10.1146/annurev-statistics-042720-125902.

[3] N.H. Williams, *AI and Healthcare: The Impact of Algorithmic Bias on Health Disparities*, Springer, Cham 2023.

[4] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, K. Lum, *Algorithmic Fairness*, cit.

[5] R.S. Baker, A. Hawn, *Algorithmic Bias in Education*, cit.

a responsible approach to AI governance is highlighted, both by companies and public institutions, including solid ethical principles, transparency, fairness. Possible strategies to mitigate bias and promote an informed use of AI are then discussed: addressing the problem of bias requires a structured approach that includes technological, ethical and regulatory solutions. It also requires focusing on the importance of greater algorithmic transparency, inclusiveness in the development of AI design, and on the education of decision makers and the population.

2. A first classification of algorithmic bias

In the digital age, the adoption of AI systems is becoming an increasingly common practice. While these tools promise increased efficiency, they also raise significant concerns about the risk of discrimination and bias. Events such as the use of algorithms for personnel selection that result in gender discrimination, or the adoption of algorithmic systems within judicial and welfare contexts that unfairly disadvantage specific social groups, highlight the risks associated with an uncritical use of artificial intelligence. The following examples explore cases of algorithmic bias in various areas of human and social experience.

The use of artificial intelligence systems in personnel selection mechanisms has been spreading for about a decade. The aim is to improve the recruiting process in qualitative terms, first of all by making it more efficient in terms of time, for example through letting the algorithm selecting the shortlist of profiles most in line with the company, relieving the recruiters from many tedious tasks such as searching for and examining dozens – if not hundreds – of CVs and sending standard responses to candidates, or placing adverts to attract candidates. This is done through the use of a wide range of tools, including, for example, chat-bots that interface with candidates, providing essential guidance and sending routine messages to those who are not in line with the company open positions. Artificial intelligence mechanisms can help create a complete profile of the candidate through the analysis of verbal and written language, often compared with that of current employees, or profile their character and emotional aspects through social media profiles[6]. Furthermore, the idea driving the process of technicalisation of selection is to get rid of human judgement as the sole criterion, in favour of a presumed greater objectivity of the technological tool.

However, the use of artificial intelligence in personnel selection has raised numerous questions. The series of problems begins with the sensitivity of the data processed by AI, often of a private and personal nature, which are processed for purposes that go beyond the interests of their owner. There is an asymmetry problem in the exchange of personal information to the extent that the candidate is somehow "forced" to provide his or her data in order to have a chance of obtaining the position.

[6] B. Dattner, T. Chamorro-Premuzic, L. Schettler, *The Legal and Ethical Implications of Using AI in Hiring*, in «Harvard Business Review», 25 April 2019: https://hbr.org/2019/04/the-legal-and-ethical-implications-of-using-ai-in-hiring.

Moreover, some personal information may be statistically deduced by algorithms without the candidate workers' knowledge[7] .

The issue of algorithmic bias is relevant here, as it potentially undermines people's careers and thus their economic prospects. AI mechanisms are not 'neutral' agents, rather the contrary. The case of the algorithm implemented by Amazon that caused a sensation in 2015 shows how the patterns identified by these systems lead to selecting candidates who are similar to the most successful employees in their company, in other words, *cloning your best people* is the outcome most likely[8]. Adopting AI mechanisms is not always rewarding for the company, as demonstrated by the case of Amazon, where the system selected candidates not so much because of technical-IT skills but on the basis of the person's gender, ending up discriminating against women. This outcome resulted from the fact that the algorithm had been trained with data from the profiles of former candidates who were now career employees, for the most part members of a particular category of people that is predominant in the hi-tech sector: males and whites. The system therefore penalised CVs that contained the word 'female' and those related to it, while it positively evaluated those containing verbs and words related to the male gender. The company then decided to abandon the programme[9].

Amazon's is probably one of the most emblematic cases, but there are numerous that have led to discrimination based on disability, ethnicity or even neighbourhood, as well as on the basis of information present on social networks[10].

The last two decades have seen an increasingly massive adoption by governments around the world of automated systems in administrative and criminal bureaucracy. While some benefits have been undoubted, the application of these systems to sensitive areas such as the health, judicial or social services systems has led to instances of injustice.

An investigation conducted by *Wired* and *Lighthouse* has shed light on the mechanisms of discrimination that have affected some entitled to municipal benefits in the Dutch city of Rotterdam. The municipal administration had devised an algorithm to classify potential defrauders of the public subsidy allocation system. Over the course of its use – between 2017 and 2021 – the machine learning algorithm generated risk indices for each of the 30,000 subsidy claimants and, city officials

[7] I. Ajunwa, R. Schlund, *Algorithms and the Social Organisation of Work*, in M. D. Dubber, F. Pasquale, S. Das, *The Oxford Handbook of Ethics of AI*, Oxford University Press, Oxford 2020, pp. 804-822.

[8] I. Ajunwa, R. Schlund, *Algorithms and the Social Organisation of Work*, cit., p. 808.

[9] J. Dastin, *Insight: Amazon scraps secret AI recruitment tool that showed bias against women*, in «Reuters», 11 October 2018: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/.

[10] For instance, if the algorithm detected a postcode linked to a run-down or bad neighbourhood, it would also negatively label the candidate. See R. Krish, *The Pros and Cons of AI in Recrutiment*, in «The Research Nest», 5 December 2018: https://medium.com/the-research-nest/the-pros-and-cons-of-ai-in-recruitment-19c141d1c4b7.

investigated the individuals from these results[11]. During its use, hundreds of people, generally belonging to categories on the margins of Dutch society, were falsely accused or suspected of fraud: the algorithm systematically attributed women and minorities a greater possibility of cheating the subsidiary system.

The article created by *Wired* presents the case of a woman of Moroccan origins, divorced mother of three children, recipient of social allowances. The woman in question was allegedly investigated for the first time after resigning from her job for health reasons; following the investigation, she was deprived of her benefit for the first time, being forced to ask for loans and food from neighbours, as well as pushing her sixteen-year-old son, still a student, to find a job in order to support himself. Moreover, after two years, she was summoned by the social services department and subjected to an "interrogation" where, since she had submitted the wrong bank statement, she was once again deprived of benefits for a few weeks. This lady had been reported as a "high risk" due to her status as a woman, single mother and of foreign origins.

Instead, as far as criminal justice is concerned, a famous investigation published in 2016 by the investigative agency *ProPublica* investigative agency revealed that the system for determining the risk of recidivism used in the United States was decidedly discriminatory against African Americans. The software, named COMPAS, systematically attributed twice the risk of recidivism to African Americans compared to whites, although it was later contradicted by the facts[12]. These cases reveal that, at the root of the accusations perpetrated against the weakest sections of the citizenry, there is a cultural problem that technology does not create but rather increases.

The impact of automatic decision making on marginalized groups is extensively treated by Virginia Eubanks in *Automatic Inequality: How High-Tech Profile, Police and Punish the Poor*, who argues that technologies reflect American culture that stigmatize poor people, regarded as the cause of their own misery, and hence criminalized and even dehumanized. Indeed, as shown by the above-discussed Rotterdam case, the approach towards social benefit recipients is often paternalistic, as people are put under investigation, often undergoing invasive practices for their privacy, divided among those who are morally deserving and those who are not; the latter being punished. Eubanks brings three cases-study that support her thesis; they display cases of erroneous benefit withdrawal, use of predictive model to target which children might be in danger of abuse or neglect while over-investigating lower-classes and black people, finally cases in which algorithms decide who get houses and who remain homeless. The way in which these systems – named "digital poorhouse" in the book – are being used threatens our democracy, warns the author, as they

[11] M. Burgess, E. Schot, G Geiger, *This Algorithm Could Ruin Your Life*, in «WIRED», 6 March 2023: https://www.wired.com/story/welfare-algorithms-discrimination/.

[12] J. Angwin, J. Larson, S. Mattu, L., Kirchner, *Machine Bias, There's software used across the country to predict future criminals. And it's biased against blacks*, in «ProPublica», May 2016: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

undermine the principles of liberty, equity of treatment and value as well as inclusion[13].

The problem of racial discrimination is all the more relevant as the current situation sees the exacerbation of intercultural conflicts and a worrying growth of racial hatred on the part of certain sections of the population. A further problem is that most systems are obscure and hidden from the eyes of citizens[14], who can hardly monitor the situation and possibly defend themselves.

The growth in the flow of information and content on online platforms in recent decades has raised the question of protecting users from offensive or harmful content. In the field of content moderation, automated AI systems are becoming collaborators of human moderators, increasing the capillarity of intervention. Furthermore, they help protect human moderators, who may suffer significant psychological and emotional damage as a result of their activity[15].

However, even in this field, algorithms can make discriminations. First of all, AI mechanisms to date are unable to grasp well the nuances of meaning and idioms of human language, nor to operate a cultural contextualisation of the content, and thus to distinguish what may be offensive in a culture and not in another. Second, they are subject to reproducing the biases of those who program them (usually Western males).

For example, as reported in an article in the *New York Times*, in 2017 the popular YouTube platform removed thousands of videos posted by Syrian activists documenting the atrocities of the Syrian war, resulting in the loss of important testimonies. The platform had recently activated a system that automatically deleted content that did not comply with its guidelines. The system was designed to identify videos posted by extremist (specifically Islamic) groups, but ended up including, quite indiscriminately, any content coming from Syria and the conflict zones where terrorists operated[16]. Similarly, in 2020, *Instagram* algorithm censored posts and profiles related to the activity of the *Black Lives Matter* movement, citing the protection of the community as the motivation. The platform immediately recognised the mistake: the exponential growth of related content had activated the mechanism to prevent spam: so that it apologised and asserted its support for the cause[17] .

There are many examples that can be given in this regard, pointing out that while social platforms give everyone the opportunity to express themselves, the

---

13 V. Eubanks, *Automatic Inequality: How High-Tech Profile, Police and Punish the Poor*, St. Martin's Press, New York 2018.

14 M. Burgess, E. Schot, G. Geiger, *This Algorithm Could Ruin Your Life*, cit.

15 *Use of AI in Online Content Moderation*, in «Cambridge Consultants», Ofcom 2019.

16 M. Browne, *YouTube Removes Videos Showing Atrocities in Syria*, in «New York Times», August 2017: https://www.nytimes.com/2017/08/22/world/middleeast/syria-youtube-videos-isis.html.

17 A. Griffin, *Instagram users trying to post about Black Lives Matter see 'action blocked' messages*, in «Independent», June 2020: https://www.independent.co.uk/tech/instagram-action-blocked-fix-get-rid-how-message-black-lives-matter-error-spam-a9543716.html.

potential for empowerment is not equally distributed or uniformly accessible[18]. Platforms have made a lot of progress and continue to improve their algorithms[19], but the issue remains sensitive because platforms have primarily economic purposes, which means that they do not necessarily have incentives to act in accordance with ideals of equality, justice, protection of minorities and collective well-being.

Speaking about empowerment, an important issue is education. Algorithms have been applied also to this field, to estimate dropout predictions, automated essay scoring, graduate admission, knowledge inference. However, these statistics have shown various degrees of inaccuracy because education is often treated as a homogenous phenomenon, while diversities in class composition should be considered. In particular, there has been insufficient research into intersectionality[20] in educational work on algorithmic bias[21]. This problem should be addressed, as education is one of the most important aspects for empowering one's life. Racism and sexism are embedded in the architecture of technology, according to Noble. Much of her work has been devoted to how technology and the information infrastructure perpetuate specific narratives and make profit from it. She began her research after having noted that the search result for "black girls" lead to hypersexualised and pornographic content[22], to then explore how people of colour are negatively or not-positively represented in the media[23], the underemployment of Black and Latinos by high-tech firms, the commercialization of identities that renders search engine a place of disinformation while deemed reliable, and how this has a role in shaping culture and society. Also, this structure challenges the very idea that marginalised people have of themselves, with a bad impact on their self-esteem, capacity for auto-determination as well as of their life's opportunities, to come back to the concept of self-empowerment. Noble argues that these "algorithmic discriminations" are not glitches in the system, as high-tech representatives claim, in fact they derived from the biases and prejudices carried by programmers – male and white – and exploited by the companies to make profits, hence being the way the system operate.

---

18 D. Endres, L. Hedler, K. Wodajo, *Bias in Social Media Content Management: What Do Human Rights Have to Do with It?* in «Cambridge University Press», June 2023.

19 *Use of AI in Online Content Moderation*, cit.

20 The term, coined by K. Crenshaw, *Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics*, «University of Chicago Legal Forum», vol. 1989, Article 8, describes the condition felt by those who belong to two or more marginalized categories, such as being female, black and lesbian. They therefore experience a discrimination that is more than the sum of the single discrimination.

21 R.S. Baker, A. Hawn, *Algorithmic Bias in Education*, cit.

22 S.U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, New York University Press, New York 2018, pp. 17-22.

23 Noble cites the comparison of African American to apes, the privileged depiction of Whites when the key words linked to "white" are searched as opposed to "black," "jew", "Asian", the association of black names to crimes, and so on.

The problem of bias codification manifests itself in all its radicality in the case of automatic weapons, where a bias can determine the life or death of a multitude of people. Automatic weapons use AI mechanisms to identify targets to hit, they are triggered by humans who, however, are in most cases unaware of what the target is, how it was identified and even when and where it will be shot down. This makes it particularly complex to control automatic weapons, also because the strategy developed by the system is often a black box for those who would try to access it[24]. Furthermore, the system continues to learn while in use, risking further escaping the goals and understanding of those who designed it[25]and amplifying the *biases* already contained within it.

In order to target the Gaza Strip in 2024, the IDF (*Israel Defence Forces*) has developed a system called "Habsora", translated into English with the term "Gospel", which identifies one hundred targets per day, half of which are hit[26]. Despite the Israeli government's declarations about the accuracy of this tool, "Gospel" does not only target Hamas militants: it identifies all individuals accused of potential collaboration, and the areas in which they are presumed to reside, accurately calculating also the number of civilian victims totally unrelated to the terrorist organisation who would be sacrificed. The result is the intensification of the massacre of the civilian population in Gaza, with hundreds of people dying every day but who have little or nothing to do with Hamas[27].

The issue of automatic weapons is problematic in many respects. For instance, if the development of artificial intelligence shows a legacy of the racist colonial system, the issue that some groups against the development of automatic weapons draw attention to is that they could represent a serious danger to ethnic or cultural minorities[28]. In other words, warfare carried out in this way would be the culmination of colonialism: programs are written by members of the dominant group, incorporating their cultural prejudices which, thanks to autonomous learning, risk being amplified by the AI escaping the control of the programmers, making certain sections of the population much more vulnerable.

Algorithmic bias has also become one of the determinants of health in so far as automated decision systems are so intertwined with people lives, influencing the social determinants of health (namely, healthcare and education access and quality,

---

[24] On the black box concept: F. Pasquale, *The black box society: The secret algorithms that control money and information*, Harvard University Press, Cambridge 2015.

[25] *What you need to know about autonomous weapons,* in «International Committee of the Red Cross», July 2022: https://www.icrc.org/en/document/what-you-need-know-about-autonomous-weapons.

[26] *The Gospel: how Israel uses AI to select bombing targets in Gaza*, in «The Guardian», December 2023: https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets.

[27] *"A mass assasination factory": Inside Israel's calculated bombing of Gaza*, in «+972 Magazine», November 2023: https://www.972mag.com/mass-assassination-factory-israel-calculated-bombing-gaza/.

[28] H. R. Jones, *Intersectionality and Racism,* in «Stop Killer Robots»: https://www.stopkillerrobots.org/wp-content/uploads/2021/09/Intersectionality-and-Racism-Hayley-Ramsay-Jones.pdf.

social and community context, economic stability, neighbourhood and built environment)[29]. In fact, although the integration of AI systems in medicine unquestionably better the quality of patient care, not everyone benefits from it: as in the case of automatic weapons, ethnical minorities and black people in particular – but also women – are discriminated because of data under-representation, implicit and explicit bias. Within this respect the research conducted by Obermeyer and alt. found that an algorithm used to foresee the enrolment in health care management programs discriminates against Black people. Even though Black patients had poorer health conditions, the Whites were predicted to need additional care services. The error was not taking into account that Black spent less on health care due to limited resources, hence using less health care management programs than Whites[30]. The evidences that life and well-being of non-White people may be under-attack by automated decision-making systems, make clear that their rights should be carefully taken into account in the development of an ethical AI.

3. The commitment to lead AI out of discriminatory bias

The topics covered highlight the challenges posed by the use of artificial intelligence in various fields, emphasizing not only the potential benefits but also the risks associated with its implementation. From personnel selection and management, through administrative and bureaucratic systems, to online content moderation and the use of automatic weapons, algorithmic biases can lead to discrimination based on gender, ethnicity, social status and other personal characteristics.

Faced with these scenarios, what reassures us is the idea that we can "pull the plug" on AI. However, this may be a misplaced hope. Not only would it be difficult to give up the extraordinary potential of AI: the real problem might be that, as time passes, human society will probably be increasingly dependent on AI, and going backward could have very high costs to face. The costs of the consequences in terms of widespread poverty, crisis of services, and fragility of the economic and social system would risk being greater than the problems for which it is thought to detach AI. This means that both politics and the organisations that produce AI systems now have a responsibility to govern AI and its effects, even if these may prove harmful.

In particular, if companies that develop technology are primarily "for-profit companies", and thus not necessarily incentivised to pursue the common good[31], it is crucial that public regulation and regulatory interventions come into play to ensure

---

[29] N.H. Williams, *Artificial Intelligence and Healthcare: The Impact of Algorithmic Bias on Health Disparities*, Springer, Cham 2023.

[30] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, *Dissecting racial bias in an algorithm used to manage the health of populations*, in «Science» 366, 2019, pp. 447-453.

[31] J. Harris, *There was all sorts of toxic behaviour: Timnit Gebru on her sacking by Google, AI's danger and big tech's biases*, in «The Guardian», May 2023: https://www.theguardian.com/lifeandstyle/2023/may/22/there-was-all-sorts-of-toxic-behaviour-timnit-gebru-on-her-sacking-by-google-ais-dangers-and-big-techs-biases.

that the evolution of AI is aligned with the broader interests of society. This includes the creation of laws and regulations that ensure transparency, fairness, accountability and safety in the use of AI, as well as effective oversight and control mechanisms. The EU seems to be moving in this direction with the Artificial Intelligence Regulation, which aims to establish rules for the development, marketing and use of artificial intelligence in the European Union, seeking to protect the safety and fundamental rights of individuals, classifying AI systems according to the risk they present, and imposing stricter requirements for those considered high-risk[32].

At the international level, however, many experts are working on the identification of principles that should inspire AI and its regulation. Recent working groups include, by way of example, the *Asilomar AI Principles* (2017), the *Montreal Declaration on Responsible AI* (2017), the *Declaration on Artificial Intelligence, Robotics and Autonomous Systems* (2018), the *Five General Principles for an Artificial Intelligence Code* (2018), and the *Ethical Guidelines for Trustworthy AI* (2019). These initiatives highlight the need for joint thinking between researchers, legislators, industries and civil society to ensure that AI development is aligned with fundamental human values. Luciano Floridi, perhaps the world's foremost expert on information and AI ethics, has attempted with his collaborators to summarise the recurring and cross-cutting principles in the documents produced in recent years[33]. They are:

1. Beneficence (AI should promote welfare, preserve dignity and sustain the planet).
2. Non-maleficence (the AI must respect privacy, be secure and avoid misuse).
3. Autonomy (AI must promote the autonomy of humans, who can always choose how and whether to delegate decisions to the machine).
4. Justice (AI must support the prosperity of peoples, preserve solidarity and avoid unfairness).
5. Explicability (the AI should be as transparent and intelligible as possible).

Implementing these principles is not an easy challenge or one with predictable outcomes. In the first place, it is still unclear how to politically promote these general principles: is self-regulation and consumer protection sufficient? Or should the state intervene more directly? Secondly, the principle of explicability is correct from a normative point of view, but it is still unclear to what extent it can be applicable. Floridi emphasises that the obscurity of AI must not provide an "excuse" for digital companies to hide their internal procedures from researchers, supervisory bodies and democratic institutions in general. Floridi is absolutely right, but he seems to forget

[32] D.E. Harris, *Europe has made a great leap forward in regulating AI*, in «The Guardian», December 2023: https://www.theguardian.com/commentisfree/2023/dec/13/europe-regulating-ai-artificial-intelligence-threat.

[33] L. Floridi, *The Ethics of Artificial Intelligence: Principles, Challenges, and Opportunities*, Oxford University Press, Oxford 2023; B. D. Mittelstadt, P. Allo, M. Taddeo, S. Wachter, L. Floridi, *The ethics of algorithms: Mapping the debate*, in «Big Data & Society», 3(2), 2053951716679679.

the fact that AI is a *black box* even for the researchers who built it[34].

Since AI is based on complex statistical relationships that are impossible for a human being to trace, it might be difficult to explain the reasons behind a machine's decision. Yet, the last principle is the most basic: only if AI is explicable can we fully check that it is beneficial, not evil, that it respects the autonomy of humans, and that it promotes social justice. This is why ways of improving the verification of AI systems are being experimented with, such as the use of "stress tests" or the creation of internal "checkpoints" where certain checks can be carried out[35]. Technically, this new field of study is called *explainable AI (XAI),* and it aims to improve the trust and transparency of AI-based systems so that AI continues to make steady progress without interruption. The degree of certainty with which we will be able to ensure that AI does not promote judgements or decisions tainted by discriminatory bias will perhaps depend on these projects.

In addition, as regard bias, there is also the problem of dealing with its harmful effects, in cases where, despite preventive controls, these should nevertheless occur. Who should be held responsible for an action of an AI system that causes harm to individuals or groups of people due to bias? We have seen that the consequences could also be very important in terms of well-being, health, even life and death. Theoretically, it becomes important for there to be "meaningful human control" (the concept "human in the loop" is also used, which envisages the presence of human will and judgement in algorithmic processes[36]). By "meaningful" is meant that purely nominal conditions and forms of human control over the machine are excluded. The control must instead be substantial, i.e. the human agent should be able to express a considered judgement about the operations the system is performing and be able to intervene in good time in the event of unforeseen events. Furthermore, the operator should be trained not to overestimate the capabilities of computer systems.

However, in practice, this is a complex task, perhaps impossible to fully realise, due to the internal opacity concerning complex computational processes and the difficulty of interpreting the information that induces the artificial system to make a certain decision or perform an action. The operator who activated the system might not be able to exercise effective control over its behaviour, due to the cognitive difficulty of understanding the AI's decision-making mechanisms (AI is a *black box*) and the long reaction times of humans that might prove ineffective for timely

[34] N. Cristianini, *The Shortcut: Why Intelligent Machines Do Not Think Like Us*, in «CRC Press», 2023.

[35] On the topic of XAI: A. Adadi, M. Berrada, *Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)*, in «IEEE access», 6, 2018, pp. 52138-52160; F. K., Došilović, M. Brčić, N. Hlupić, *Explainable artificial intelligence: A survey*, in «41st International convention on information and communication technology, electronics and microelectronics», 2018, pp. 0210-0215; G. Vilone, L. Longo, *Notions of explainability and evaluation approaches for explainable artificial intelligence*, in «Information Fusion», 2021, 76, 2021, pp. 89-106.

[36] F. M. Zanzotto, *Human-in-the-loop artificial intelligence*, in «Journal of Artificial Intelligence Research», 2019, pp. 64, 243-252; E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, Á. Fernández-Leal, *Human-in-the-loop machine learning: A state of the art*, in «Artificial Intelligence Review», 56(4), 2023, pp. 3005-3054.

intervention. Ultimately, who may be primarily responsible for bias damage? Potential candidates include: the software engineers who created the AI system, the executives of the company that developed it, the consultants or staff members of the entity that implemented the system, the leaders of the organisation that adopted the system, the head of the department that used the system, or the individual who responsible for directly supervising the system. The risk is to conclude that no one made a truly significant contribution to the occurrence of the damage[37]. Or there is the risk of attributing this diffuse responsibility by law to a single figure, perhaps paid precisely to legally take on any negative responsibility.

These issues are under discussion, but several strategies are beginning to emerge that will need to be explored and tested in the near future. Some strategies are mainly technological. As we have seen, it will be crucial to understand whether and how it is possible to increase the transparency of AI-based algorithms and systems, allowing for independent review and analysis to identify bias. This is obviously related to the importance of investing in research and development of methodologies capable of mitigating and removing existing biases in AI systems. Other strategies, however, concern human beings more directly. We are not only referring to the identification of solid ethical principles and public regulation that place respect for human rights and the protection of minorities and vulnerable groups at the centre. It is also about ensuring that developers of AI systems come from diverse backgrounds to ensure that a variety of perspectives are considered in the design of algorithms, thus reducing the risk of unintended bias. Most importantly, raising awareness and education about the potential pitfalls and biases of AI algorithms will be essential for those who develop them, those who use them and those who are affected by them.

[37] H. Nissenbaum, *Accountability in a Computerised Society*, in «Science and Engineering Ethics», 2, 1, 1996, pp. 25-42.

AI: inevitabile o evitabile, questo (non) è il problema. Ciò che precede la trasparenza algoritmica[a]

Emanuela Tangari[*]

Abstract

L'articolo esplora la relazione tra Intelligenza Artificiale (IA) e fiducia, ponendo l'accento sulla trasparenza e la "trasparibilità" come elementi chiave per l'analisi di un utilizzo etico e responsabile, di cui il contesto medico si pone come caso d'uso privilegiato. Attraverso riferimenti a teorie filosofiche e psicologiche, si analizzano le sfide e le implicazioni delle decisioni autonome delle IA, mettendo in luce il loro impatto sul ragionamento umano. Viene preso in esame il progetto europeo MES-CoBraD per evidenziare i benefici e i limiti dell'applicazione dell'IA in medicina. Il tema centrale rimane la necessità di una trasparenza che superi la mera comprensione tecnica, per abbracciare una comprensione relazionale capace di sostenere una fiducia autentica e un utilizzo della ragione *tout court* nelle decisioni.

Parole chiave: Medicina e Tecnologie, Etica dell'Intelligenza Artificiale, Intelligenza Artificiale, Trasparenza Algoritmica, Fiducia e Affidabilità

Abstract

The article explores the relationship between Artificial Intelligence (AI) and trust, emphasizing transparency and "traceability" as key elements in analyzing the ethical and responsible use of AI, with the medical field serving as a prime use case. Drawing on philosophical and psychological theories, it examines the challenges and implications of AI-driven autonomous decisions, highlighting their impact on human reasoning. The European MES-CoBraD project is analyzed to showcase the benefits and limitations of AI applications in medicine. The central theme remains the necessity for transparency that goes beyond mere technical understanding, embracing

[a] This paper is part of a project that has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 965422. Saggio ricevuto in data 31/05/2024 e pubblicato in data 22/01/2025.

[*] Docente a contratto, Università di Roma "Tor Vergata", email: emanuela.angela.tangari@uniroma2.it.

a relational comprehension capable of fostering genuine trust and the application of reason in decision-making.

Keywords: Medicine and Technologies, Ethics of Artificial Intelligence, Artificial Intelligence, Algorithmic Transparency, Trust and Reliability

1. Introduzione

Nel 2013 veniva pubblicato un articolo dal titolo "The Inevitable Application of Big Data to Health Care"1; a distanza di quasi 10 anni, nel 2022 viene pubblicato "Artificial Intelligence for Health and Care Is Not Inevitable: Introduction and Critical Vocabulary"2. Indipendentemente dalla divergenza di prospettive e dalla posizione – e dalle teorie a supporto – che si vuole assumere, una cosa ci appare certamente inevitabile: la domanda aperta sull'Intelligenza Artificiale e sulle sue implicazioni, oggi il tema forse più presente nel dibattito filosofico, ingegneristico, matematico, sociale, giuridico.

Tra i campi d'applicazione più discussi – in cui da sempre la cooperazione tra esseri umani e tecnica è al centro – c'è senza dubbio quello della medicina. Non vogliamo qui indagare l'inevitabilità dell'utilizzo della AI in medicina; piuttosto, una premessa nota ma non così ovvia può essere quella di domandarsi che cosa possa significare non tanto una "intelligenza artificiale", ma piuttosto a quali condizioni le pratiche (mediche e non) sostenute dai sistemi di Intelligenza Artificiale o dai sistemi basati sugli algoritmi, o le norme, le decisioni – al di là che esse siano automatizzate o no – possano definirsi intelligenti, e soprattutto a partire da quali criteri e quali principi. Su tali principi è necessaria un'analisi che sposti l'attenzione da un piano formale e normativo a un piano fondamentale, cognitivo, che parta (e ritorni) alla condizione umana nel suo accadere.

Uno degli elementi caratteristici di tale condizione umana nel suo accadere è l'essere in relazione: il fatto che gli esseri umani esistono in un contesto di reciprocità, in un contesto esistenzialmente e socialmente correlato, e in una situazione storico-sociale in cui l'individualità sorge e si sviluppa. A ciò si aggiunge il fatto che il contesto storico-sociale in questione è un contesto permeato dalla tecnica-tecnologia; è quindi un contesto storico-tecno-sociale, in cui la tecnica non ricopre più solo il ruolo di artefatto, ma acquista lo statuto – almeno fenomenologico, anche se non si voglia considerare quello assiologico – di agente tra agenti, e non solo di un prodotto. E così perveniamo al fattore forse più rilevante della riflessione: a comporre il sostrato che regola il vivere comune, le relazioni, vi sono numerose dinamiche. Prendiamo qui in

[1] T.B. Murdoch, A. S. Detsky, *The inevitable application of big data to health care*, in «JAMA», 309, n. 13, 2013, pp. 1351-1352. https://doi.org/10.1001/jama.2013.393.

[2] R. Walker, *Artificial Intelligence for Health and Care Is Not Inevitable: Introduction and Critical Vocabulary,* in J. Dillard-Wright, J. Hopkins-Walsh, B. Brown (a cura di), *Nursing a Radical Imagination Moving from Theory and History to Ac-tion and Alternate Futures,* Routledge, London 2022.

esame una di queste, che si pone come centrale in riferimento al contesto storico-tecno-sociale.

La questione della fiducia e dell'affidabilità diventa uno snodo essenziale nel dibattito sulla tecnologia3, pur non trovando in questo dibattito il suo inizio; prende invece in causa problemi morali, virtù, predisposizioni, dinamiche sociali e intersoggettive che vengono molto prima delle loro applicazioni. Su che cosa si fonda la fiducia? quando possiamo dire che qualcosa o qualcuno è affidabile? e che rapporto c'è tra la categoria di affidabilità e l'esperienza – fenomenologicamente intesa – della fiducia? Non basta che un soggetto o un agente sia ritenuto (da altri o da noi stessi) affidabile affinché si generi l'*esperienza* della fiducia. Questo mostra l'intricato rapporto tra i due termini e apre ad un'altra dimensione fondamentale, quella dell'intenzionalità, delle volontà, dei fini. Nel campo dell'ingegneria tecnologica, sono proprio tali intenzioni – e dunque la costruzione e la strutturazione dei modelli, dei software, degli strumenti – a coinvolgere, a discesa, quegli elementi che vengono poi messi al centro della discussione: la trasparenza, la responsabilità, l'equità, il diritto alla privacy, l'inclusione, l'imparzialità, e le numerose altre sfere implicate.

La trasparenza nell'uso delle tecnologie digitali, specialmente in ambito medico, è fondamentale non solo per la fiducia degli utenti ma anche per garantire una partecipazione informata e consapevole da parte degli individui coinvolti: non si tratta dunque di un mero requisito tecnico ma anche un principio etico che deve guidare lo sviluppo e l'implementazione delle nuove tecnologie. La trasparenza, nel contesto delle tecnologie digitali, si riferisce alla disponibilità di informazioni sui processi interni di un sistema. Un sistema trasparente permette agli utenti di capire come le decisioni vengono prese, quali dati vengono utilizzati e quali criteri vengono applicati. La trasparenza è essenziale per garantire che i sistemi siano utilizzati in modo equo e responsabile. La trasparibilità – cui si farà cenno come fattore distinto dalla trasparenza e maggiormente in riferimento al funzionamento del ragionamento umano – si riferisce alla capacità di un sistema di rendere visibili i suoi processi interni e i suoi risultati. Questo include la possibilità di tracciare le fonti dei dati, comprendere le logiche di funzionamento degli algoritmi o dei processi psichici, e accedere ai dati che spieghino le decisioni. Nel contesto delle AI – ma anche del ragionamento umano –, della loro verifica e validazione, la trasparibilità diventa cruciale per descrivere il tema della fiducia. Nel campo delle tecnologie digitali, tale fiducia è strettamente legata alla trasparenza e alla spiegabilità, vale a dire alla possibilità di interagire con sistemi che operino in modo prevedibile, sicuro, comprensibile, responsabile, socialmente accettabile.

2. L' εὐδαιμονία aristotelica e l'euristica della scelta

Un passo indietro (o di lato): si è iniziato chiedendo quando la medicina, o un altro dominio, può definirsi "intelligente". La domanda può qui riguardare una certa

[3] O. O'Neill, *Trust and Accountability in a Digital Age*, in «Philosophy», 95, n. 1, 2020, pp. 3-17.

declinazione dell'intelligenza, cioè la razionalità: quando, cioè, una decisione può definirsi razionale. La storia della filosofia è costellata da tale problema, dalla filosofia antica a quella contemporanea, passando per la filosofia moderna e l'analisi della ragione, da Cartesio a Kant, senza conoscere la fine del problema. Già Aristotele nell'*Etica Nicomachea* descriveva la razionalità – la razionalità "pratica" – come mezzo, mediazione, per conseguire uno scopo o per soddisfare un desiderio. Prendiamo in prestito proprio la prospettiva aristotelica, particolarmente utile ad evidenziare quanto l'accezione che si attribuisce ad un termine o alla sua realizzazione sia significativa nella costruzione del giudizio "morale".

Sono necessarie due precisazioni: 1. qual è, per Aristotele, il fine ultimo di ogni azione. Questa domanda appare secondaria nella riflessione attuale sull'Intelligenza Artificiale: si trova dunque un ampio dibattito su che cosa sia quest'ultima e quali siano le sue implicazioni e i suoi impatti, ma la questione si allontana sempre di più dalla domanda originaria su quale sia l'"intelligenza" (cioè la ragione, il fine) che guida l'azione intelligente. 2. Il sillogismo che descrive l'azione razionale non può prescindere dall'elemento onnipresente in ogni movimento umano: quello della libertà, della volontà, o anche solo del libero arbitrio. Libertà di che cosa? verso dove si volge (dovrebbe volgere) la deliberazione? Verso la felicità, εὐδαιμονία. Questa εὐδαιμονία è l'accordo con l'ἀρετή, l'eccellenza: proprio dell'umano è quindi l'esercizio delle virtù, secondo Aristotele – etiche e dianoetiche –, l'esercizio razionale accordato all'eccellenza, e volto al bene. Tralasciando qui la riflessione sull'interpretazione descrittiva o normativa del modello deliberativo, è utile considerare, dopo aver sottolineato l'importanza del tema dei fini, un ulteriore aspetto nel discorso della razionalità, e quindi dell'intelligenza, sia essa umana o "artificiale": l'atteggiamento, il comportamento, la predisposizione specificamente umana nell'esercizio della razionalità. Se le azioni umane intelligenti o razionali seguissero un fine preciso e identificato (qualsiasi esso sia, e a maggior ragione se esso fosse il "bene"), non sarebbe possibile spiegare la gamma di contraddizioni, dissidi, controsensi che le scelte e le decisioni umane comportano.

Gli studi sulla "razionalità limitata" (bounded rationality) e sui meccanismi cognitivi che sono alla base delle decisioni umane e dei loro biases4 contraddicono l'idea di una razionalità logicamente intesa, ed evidenziano invece la potenza di una euristica irrazionale, tipicamente umana, che sottende l'agire individuale e collettivo. Secondo Daniel Kahneman5, meccanismi cognitivi come l'ancoraggio, la rappresentatività, la disponibilità, sono i motori dei giudizi ritenuti razionali, ma che poco hanno a che fare con il "sistema 2", quello logico, deliberativo, probabilistico, e molto di più ineriscono al "sistema 1", il pensiero "veloce", intuitivo. La stessa capacità computazionale e statistica della mente umana è soggetta necessariamente alla rappresentazione che una parte dell'io – o del sé – le restituisce: questo è il motivo per cui la memoria non raccoglie informazioni secondo la loro quantità, ma secondo la loro potenza: il

[4] A. Tversky, D. Kahneman, *Judgment under Uncertainty: Heuristics and Biases*, in «Science», 185, n. 4157, 1974, pp. 1124-1131.

[5] D. Kahneman, *Thinking, Fast and Slow*, Penguin Books, London 2012.

ricordo di un singolo evento può essere assai più vivido – e quindi più determinante e statisticamente più probabile in una futura scelta – di una somma di informazioni ripetute, ma ritenute meno incisive.

In chiave più analitica, Donald Davidson nel 1986 si dedicava alla delineazione del rapporto tra il fondamento dell'interpretazione e quello della conoscenza, affermando che la «conoscenza empirica […] nasce, piuttosto, dalla natura dell'interpretazione. Come interpreti dobbiamo trattare l'attribuzione autonoma di credenze, dubbi, desideri e preferenze come privilegiata; questo è un passo essenziale nell'interpretare il resto di ciò che una persona pensa e dice. La fondazione dell'interpretazione non è la fondazione della conoscenza, anche se una comprensione della natura dell'interpretazione può portare alla comprensione della natura essenzialmente veridica delle credenze»[6]; a distanza di anni, nel 2001, uno dei suoi lavori vede emergere un contrasto con la razionalità pratica basata sulla decision theory: «la teoria della decisione corrisponde alle nostre intuizioni su come le reali decisioni vengono prese, ed è parte del nostro apparato di senso comune per spiegare il comportamento intenzionale»7; ciò si fonda sul presupposto che le azioni che gli esseri umani compiono in maniera intenzionale sono guidate da credenze e desideri, che conferiscono valore all'azione stessa.

3. L'impatto della AI sul ragionare umano

Alla luce di questi brevi cenni su alcune delle prospettive concernenti la teoria della scelta e il funzionamento della razionalità umana, si può tornare a chiedersi che cosa accade quando l'agire (o il ragionare) umano può servirsi di una "intelligenza", di un ragionare – se così può dirsi – artificiale, che non possiede intenzionalità alcuna. Il problema sembra situarsi non solo nell'ipotetica decisione che un sistema decisionale autonomo può prendere, ma ancor prima – e in maniera ancora più determinante – nel modo in cui il ragionare umano può essere inficiato, alterato dalle decisioni autonome. Non si tratterebbe allora di un parallelismo, classicamente inteso, tra la decisione umana e quella artificiale, ma di un circuito in cui il sistema decisionale autonomo interferisce, "compromette" non solo e non necessariamente la singola decisione, ma il modo stesso di ragionare.

Qui appare in maniera predominante il tema della trasparenza, di seguito discusso, che non intendiamo allora solo nei riguardi delle singole decisioni o dello

[6] D. Davidson, *A coherence theory of truth and knowledge*, in E. LePore (ed), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, Blackwell, Cambridge 1986, p. 332: «[…] empirical knowledge […] springs, rather, from the nature of interpretation. As interpreters we have to treat self-ascriptions of belief, doubt, desire and the like as privileged; this is an essential step in interpreting the rest of what the person says and thinks. The foundations of interpretation are not the foundations of knowledge, though an appreciation of the nature of interpretation can lead to an appreciation of the essentially veridical nature of belief».

[7] D. Davidson, *Subjective, Intersubjective, Objective*, Oxford University Press, Oxford 2001, p. 126. https://doi.org/10.1093/0198237537.001.0001.

specifico processo del particolare strumento, ma nei riguardi dell'interazione tra le due intelligenze, tra le due razionalità: quella umana e quella artificiale.

Secondo Kahneman la stessa razionalità umana è assai complessa da descrivere; le scelte apparentemente razionali sono sovente deliberate dal sistema veloce, quello dell'intuizione, del ragionamento "pratico", che si basa o può basarsi su elementi anche ancestrali, inconsapevoli, non risolvibili da una logica linguisticamente descrivibile. Se è così complicato decifrare la natura del ragionamento umano, che ne è di quello (programmato – almeno finora – da individui umani, ma con una crescente dose di "autonomia") artificiale? Individuare, e tanto più descrivere, la *ratio* dei sistemi decisionali autonomi sembra impresa ardua, quanto più questi sistemi sono stratificati e complessi, non solo per la quantità di dati ma soprattutto per la loro tipologia e per i fattori discriminanti che dettano la scelta.

Fortemente interessante sarebbe indagare, da un punto di vista filosofico e psicologico, il modo in cui i sistemi decisionali artificiali compromettono la decisione umana, o meglio il modo in cui essi interagiscono col sistema decisionale umano; in che modo, quindi, la relazione umani-macchine modifica non tanto il risultato finale, ma il processo stesso in cui le decisioni umane procedono. Sembra quindi configurarsi, nell'universo del "Sistema 1 e 2" di Kahneman, un terzo sistema, che non si configura come sintesi o una negazione dei primi due e la cui rilevanza sta invece nel fatto di intercettare, modificare, penetrare intrinsecamente il funzionamento stesso dei primi due sistemi. Per questo il tema dell'Intelligenza Artificiale – in qualsiasi dominio di applicazione la si osservi – diventa decisivo, poiché non si pone al livello dei risultati o degli esiti, ma della strutturazione stessa del ragionamento, come anche della percezione umana, e della relazione degli esseri umani con il mondo, con gli altri, e con sé stessi. Sarebbe interessante per esempio osservare in che modo le AI mediche capaci di generare diagnosi o di proporre protocolli di cure si interpongano nel giudizio (presente e futuro) dei medici; in che modo cioè vadano a modificare non solo la decisione singola che di volta in volta viene presa, ma l'autocoscienza e l'abilità stessa dei professionisti, il rapporto tra le conoscenze acquisite e la capacità di decisione, l'esperienza di osservazione dei casi clinici, la strutturazione delle competenze.

Tralasciamo qui le questioni, note quanto importanti, del dibattito, che comprendono la considerazione della responsabilità8, dei pregiudizi, del controllo finale sui dati, o dell'equità delle scelte effettuate dai sistemi di Intelligenza Artificiale; di quanto, inoltre, siano realmente autonomi i sistemi di AI. Sulla scia di quanto accennato, soffermiamoci invece sull'aspetto della *trasparenza* e della spiegabilità delle decisioni, degli algoritmi che guidano le decisioni.

[8] K. Baum, S. Mantel, E. Schmidt, T. Speith, *From Responsibility to Reason-Giving Explainable Artificial Intelligence*, in «Philosophy & Technology», 35, n. 12, 2022. https://doi.org/10.1007/s13347-022-00510-w.

4. Trasparenza e "trasparibilità": una visione e non solo una ragione

La trasparenza è un tema centrale della discussione sulla creazione e sul funzionamento dei sistemi di AI9, specialmente di quelli decisionali, che possono per loro "natura" produrre discriminazioni e acuire divari già esistenti. La considerazione dei bias e degli algoritmi *black box* ripropone continuamente la necessità di lavorare ad una trasparenza che lasci intendere all'utente i criteri attraverso i quali il sistema giunge alla scelta10, e che tali criteri pongano alcuni fini o diritti umani – come l'equità, la "non-maleficenza", la responsabilità – come priorità del sistema stesso.

Di nuovo, emerge anche qui la difficoltà di rintracciare una definizione univoca e universale di "equità", per esempio, o di livello di spiegabilità che il sistema deve prevedere, o i destinatari per i quali esso deve risultare spiegabile. Si rinviene anche qui la difficoltà di pervenire ad un significato condiviso (e quindi applicabile) di trasparenza algoritmica, perché questo prevede in primo luogo la definizione di quali siano gli scopi e i soggetti coinvolti, e contemporaneamente dei criteri convenzionalmente condivisi sia per il processo di lettura delle previsioni/decisioni, sia soprattutto per un accordo, se possibile, su che cosa significhi l'interpretabilità. Se questa cioè ha a che fare con la possibilità di descrivere il processo di scelta e selezione, a partire dagli elementi forniti; e se tale descrizione dovrebbe garantire la comprensione, per esempio. Sorge allora la domanda su che cosa significhi comprendere, e ancor meglio su che cosa significhi comprendere per un essere umano (o per un gruppo, una società di individui). La comprensione, come la fiducia, non può essere facilmente disgiunta da fattori personali e spesso inconsci; tuttavia è proprio da questa comprensione – o comprensibilità – che prendiamo delle decisioni, che direzioniamo la scelta; la comprensione, come la fiducia che si accorda a qualcosa o qualcuno, sembra possedere i tratti di una intuizione, o di una "visione", propria degli individui umani, che non sempre e non facilmente è possibile descrivere con una logica algoritmica, matematica.

Si tratta di una conoscenza per così dire "morale", che non riguarda la conoscenza dei dati e delle informazioni ma di quelle intenzioni e fini cui lo strumento tende e da cui dipende. La trasparenza, dunque, come la fiducia, implica una relazione e non semplicemente una nozione, una disposizione di informazioni. Ciò è implicito nel concetto stesso di conoscenza, e dunque di comprensione e di spiegabilità: non basta che qualcosa sia spiegabile, o spiegato, affinché sia compreso. La conoscenza richiede anch'essa un fatto – il fatto di un rapporto che nasce tra il soggetto che conosce e l'oggetto conosciuto, nella sua natura e nel suo fine più profondo — e non semplicemente un dato (la potenziale spiegabilità o conoscibilità dell'oggetto).

---

[9] Non si porrà qui una distinzione particolare tra Machine Learning e Deep Learning; l'attenzione è invece su questioni di fondamento, su dilemmi e sfide per una riconfigurazione non solo tecnologica, ma anzitutto soggettiva, che ha a che fare con la professionalità e con il compito soggettivi.

10 W.J. von Eschenbach, *Transparency and the Black Box Problem: Why We Do Not Trust AI*, in «Philosophy & Technology», 34, 2021, pp. 1607-1622. https://doi.org/10.1007/s13347-021-00477-0.

Il tema problematico della trasparenza, e della trasparenza tecnologica, fonda su questo terreno le sue radici. La prima domanda, ancora una volta, è che cosa si intende per trasparenza. Quando nel processo conoscitivo e comprensivo umano qualcosa si ritiene trasparente; se sia una questione di accessibilità, di presa di possesso dei dati che costituiscono il contenuto dell'oggetto (sia esso una decisione, un algoritmo, un processo), o piuttosto di intelligibilità tecnica di questi. Forse entrambi questi fattori – insieme ad altri – contribuiscono alla percezione della trasparenza, fermo restando che 1. così come per il binomio fiducia-affidabilità, non è sufficiente che qualcosa sia *de facto* trasparente da un punto di vista tecnico perché sia anche comprensibile; 2. tale intelligibilità diviene problematica nel momento in cui l'elaborazione dei dati avviene attraverso l'apprendimento automatico, e non solo come esito della programmazione operata dall'essere umano: da qui l'idea di opacità o di inaccessibilità delle black-box11.

5. Un caso studio europeo: MES-CoBraD

Tra i numerosi casi di studio e di applicazione dell'AI in campo medico in cui le suddette questioni trovano spazio e necessità di essere attenzionate[12]. prendiamo in esame quello condotto nel progetto europeo MES-CoBraD Multidisciplinary Expert System for the Assessment & Management of Complex Brain Disorders), un interessante caso di Deep Learning utilizzato per la classificazione delle immagini (ad esempio espressioni geniche, scansioni MRI, ecc.). Coordinato dalla National Technical University of Athens, il progetto ha l'obiettivo di migliorare la precisione diagnostica e i risultati terapeutici nelle persone affette da disturbi cerebrali come i disturbi neurocognitivi (demenza), i disturbi del sonno e le crisi epilettiche (epilessia), nonché le loro interconnessioni.

L'Intelligenza Artificiale può porsi come strumento per migliorare e potenziare le funzioni cognitive umane dei medici che trattano pazienti affetti da malattie complesse13. Questi progressi richiedono un'analisi approfondita del modo in cui l'adozione di tali tecnologie sta influenzando i concetti di malattia, cure mediche, pratica clinica e la relazione di cura all'interno delle società. Di particolare rilievo sono i nuovi approcci basati sulle Reti Generative Avversarie14. che mirano ad ampliare il pool di dati medici disponibili per l'addestramento dei sistemi di apprendimento automatico capaci di riconoscere specifici tipi di cancro tramite immagini. Queste tecniche, note

---

[11] J. Burrell, *How the machine 'thinks': understanding opacity in machine learning algorithms*, in «Big Data & Society», 2016, pp. 1-12. https://dx.doi.org/10.2139/ssrn.2660674.

[12] J. Hatherley, *Limits of trust in medical AI*, in «Journal of Medical Ethics», 46, n.7, 2020, pp. 478-481. https://doi.org/10.1136/medethics-2019-105935.

[13] S.A. Bini, *Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: what do these terms mean and how will they impact health care?*, in «The Journal of Arthroplasty», 33, n. 8, 2018, pp 2358-2361. https://doi.org/10.1016/j.arth.2018.02.067.

[14] F.H.K. do Santos Tanaka, C. Aranha, *Data augmentation using GANs*, in «arXiv», 2019. https://doi.org/10.48550/arXiv.1904.09135.

come Data Augmentation15. presentano sfide significative dal punto di vista normativo, poiché i dati utilizzati per l'addestramento dei sistemi di Intelligenza Artificiale sono generati sinteticamente da altri sistemi di Intelligenza Artificiale.

La piattaforma MES-CoBraD è pensata per essere fruibile dalla più ampia comunità di utenti, ed essere conforme agli standard medici e sociali, legali ed etici suggeriti dalla Commissione Europea. Tra gli obiettivi del progetto vi è proprio la promozione del benessere e l'attenzione (e prevenzione) ai potenziali rischi, che comprende anche la condivisione dei dati reali (RWD) in una prospettiva di trasparenza e affidabilità, perseguite con la riduzione dei bias nei processi decisionali e garantendo che la responsabilità rimanga centrata sull'essere umano e sulle strutture sanitarie coinvolte, e incoraggiando i fruitori a seguire i codici internazionali di etica medica e delle pratiche per l'Intelligenza Artificiale.

Tra le altre ambizioni del progetto, vi è quella di supportare lo sviluppo e l'uso dell'AI per garantire che tutti possano essere curati e assistiti in strutture sanitarie che facciano uso di sistemi di Intelligenza Artificiale "sani", anche al di fuori del campo dell'imaging sanitario, attraverso un approccio "human-in-the-loop", attraverso il quale i professionisti medici e i pazienti possano essere coinvolti in un processo in cui la soluzione sia regolata, addestrata e testata su misura.

6. Trasparenza razionale, trasparenza artificiale

È doveroso chiedersi quale sia la distinzione tra la trasparenza (o meglio trasparibilità) artificiale e quella umana nei processi decisionali16. Secondo l'*argomento di uguale opacità*17 18, non dovrebbe sussistere una differenza radicale tra il modo in cui vengono giustificate le decisioni umane, in cui – data l'impossibilità di descrivere ogni passaggio della scelta umana – vi è una razionalizzazione a posteriori, e quello utilizzato dai sistemi ADM (Automated Decision-Making). Non si intende qui entrare nel merito di una critica dell'argomento di eguale opacità19 – molto interessante, perché pone tra gli altri il problema dell'autoregolazione e della modellazione del pensiero umano – ma si vogliono indicare i problemi di fondo e le implicazioni, e suggerirne una chiave di lettura.

Come accennato, l'opacità di alcuni sistemi di Intelligenza Artificiale è attribuita all'impossibilità di controllare o di seguire passo dopo passo il codice che produce un certo risultato, generando una sorta di "irriducibilità strutturale" di

---

[15] C. Shorten, T.M. Khoshgoftaar, *A survey on image data augmentation for deep learning*, in «Journal of big data», 6, n. 1, 2019, pp. 1-48. https://doi.org/10.1186/s40537-019-0197-0.

[16] J. Zerilli, A. Knott, J. Maclaurin, C. Gavaghan, *Transparency in algorithmic and human decision-making: is there a double standard?* in «Philosophy & Technology», 32, n. 4, 2019, pp. 661-683. https://doi.org/10.1007/s13347-018-0330-6.

[17] C. Buckner, *Black boxes or unfattering mirrors? Comparative bias in the science of machine behaviour*, in «British Journal for the Philosophy of Science», 74, n. 3, 2023, pp. 681-712.

[18] Burrell, *How the machine 'thinks'*, art. cit.

[19] U. Peters, *Explainable AI lacks regulative reasons: why AI and human decision-making are not equally opaque*, in «AI Ethics», 3, 2023, pp. 963-974. https://doi.org/10.1007/s43681-022-00217-w.

quest'ultimo e del suo processo20. Questo è direttamente proporzionale alla stratificazione e alla complessità del codice e della computazione algoritmica chiamata in causa nel sistema dell'AI e nel suo scopo, e spesso direttamente proporzionale anche alla sua efficienza. In questi casi diviene difficile individuare la correlazione tra i dati, le informazioni, e i risultati proposti, malgrado la natura intrinsecamente intellegibile di quei dati. Si dirà che lo stesso – su un altro livello – accade con la computazione umana, con i processi e con le scelte umane: abbiamo infatti notato come la razionalità o l'apparato decisionale non procede in modo lineare, logico; i movimenti psichici e della ragione si servono di dinamiche spesso inaccessibili alla coscienza o alla memoria umana, e tuttavia si dimostrano nella maggior parte dei casi *efficaci*.

D'altro canto, contrastare l'opacità del sistema decisionale (sia esso artificiale o soggettivo) necessita di considerare l'elemento dell'intenzionalità, tipico degli esseri umani, a cui inerisce la fiducia e la comprensibilità[21].

Valutare l'accuratezza delle previsioni o delle decisioni tra l'uno e l'altro sistema appare estremamente complesso: da un lato vi è la scarsa trasparenza dei sistemi decisionali autonomi per le ragioni già esposte; d'altro canto, però, le decisioni prodotte dall'Intelligenza Artificiale non possiedono il carattere di "autosabotaggio" o di condizionamento che la mente umana è capace di produrre e spesso (se non sempre) produce rispetto alle proprie scelte, inficiate da fattori psicologici, storici, esperienziali; l'opacità si colloca quindi negli esseri umani, e nel processo decisionale della mente, su un altro livello, ma è tuttavia presente e non facilmente calcolabile o prevedibile, data l'estrema complessità dei meccanismi psichici. A ciò si aggiunge un altro fattore discriminante tra i due sistemi: malgrado la possibile e frequente incoerenza tra la decisione umana, le sue spiegazioni, e i dati, gli elementi reali da cui la decisione è scaturita, il fatto stesso che i soggetti giustifichino una specifica decisione delineandone le ragioni produce una sorta di coerenza *post-hoc*, per la quale gli individui sono portati ad agire in linea con le spiegazioni fornite.

L'impegno alla coerenza – alla stregua dell'"effetto alone" descritto da Kahneman, per il quale la mente tende ad ignorare le informazioni che non sono allineate con la storia o con la percezione costruita – non può evidentemente modificare l'origine delle cause da cui le decisioni iniziali provengono, ma può certamente intervenire in questo processo e alterare l'automatismo di quelle cause. Da ciò deriverebbe, in ultima analisi, una maggiore trasparenza fondata su un movimento contemporaneamente passato e futuro in cui le intenzioni, le autoaffermazioni e le convinzioni (anche a posteriori) che gli individui umani attribuiscono alle scelte incidono concretamente sulla stabilità e sulla coerenza futura della decisione, e dunque della sua prevedibilità22.

---

20 P. Zellini, *La dittatura del calcolo*, Adelphi, Milano 2018.

21 A. Ferrario, M. Loi, E. Viganò, *Trust does not need to be human: it is possible to trust medical AI*, in «Journal of Medical Ethics», 47, 2021, pp. 437-438. https://doi.org/10.1136/medethics-2020-106922.

22 L. De Bruin, D. Strijbos, *Does confabulation pose a threat to first-person authority? Mindshaping, self-regulation and the importance of self-know-how*, in «Topoi», 39, 2020, pp. 151-161. https://doi.org/10.1007/s11245-019-09631-y.

7. Razionalità o ragionevolezza: un divario dialogico

Ancora, ma su un altro piano, è possibile mostrare che – per quanto opaca – la ragione e l'intuizione umana possiedono una sofisticatezza forse non sufficientemente spiegabile, ma ancora ragionevole e affidabile.

L'elemento problematico sottostante a queste complesse considerazioni e valutazioni sembra essere sempre il tipo di attribuzione di significato, di definizione condivisa o di valore che si associa agli elementi e ai termini presi in esame. È possibile, per esempio, paragonare la "spiegazione" fornita o ricavabile da un sistema di AI a quella che si può *esigere* da un essere umano? Se non ci si sofferma qui sulla richiesta – sull'aspettativa, strettamente umana in questo caso – che si ha nei confronti dell'uno e dell'altro sistema, qualsiasi comparazione risulta peregrina. Se la costruzione di un sistema di AI può poggiare su strutture designate a rispettare alcuni elementi essenziali (*ethics by desing*, etc.) che a loro volta rendono possibile una sorta di comparazione dei risultati, è altrettanto vero, o reale, che ad oggi le aspettative e le relazioni che le persone hanno con le Intelligenze Artificiali non sono (ancora?) le stessa di quelle che richiedono agli altri esseri umani. Questo si manifesta in termini di responsabilità e sensibilità che ci si aspetta. Rispetto agli esseri umani, non ci si limita a cercare razionalità o spiegabilità, ma anche una una *ragionevolezza* da intendersi come una comprensione profonda e completa dell'oggetto o soggetto coinvolto. Tale ragionevolezza è una caratteristica tipica delle interazioni umane, in quanto coinvolge proprietà sia personali che interpersonali, che emergono nella relazione tra individui.

Non lo sono, inoltre, dal punto di vista dell'interazione stessa: il rapporto che gli esseri umani intrattengono – e il campo medico ne è o ne dovrebbe essere un campo esemplificativo – è un rapporto dialogico, relazionale. Il rapporto dialogico richiede la possibilità di porre domande e ricevere risposte; ma, soprattutto, di porre *reciproche* domande.

Al di là, dunque, del funzionamento del sistema mentale in confronto a quello della AI, vi è dunque il funzionamento della relazione umana in confronto a quella che si può stabilire con i sistemi di AI; questi ultimi non domandano: o meglio, non domandano in modo *creativo*. Non vi è per l'AI la possibilità di porre domande affettive (se non per simulazione), e dunque *affettivamente intuitive* – quali sono quelle sullo stato emotivo e sul modo in cui ciascuno significa la propria esperienza – che nascono cioè dalla condizione stessa della relazione e sono impossibili al di fuori di essa.

Con questo arriviamo ad una maggiore delineazione della suddetta questione della ragionevolezza come capacità di osservazione della complessità dei fattori. Tale ragionevolezza, a differenza della razionalità, comprende quell'alveo di affetto (nel senso latino dell'*affectus*, dell'essere *affetti-da*) che genera una imprevedibile capacità di comprensione, impossibile alla sola razionalità. Da qui una abilità a porre domande che non derivano dalla sola associazione di dati e informazioni, ma da una esperienza che trova nella relazione stessa una sempre nuova vitalità, una ri-generazione continua

che si fonda sulla ragione affettiva, e che quindi risulta impossibile all'Intelligenza Artificiale la quale, per ora, può solo simulare.

È proprio la ragionevolezza a segnare anche la distinzione tra la comprensibilità, l'accettabilità delle decisioni che derivano da sistemi di AI e quelle prodotte dagli esseri umani, o tra un essere umano ed un altro; così come per il binomio affidabilità/fiducia, la ragionevolezza della decisione non dipende solo dalla coerenza intrinseca ai dati che la sottendono: una decisione può essere perfettamente logica se derivata dai suoi elementi, eppure parziale, insoddisfacente e, in ultima analisi, "disumana". Disumana non nell'accezione di "sbagliata" o "malevola", ma poco consona alla complessità – alla ragionevolezza, appunto – che gli esseri umani esigono e di cui sono capaci. Qual è allora l'elemento – o gli elementi – costitutivi di questa ragionevolezza? La possibilità che il risultato, di qualsiasi ordine e contesto esso sia, possieda appunto una "ragione". Una ragione che abbia i tratti non solo di una spiegazione, ma di un significato, di un motivo. La possibilità, dunque, di comprendere, o almeno intuire, qualcosa che non rientri nella categoria della casualità.

Una siffatta ragione non può essere altro che una ragione *poetica*, laddove per poetica si intende ciò che arriva al fondo di una esigenza, a cui solo l'esperienza – e non solo il linguaggio, il calcolo, la spiegazione può rispondere. Qui sembra esservi la radice della vera responsabilità: capace di rispondere sottraendosi al dominio della casualità, all'inquietudine che da essa si genera, di rendersi portavoce di una risposta, di una intuizione, di una proposta, non solo di un *come*, ma di un *perché* che sostenga l'esigenza di ragionevolezza della decisione, del fatto, del dato, del fine.

The Impossibility of Transparent Social Robots[a]

Giovanna Di Cicco[*]

Abstract

La trasparenza è emersa come uno dei concetti più rilevanti nel dibattito etico che circonda diversi ambiti, tra cui la robotica sociale. Questo articolo esplora il modo in cui la trasparenza si applica ai robot sociali e se possa essere uno strumento efficace per proteggere gli interessi degli utenti da potenziali inganni e dinamiche ambigue implicate nelle interazioni tra esseri umani e robot. L'articolo traccia una distinzione preliminare tra la trasparenza intesa come proprietà della robotica sociale e la trasparenza intesa come attributo dei robot sociali, evidenziandone i diversi significati e implicazioni. La discussione si concentra poi sulla trasparenza dei robot sociali e viene fatta un'ulteriore distinzione tra *trasparenza sui robot sociali* e *trasparenza attraverso i robot sociali*. Partendo dalla descrizione dei tre tipi di inganno proposti da John Danaher, l'inganno di stato interno, messo in atto da robot sociali che mostrano facoltà e stati emotivi che in realtà non hanno, viene identificato come la forma più costitutiva di inganno coinvolta nelle interazioni con i robot sociali. Questo aspetto viene poi considerato alla luce dell'antropomorfismo, per esaminare la progettazione di robot trasparenti, che dovrebbero attenuare le risposte antropomorfiche come possibile rimedio per proteggere gli interessi degli individui ed evitare l'inganno. Tuttavia, poiché l'antropomorfismo sembra essere il fondamento stesso della socialità percepita dai robot, è impossibile rinunciare al loro comportamento ingannevole senza rinunciare anche al loro ruolo sociale. Ciò porta, infine, a sostenere che un robot sociale veramente trasparente non è realizzabile e che la trasparenza non è sufficiente a garantire una robotica sociale responsabile.

Parole chiave: robot sociali, trasparenza, antropomorfismo, pregiudizi cognitivi, inganno dei robot, teoria della mente, interazione umano-robot, etica della tecnologia, roboetica, implicazioni etiche.

[a] Saggio ricevuto in data 30/05/2024 e pubblicato in data 22/01/2025.
[*] Dottoranda, Università degli Studi di Genova – Northwest Italy Philosophy PhD Program (FINO), email: giovanna.d.cicco@gmail.com.

Transparency has emerged as one of the most relevant concepts in the ethical debate surrounding several fields, and social robotics is one of them. This paper explores how transparency relates to social robots and whether it could be an effective tool to protect users' interests from potential deception and misleading dynamics involved in human-robot interactions. The paper outlines a preliminary distinction between transparency understood as a property of social robotics and transparency understood as an attribute of social robots, highlighting their different meanings and implications. The discussion, then, focuses on the transparency of social robots, where a further distinction is drawn between *transparency on social robots* and *transparency through social robots*. Starting from the description of three types of deception proposed by John Danaher, internal state deception, enacted by social robots that display faculties and emotional states they do not really have, is identified as the most constitutive form of deception involved in interactions with social robots. This is then considered in the light of anthropomorphism, to examine the design of transparent robots, which should mitigate the anthropomorphic responses as a possible remedy to protect the interests of individuals and avoid deception. However, since anthropomorphism appears to be the very foundation of robots' perceived sociality, it is impossible to forego their deceptive behaviour without also foregoing their social role. This leads, finally, to argue that a genuinely transparent social robot is not achievable, and that transparency is not enough to ensure a responsible social robotics.

Keywords: social robots, transparency, anthropomorphism, cognitive bias, robot deception, theory of mind, human-robot interaction, ethics of technology, roboethics, ethical implications.

Introduction

Recent advancements in the field of social robotics and artificial intelligence are bound to change the way human beings engage with reality and with each other. This novel context challenges the traditional tools we use to understand others' behaviour and discern between genuine and fake, reality and simulation. While regulations and institutions seem to struggle to cope with evolving technologies and to defend the interests of users, the ethical debate faces unprecedented issues and questions.

One main ethical concern raised about social robotics is that human-social robot interactions might be inherently deceptive and inauthentic, as they provide the illusion of robots being something they are not and having attributes they do not actually have. Although social robots are not necessarily humanoid or human-like, they usually display evocative features, such as certain facial expressions, proxemic and postural attitudes or vocal tones. They are designed to perform tasks focused on interacting with human beings, behaving as credible social actors, and eliciting empathic and emotional reactions. Therefore, they behave *as if* they had emotions, intentions, preferences, or goals, where the words "*as if*" precisely reflect the dimension of simulation involved. Some of the risks of such deception are those

related to privacy and information sharing, the building of unidirectional bonds or trust, the misinterpretation of robots' behaviour, and the information and power asymmetry between users and companies[1].

Understanding the implications of deceptive practices involved in social robotics and developing strategies to defend the interests of users has become a key issue in the technology ethics debate. *Transparency* has then emerged as an increasingly relevant concept and one of the most advocated strategies[2]. Indeed, the AI HLEG, established by the European Union, identifies it as one of the principles for a sustainable and trustworthy artificial intelligence[3].

However, the great success of this concept comes with an equal amount of uncertainty regarding its understanding and applications. This results in what Emmanuel Alloa describes as a *magic concept*, characterised by a great normative attractiveness and an exceedingly positive connotation, yet presenting multiple and overlapping definitions[4]. As scholars point out, despite the relevance assigned to transparency in the ethical debate on social robotics, there is currently a lack of an extensive literature or agreed definitions[5].

Therefore, in order to understand whether transparency is a suitable and sufficient strategy to avoid deception and ensure a sustainable development for social robots, this paper will try to clearly define what transparency means for social robotics, how it interacts with different forms of deception perpetrated by social robots, and what are the limits of its application.

1. Why transparency and which transparency

Firstly, it is worth noting that the relevance of transparency in the technology ethics debate is part of a broader flourishing of this concept in the 21st century. Such notion has been regarded as a socio-political tool to serve democracy, playing a major role in the fight against corruption and stimulating responsible and informed decision-making.[6]. However, Alloa highlighted that transparency can be associated with

---

[1] R. Wullenkord & F. Eyssel, *Societal and Ethical Issues in HRI*, in «Current Robotics Reports», 1, 2020, pp. 85-96.

[2] See S. Turkle, *Authenticity in the age of digital companions*, in «Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems», 8, n. 3, pp. 501-517; P.G.R. de Almeida, C.D. dos Santos & J.S. Farias, *Artificial Intelligence Regulation: a framework for governance*, in «Ethics and Information Technology», 23, 2021, pp. 505-525; and A. Jobin, M. Ienca & E. Vayena, *The global landscape of AI ethics guidelines*, in «Nature Machine Intelligence», 1, 2019, pp. 389-399.

[3] AI HLEG - High-Level Expert Group on Artificial Intelligence, *The Assessment List for Trustworthy Artificial Intelligence* (ALTAI), 17 July 2020.

[4] E. Alloa, *Transparency: A magic concept of modernity*, in E. Alloa & D. Thomä (eds.), *Transparency, Society, Subjecticity. Critical Perspectives*, Palgrave Macmillan, London, 2018, pp. 21-55: 29.

[5] A. Theodorou, R.H. Wortham & J.J. Bryson, *Designing and implementing transparency for real time inspection of autonomous robots*, in «Connection Science», 29, n. 3, pp. 230-241.

[6] J.C. Bertot, P.T. Jaeger, J.M. Grimes, *Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies*, in «Government Information Quarterly», 27, n. 3, 2010, pp. 264-271: 264.

different aspirations and contexts, such as gaining access to information, safeguarding justice, providing accountability, and encouraging virtuous conduct.[7]. Therefore, when we talk about transparency in social robotics, we can group the various nuances of the concept into two main original understandings: *transparency of social robotics*, as an attribute of such field of research and production, and *transparency of social robots*, as an attribute of social actors.

Transparency of social robotics refers more precisely to information transparency, achieved through disclosure procedures that allow information, data, or behaviour to be visible and understandable. It thus relates to the possibility for governments and general public to see practices and activities underlying the research and production of social robots. Some authors have emphasised, for instance, the importance of knowing the working algorithms and rules of AI[8] and the methods and data that have been used in their training[9]. Other significant information concern how the data collected by social robots is processed, the business operations of the producing companies, the power dynamics in which they are involved and the operational and procedural processes of their activities.

On the one hand, this openness provides a greater understanding of robotic technologies and related risks, allowing individuals to be more conscious in their use. On the other hand, it exposes the actions of companies to the judgement of public and laws, allowing for a greater scrutiny of their legitimacy and to hold them accountable for their decisions. At the same time, it seems to be a potential instrument of moralisation and self-regulation that could induce restraint and best practices[10]. If companies are forced to provide details and reasons for their actions, then what they do is there for all to see and they are much more likely to act in a virtuous manner.

Understanding transparency in this way does not present unique peculiarities related to social robotics and can be traced back to the debate on transparency in the socio-political perspective. Moreover, it is worth noting that Transparency as an alternative to stricter regulation or as a regulation in itself has been questioned and shows limits in its application and outcomes[11].

Transparency of social robots, instead, refers to the user's ability to clearly understand the artificial social partner, so as to accurately grasp the functioning of the robot he or she interacts with. On a pragmatic level, this could foster human-robot cooperation by ensuring a safer and more effective use of the robot, such as knowing when to consider it reliable or is acting unexpectedly, how it makes decisions, or how

---

[7] E. Alloa, *Transparency: A magic concept of modernity*, cit., pp. 31-32.

[8] M.C. Buiten, *Towards Intelligent Regulation of Artificial Intelligence*, in «European Journal of Risk Regulation», 10, n. 1, 2019, pp. 41-59.

[9] M. Butterworth, *The ICO and artificial intelligence: The role of fairness in the GDPR framework*, in «Computer Law & Security Review», 34, n. 2, 2018, pp. 257-268.

[10] E. Alloa, *Seeing Through a Glass Darkly. The Transparency Paradox*, in E. Alloa (Ed), *This Obscure Thing Called Transparency. Politics and Aesthetics of a Contemporary Metaphor*, Leuven University Press, Leuven, 2022, pp. 9-25: 12.

[11] A. Etzioni, *The Limits of Transparency*, in E. Alloa & D. Thomä (eds.), *Transparency, Society, Subjecticity. Critical Perspectives*, Palgrave Macmillan, London, 2018, pp. 179-201.

to interpret its behaviour.[12]. On a more ethically relevant level, the main benefit of transparent robots is considered to be the elimination, or reduction, of users' deception, making him less vulnerable to the possible exploitation of his trust and emotional response.

Schött, Amin and Butz points out that transparency of social robots can be understood either as *transparency on the robot*, where information is provided from outside, or as transparency *through the robot*, where information is integrated into the design itself[13]. Transparency on the robot can be achieved, for instance, through what is conveyed by marketing, advertising, instruction manuals or websites. Transparency through the robot refers to constitutive elements embedded in the robot design itself, which thus becomes the source of transparent information. This can be done in an explicit manner, such as by having the robot remind the user that it is an artificial entity, that it has no feelings, that it does not belong to any gender or that it cannot answer questions about its pretended past or its emotional states[14]. But it can also occur implicitly, when the robot is constructed in such a way that it does not resemble a human being, when it appears obviously mechanical, or has a voice clearly recognisable as artificially synthesised[15].

In both transparency on the robot and transparency through the robot, then, the goal is to provide a look at the reality that lies beyond the social appearance of the artificial agent, beyond the *as if* it performs. To aim for a transparent robot means to aim for a robot that is accurately perceived by the user, who can clearly recognise its artificial nature and real properties. In this way, transparency would become the way to avoid manipulation of users and preserve them from developing inappropriate and one-sided emotional responses or bonds. However, to understand whether this is the case, it is worth investigating what we mean when we talk about deception involved in interactions with social robots and how this responds to attempts at transparency.

2. Social robots, deception and anthropomorphism

There are several ways for a robot to deceive users and John Danaher has specifically outlined three[16].

1) *External state deception* occurs when the robot deceives the user on something that does not concern the robot itself by providing false information. As Danaher points out, external state deception is similar to cases where humans lie, so it follows

[12] A. Theodorou, R.H. Wortham & J.J. Bryson, *Designing and implementing transparency for real time inspection of autonomous robots*, cit., pp. 232-234.
[13] S.Y. Schött, R.M. Amin, R.M. Butz, *A literature survey of how to convey transparency in co-located human-robot interaction*, in «Multimodal Technol. Interact.», 7, n. 25, 2023, p. 9.
[14] B. Leong & E. Selinger, *Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism*, in «Proceedings of the Association for Computing Machinery's Conference on Fairness, Accountability, and Transparency», 2009, pp. 299-308: 307.
[15] C. Balkenius & B. Johansson, *Almost Alive: Robots and Androids*, in «Frontiers in Human Dynamics», 4, 2022, pp. 1-7: 5-6.
[16] J. Danaher, *Robot Betrayal: a guide to the ethics of robotic deception*, «Ethics Inf. Technol.», 22, 2020, pp. 117-128: 121.

the same moral principles[17]. Therefore, in this context transparency, in terms of information about the design process, can be crucial in ensuring that artificial agents are constructed in such a way that they never lie to the user.

2) *Hidden state deception*, instead, occurs when the robot possesses certain capabilities, but it keeps them hidden from users by omitting or denying them. They might include hidden recording devices or undeclared personal data retention. Again, transparency on the robot plays a key role here, as it allows users to have full knowledge about all the functionalities of the robot they interact with. Users should be aware of the possibility of audio or video recording, of how the robot handles personal information it collects, and what kind of physical force it may exert. As with the previous case, we do not consider it morally acceptable to take advantage of trust or naivety of individuals in order to covertly act against their interest, and neither should it be acceptable for social robots.

In both cases, if users have access to comprehensive and meaningful information about the robot, they are given the tools to rationally choose the best way to deal with it.

3) *Superficial state deception*, finally, includes cases where the robot pretends to have abilities or internal states that it does not actually possess. This form of deception is particularly relevant for social robots, since their very ability to pose as social actors, and create relationships with humans, relies on the simulation of feelings and emotional states capable of evoking an empathic response. Such simulation is not always necessarily a malicious tactic against users, but represent a fundamental design element, which is ultimately useful for the legitimate tasks the robot is designed to perform. In recent years, studies of human-robot interaction (HRI) have played a vital role in understanding how individuals respond to artificial agents and how to improve their interactions so that they become as friendly and natural as possible[18]. To understand how and whether transparency can be useful in defending individuals against superficial state deception, it is then imperative to look at the cognitive and behavioural dynamics it involves.

3. Superficial state deception is not a choice

HRI pragmatic experiments have shown that humans tend to apply to interactions with robots the same social norms and inferences they apply to interactions with living beings[19]. Subjects were shown to attribute meaning and intention to the behaviour of

[17] *Ibid.*

[18] L.T. Cordeiro Ottoni & J. de Jesus Fiais Cerqueira, *A Review of Emotions in Human-Robot Interaction*, 2021 Latin American Robotics Symposium (LARS), 2021 Brazilian Symposium on Robotics (SBR), and 2021 Workshop on Robotics in Education (WRE), Natal, Brazil, 2021, pp. 7-12.

[19] C. Nass, J. Steuer & E.R. Tauber, *Computers are Social Actors*, in «Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 1994)», Association for Computing Machinery, New York, NY, USA, pp. 72–78.

social robots[20] and to place them within social categories, to the point of transferring prejudices or notions, such as gender or identity, onto them[21]. By encouraging such tendencies, social robotics hopes to produce artifacts capable of performing roles traditionally reserved to conscious beings, such as those of care robots, hospitality robots or sex robots. At the same time, it seems to somehow encourage subjects' inaccurate representation of reality, potentially leading them to experience inauthentic relationships.

This has led some authors to believe that social robotics engages in outright deception to the detriment of the user, leading them to indulge in empathic and emotional attachments that are not justified. Robert Sparrow describes it as an excessive sentimentalism of users, who are induced to violate the *prima facie* duty to pursue an accurate representation of reality[22]. The perspective held by Sparrow, however, is challenged by Mark Coeckelbergh, who suggests considering the illusion carried on by social robots not as a deception but as a performance, similar to the one of magic shows[23]. In this view, thus, designers and users are in a relationship resembling the one between magicians and their audience, where they cooperate in maintaining the illusion they voluntarily take part in.

Both sides of the argument, anyway, consider the successful illusion operated by social robots to some voluntary disposition of the subject and thus fail to grasp the problematic core of the issue. HRI studies point out that the creation of empathic and emotional bridges with social robots largely rests on human cognitive mechanism of anthropomorphism: the tendency to attribute human properties and mental states, such as emotions, motivations, or intentions, to nonhuman entities. This emerges both through the analysis of behavioural results and by looking at neurophysiological findings and brain activity reports[24]. Understanding the dynamics of anthropomorphism, then, highlights that the illusion underlying human-robot interactions is not resulting from a choice, but from a conditioned response.

---

[20] E. Schellen, F. Bossi & A. Wykowska, *Robot Gaze Behavior Affects Honesty in Human-Robot Interaction*, in «Front. Artif. Intell.», 4, 2021; M. Salem, F. Eyssel, K. Rohlfing, S. Kopp & F. Joublin, *To Err is Human(-like): Effects of Robot Gesture on Perceived Anthropomorphism and Likability*, in «Int. J. Soc. Robot», 5, 2013, pp. 313- 323.

[21] See S.J. Stroessner & J. Benitez, *The Social Perception of Humanoid and Non-Humanoid Robots: Effects of Gendered and Machinelike Features*, in «International Journal of Social Robotics», 11, 2019, pp. 305-315; J. Bernotat, F. Eyssel & J. Sachse, *The (Fe)male Robot: How Robot Body Shape Impacts First Impressions and Trust Towards Robots*, in «International Journal of Social Robotics», 13, 2021, pp. 477-489.

[22] R. Sparrow, *The March of the robot dogs*, in «Ethics and Information Technology», 4, n. 4, 2002, pp. 305-318.

[23] M. Coeckelbergh, *How to describe and evaluate "deception" phenomena: recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn,* in «Ethics Inf. Technol.», 20, 2018, pp. 71-85.

[24] G. Di Cesare, F. Vannucci, F. Rea, A. Sciutti & G. Sandini, *How attitudes generated by humanoid robots shape human brain activity,* in «Scientific Reports», 10, n. 16928, 2020.

4. *Anthropomorphism and theory of mind: how social robots trick us*

Anthropomorphism is usually regarded as a cognitive *bias* that leads us to misinterpret the behaviour of nonhumans, inferring inaccurate causes for it. As human beings, we tend to detect something human in every thing, and this constitutes a structural element of how our minds work, which can be even found in ancestral forms, such as pareidolia[25].

Anthropomorphism is understood as a part, typically considered improper, of the more general human faculty of adopting others' point of view and to imagine what they might be feeling or thinking. This ability, often called "theory of mind" (ToM), refers to the possibility of creating explicit meta-representations of others' mental states, inferring beliefs, motivations, or goals, so that their condition can be evaluated according to their own parameters[26]. To imagine how others might experience a certain situation is different from imagining how we will experience that same situation[27]. It is this very act of perspective taking that underlies the peculiarity of empathic experiences in human beings.

Empathy can be understood as a neurobehavioural process with evolutionary underpinnings, emerging in a variety of human and non-human animals, and consisting of a spontaneous response to specific external stimuli[28]. However, human empathy seems to extend to more situations and individuals. Sometimes we experience empathy for strangers, individuals who are distant in time and space, or even fictional characters and animals of other species. When we adopt the other's point of view, it is no longer relevant who we are or what relationship we have with our social counterparts. ToM allows us to access what lies behind their external behaviours through a spontaneous inferential mechanism, which associates those behaviours with the inner states generating them[29]. However, since we can never have immediate access to internal states of others, this inference is inevitably grounded in our own personal existence, in how we experience those internal states as individuals and as human beings.

On the one hand, this implies that, although excessive human-likeness is shown to produce a feeling of uncanny and discomfort[30], a general resemblance to human beings is more likely to trigger anthropomorphism. Indeed, HRI research

---

[25] L.F. Zhou & M. Meng, *Do you see the 'face'? Individual differences in face pareidolia*, in «Journal of Pacific Rim Psychology», 14, 2020.

[26] V. Stone, *The moral dimensions of human social intelligence*, in «Philosophical Explorations: An International Journal for the Philosophy of Mind and Action», 9, n. 1, pp. 55-68.

[27] C.D. Batson, S. Early & G. Salvarani, *Perspective taking: Imagining how another feels versus imagining how you would feel*, in «Personality and Social Psychology Bulletin», 23, n. 7, 1997, pp. 751-758.

[28] J. Decety, G.J. Norman, G.G. Berntson & J.T. Cacioppo, *A neurobehavioral evolutionary perspective on the mechanisms of underlying empathy*, in «Prog. Neurobiol.», 98, 2012, pp. 38-48.

[29] V. Stone, *The moral dimensions of human social intelligence*, op. cit.

[30] M. Mori, K.F. Macdorman & N. Kageki, *The Uncanny Valley*, in «IEE Robotics & Automation Magazine», 19, n. 2, 2012, pp. 98-100.

reports that we relate, empathise, and trust artificial agents more easily when they have humanoid attributes, both on an aesthetic and behavioural level[31].

On the other hand, the more distant the entity is from us, either socio-culturally or evolutionarily, the more likely it is that the inference is inaccurate and that he or she experiences or externalizes those inner states differently from us. But whereas in the case of other living beings we are faced with the unrealizable attempt to understand the nature of their inner life, social robots are produced by us, so we may know how they operate.

When it comes to social robots, anthropomorphism leads us to an incorrect inference, since, as Paul Dumouchel points out, their behaviours cannot really be interpreted as read-outs of any internal state but are «signs without referents»[32]. The empathic and emotional response of individuals towards social robots does not depend on a defect of reason or will, nor on false beliefs or intentional participation in an illusory reality[33]. Subjects involved in empirical experiments and users of social robots are aware that they are dealing with machines without internal states but tend to respond empathically to them regardless. Such responses, therefore, cannot be seen as a choice. Instead, they are to be understood as spontaneous and pre-reflexive cognitive mechanisms, which are deliberately elicited through specific design and marketing strategies.

5. The impossible transparent robot: inherent limits of transparency in social robotics

Having clarified the effects of anthropomorphically inspired design on our cognitive biases, we can evaluate how effective transparency of social robots is in preventing individuals from being manipulated.

Transparency on the robot can certainly be a regulatory requirement to ensure that companies and research do not convey misleading information and provide a truthful representation of social robots, at least on a theoretical level. Authors have often expressed this need and highlighted problematic human-washing (or machine-washing) practices carried out by companies[34]. In analogy to the concept of greenwashing, human-washing describes the strategy of companies to deliberately manipulate their communications by creating a symbolic veil; a misleading façade that generates information asymmetry and portrays social robots as more competent, harmless, or similar to us. Demanding transparency from companies about the real properties of social robots, therefore, means removing the opacity of this façade,

---

[31] See M. Li & A. Suh, *Machinelike or Humanlike? A literature Review of Anthropomorphism in AI-Enabled Technology*, in «Hawaii International Conference on System Sciences», 2021; A. Sacino et al., *Human- or object-like? Cognitive anthropomorphism of humanoid robots*, in «PLoS ONE», 17, n. 7, 2022.

[32] P. Dumouchel, *Making Faces*, in «Topoi», 41, 2022, pp. 631-639: 637.

[33] L. Damiano, P. Dumouchel, *Anthropomorphism in Human-Robot Coevolution*, in «Frontiers in Psychology», 9, n. 468, 2018.

[34] G. Scorici, M.D. Schultz & P. Seele, *Anthropomorphization and beyond: conceptualizing humanwashing of AI-enabled machines*, in «AI & Society», 39, pp. 789-795, 2024.

making it transparent, so that we can access the reality it conceals. However, two considerations should be taken into account.

First, as Alloa points out, just because a medium is transparent does not mean that there is no mediation[35]. Transparency of information is not a given property but something that is made, the result of a process that is never ethically neutral[36]. When information is disclosed, someone has decided on which information, as well as how to interpret and elaborate it to make it understandable and accessible. This process needs to be understood and regulated, so that it does not become a new façade for the sake of marketing.

Second, we have observed that accurate theoretical knowledge is not sufficient to empirically avoid the emergence of misleading dynamics between individuals and social robots. In a sense, we are evolutionarily programmed to interpret robot behaviour through the lens of anthropomorphism.

Transparency through the robot is thus proposed as a strategy to mitigate the effects of anthropomorphism throughout the interaction itself. One example of this strategy has been highlighted by van Straten and Kühne in a study on the interaction between children and social robots, where children's tendency to anthropomorphise and trust robots was found to decrease when they interacted with robots consistently communicating the absence of human psychological capacities[37].

Illusion of transparency, in social psychology, refers to the biased perception that our internal experiences are more visible to others than they really are and that others can perceive our actual personal thoughts, emotions, or mental states. Human beings are never transparent, but we have access to their internal experience through the correct interpretation of their behaviour. Therefore, applying the same notion to social robots, we can conclude that the more the robot is transparent, the more our interpretation of its behaviours should allow us to perceive its lack of inner experience. The design of an entirely transparent robot, then, should convey by the interaction itself that those behaviours are mere simulacra.

However, as seen above, the possibility for the robot to be perceived as a social actor and to create an empathic and engaging interaction is based on the very triggering of anthropomorphism. This means that transparency and perceived sociality of the robot are inversely proportional and to forego the design of social robots with misleading features triggering anthropomorphism is to forego the design of social robots altogether. A genuine transparent social robot is hence not really possible.

---

35 E. Alloa, *Transparency: A magic concept of modernity*, op. cit., p. 36.

36 M. Turilli & L. Floridi, *The ethics of information transparency*, in «Ethics. Inf. Technol.», 11, pp. 105-112: 109.

37 C.L. van Straten, J. Peter, R. Kühne, *Transparent robots: How children perceive and relate to a social robot that acknowledges its lack of human psychological capacities and machine status*, in «Int. J. Human-Computer Studies», 177, 2023.

Conclusion

In the end, transparency alone does not appear to be able to protect individuals from one of the major sources of exploitation risk: the one that comes from directly interacting with social robots and misinterpreting their behaviours. In fact, since such interactions are mediated by the pre-reflective cognitive mechanism of anthropomorphism, transparency of information about how the robot is constituted does not prevent us from empathising with it and ascribing it mental states it does not possess.

I argue that we should embrace this impossibility and use such awareness to shape adequate strategies for regulating social robots in the world. The use of social robots has been shown to be a potential resource in controlled settings, such as in investigating the functioning of human relationships or in treating social disorders[38]. And there might be other cases where social robots are beneficial, so much so that we agree «to conscientiously harness our weird sensibilities so that our instinctual responses work for us and not against our best interests»[39]. If we accept that some degree of deception is always involved in the relationships between humans and social robots, we can begin to engage in discussions about whether, when and how such deception is something we are willing to allow as a society.

This means further investigating the limits of transparency and identifying the empirical consequences of human-robot interactions that are not free of deception. Additional studies and cognitive experiments are needed to determine the potentially disruptive effects of manipulating cognitive biases on how we interact with each other and the possible benefits that might emerge from deception in specific settings. Furthermore, an interdisciplinary dialogue between HRI, engineering, cognitive science and ethics needs to be developed in order to reach a coherent definition of transparency and a viable implementation strategy. Finally, governments and institutions need to produce strong regulations where transparency is not the goal, but a tool to ensure that social robotics meets the standards of such regulations. To do so, transparency is necessary, but is not enough. Instead, we need to regulate how social robot should be designed, for what purpose, and what the alternatives are.

[38] A. Kouroupa, et al., *The use of social robots with children and young people on the autism spectrum: A systematic review and meta-analysis*, in «PLoS One», 17, n. 6, 2022.

[39] B. Leong & E. Selinger, *Robot Eyes Wide Shut: Understanding Dishonest Anthropomorphism*, in «Proceedings of the Association for Computing Machinery's Conference on Fairness, Accountability, and Transparency», 2009, pp. 299-308: 308.

« Qu'est-ce que tu ne comprends pas ? » Jeux de langage et algorithmes boîte noire[a]

Rémy Demichelis[*]

Abstract
L'enjeu de cet article est de déterminer ce qui pose problème dans notre compréhension des algorithmes dits « boîte noire », une problématique propre à la jeune discipline de l'*Explainable Artificial Intelligence* (XAI). Car, s'il est aisé de comprendre quelque chose que quelqu'un nous explique, c'est plus délicat lorsque personne n'arrive à saisir le problème. Cependant, notre propos consiste à souligner : (1) qu'il convient de parler d'*interprétabilité* plutôt que d'*explicabilité* lorsque nous cherchons à comprendre les modèles, principalement parce que nous n'avons jamais un accès complet et sans ambiguïté à l'information ; (2) que la machine fait face au problème de l'inscrutabilité de la référence, de la même manière que le linguiste imaginé par Willard Van Orman Quine ne peut pas déterminer précisément ce que désigne le terme « *gavagai* » dans une situation de traduction radicale ; (3) qu'il n'y a pas de règle pour l'application de la langue, si ce n'est des « *language games* », comme la linguistique de Ludwig Wittgenstein nous l'enseigne. Il en découle que l'espoir d'arriver à une explicabilité des algorithmes, et donc à la transparence attendue, est sans doute vain : nous ne pouvons nous contenter que d'interprétations qui ne mentionneront jamais la règle de la règle.

Keywords: XAI, Explicabilité, Interprétabilité, Linguistique, Jeux de langage, Ethique.

Abstract
The aim of this article is to understand the problem of "black box" algorithms, an issue inherent to the nascent field of Explainable Artificial Intelligence (XAI). While it is relatively easy to understand something someone explained to us, it becomes more complicated when no one can fully grasp the issue. Our purpose is however to highlight: (1) that we should speak of *interpretability* rather than *explainability* when we seek to understand models, mainly because we never have complete and unambiguous

[a] Saggio ricevuto in data 31/05/2024 e pubblicato in data 22/01/2025.
[*] Giornalista, dottore di ricerca in filosofia e docente a contratto, email: r_demichelis@parisnanterre.fr.

access to information; (2) that the machines face the problem of the inscrutability of reference, in the same way that the linguist imagined by Willard Van Orman Quine cannot precisely determine what the term "gavagai" refers to in a situation of radical translation; (3) that there is no rule for the application of language, except for "language games", as Ludwig Wittgenstein's linguistics teaches us. The hope of achieving complete explicability and transparency of algorithms is undoubtedly in vain: we can only rely on partial and broad interpretations that will never fully explain the underlying rules.

Keywords: XAI, Explicability, Interpretability, Linguistics, Language games, Ethics.

1. Un souvenir d'école

Lequel était-ce ? Je ne sais plus. Ils étaient plusieurs certainement, ces enseignants qui me demandaient : « Qu'est-ce que tu ne comprends pas ? » Une phrase si souvent entendue qu'elle se noue à l'imaginaire de l'éducation nationale, dans un mélange de bienveillance et de terreur. Comment répondre à cette question d'apparence innocente ? Pour comprendre ce que je ne comprends pas, encore faut-il avoir un indice, une piste, subodorer que je sais vers où me diriger. Bref, encore faut-il avoir un peu compris pour savoir ce que l'on ne comprend pas. Mais la honte surgit de l'incapacité même à formuler son ignorance ; on se trouve plus bête que bête.

Aujourd'hui, nous sommes en proie à cet embarras lorsque nous cherchons à expliquer certains algorithmes d'intelligence artificielle (IA), car leurs raisonnements échappent à une formulation mathématique. Ce n'est pas entièrement un hasard à en croire cette définition de Bruno Bachimont : « L'IA veut traiter informatiquement (c'est la méthode) des problèmes qui nécessitent des connaissances non formalisables pour être résolus (c'est l'objet)[1]. » Si ces connaissances ne sont pas formalisables, vouloir en formaliser une explication – donc une connaissance –, c'est se lancer dans une tâche paradoxale. Ce sont d'ailleurs les systèmes les moins explicables, ceux d'apprentissage automatique fondés sur des inférences statistiques, et particulièrement les réseaux de neurones formels (apprentissage profond), qui se sont avérés les plus utiles à des fins aussi variées que la reconnaissance d'images, la traduction ou la génération de contenus. Ces algorithmes ne répondent plus, par définition, à une formulation en logique formelle de type mathématique, ce qui leur donne une plus grande malléabilité et donc capacité d'adaptation, mais ils échappent ainsi à notre compréhension.

Il est devenu courant de parler de *boîtes noires*, dans le sens où les résultats fournis ne s'expliquent pas facilement. Tout juste pouvons-nous nous appuyer sur des estimations grâce à d'autres méthodes statistiques, mais pas sur l'*explicabilité* qu'offrent les mathématiques. Nous considérons d'ailleurs qu'il convient plutôt de parler d'*interprétabilité* lorsque nous ne pouvons pas arriver à une absence d'ambiguïté.

[1] B. Bachimont, *Le Contrôle dans les systèmes à base de connaissance. Contribution à l'épistémologie de l'intelligence artificielle*, 2nde éd., Hermès, Paris 1994, p. 181.

Cependant, des esprits optimistes ne relâchent pas leurs efforts et espèrent encore parvenir à l'*explicabilité*, au point d'avoir fait émerger la nouvelle discipline de l'*Explainable Artificial Intelligence* (XAI).

Mais avant toute entreprise de sauvetage, la question primordiale devrait être de savoir pourquoi exactement nous n'arrivons pas à atteindre cette absence d'ambiguïté ? Bref, affrontons nos démons infantiles et demandons-nous : qu'est-ce que je ne comprends pas ?

Notre propos consiste à mettre en évidence que la réponse est certainement à chercher du côté de la linguistique, dans le sens où il n'y a pas de possibilité de dire avec le langage les règles qui nous permettent de nommer les choses et qu'il en va de même pour les systèmes d'IA fondés sur l'apprentissage profond. Nous nous situons du côté des pessimistes.

Pour étayer cette thèse, nous commencerons par détailler le fonctionnement des systèmes d'apprentissage profond afin de mieux cerner la problématique. Nous reviendrons ensuite sur l'état de l'art dans l'XAI. Ensuite, nous nous pencherons sur la difficulté d'identifier le critère de dénomination dans la tradition philosophique, principalement avec Wittgenstein. Puis, nous nous inspirerons de son propos pour développer notre propre thèse.

2. Quel est le problème ?

A la question « pourquoi le système d'IA a donné tel résultat plutôt que tel autre ? », il devient difficile de répondre depuis la révolution de l'apprentissage automatique par réseaux de neurones artificiels, dit IA « connexionniste ». Jusqu'aux années 2010, l'IA « symbolique » prédominait le champ de l'informatique[2] et le problème de l'explicabilité n'avait pas vraiment lieu d'être : les opérations étaient appliquées sur des symboles qui signifiaient quelque chose pour nous (rouge, chien, grand, etc.). Puis, quelques innovations[3] en vision par ordinateurs vinrent changer le paradigme et redonner du lustre aux réseaux de neurones artificiels. Elles furent principalement soutenues par les grandes masses de données nouvellement à disposition grâce au développement d'Internet. Des masses de données nécessaires à l'apprentissage machine, car les réseaux de neurones artificiels ont besoin d'importantes quantités d'informations pour s'entraîner et induire automatiquement le meilleur paramétrage afin de répondre à un problème.

Seulement, cela signifie que le système reste fortement influencé par ses données d'apprentissage sans que nous puissions identifier dans l'algorithme où ni comment cette influence s'exerce. C'est ainsi qu'un système de catégorisation des images peut se mettre à assimiler par erreur des paysages enneigés à des photos de

[2] M. Mitchell, *Artificial Intelligence: A Guide for Thinking Humans*, Farrar Straus & Giroux, New York 2019, emplact. 366.

[3] A. Krizhevsky, I. Sutskever, G. E. Hinton, *ImageNet classification with deep convolutional neural networks*, in « Communications of the ACM », 60, n. 6, 2012, pp. 84-90 ; Q. V. LE, M.A. RANZATO, R. MONGA, et al., *Building high-level features using large scale unsupervised learning*, in *arXiv*, 12 juillet 2012.

loups[4] ; parce que toutes les occurrences de loups dans sa bases de données d'apprentissage montraient les canidés dans la neige. C'est encore de cette manière qu'un système peut se mettre à adjoindre systématiquement un bras humain à un haltère quand on lui demande simplement « un haltère[5] » ; car les haltères étaient tous tenus à bout de bras dans les photos d'entraînement. Nous disons ainsi souvent qu'il y a un *biais* dans ce genre de situation. Le problème de fond est que peuvent se créer ainsi des associations d'idées ou des anticipations automatisées entre des personnes et des faits ou des comportements. Il a ainsi été observé avec des systèmes d'IA une multitude de biais discriminatoires, reproduisant des préjugés sexistes, racistes, islamophobes ou validistes[6].

Identifier ces écueils, les isoler et les déjouer est devenu un enjeu de premier ordre dans nos démocraties. Le règlement européen dit « AI Act », dont la publication a eu lieu en 2024, prévoit ainsi que les systèmes dits « à haut risque », utilisés notamment pour les ressources humaines ou l'attribution de crédits, soient audités afin de « de repérer d'éventuels biais qui sont susceptibles de porter atteinte à la santé et à la sécurité des personnes, d'avoir une incidence négative sur les droits fondamentaux ou de se traduire par une discrimination interdite par le droit de l'Union, en particulier lorsque les données de sortie influencent les entrées pour les opérations futures[7] ». Il conviendra donc pour les éditeurs de prendre « des mesures appropriées visant à détecter, prévenir et atténuer les éventuels biais repérés[8] ».

La grande difficulté pour les informaticien·nes est de réussir à comprendre le fonctionnement des algorithmes d'apprentissage profond, car la valeur de chaque paramètre est rarement compréhensible. Dans les couches inférieures, les systèmes connexionnistes ne manipulent pas des symboles, comme un code couleur tel que #FF0000 pour le rouge, mais des valeurs numériques issues de ce code transformées au cours de multiples opérations. Ces calculs sont parfois très simples, mais comme ils sont effectués de façon croisée, la référence est diluée au fil des calculs : au neurone 456 il est impossible de savoir ce que veut dire le chiffre 0,42. Cela ne signifie rien et c'est à l'image de l'ordinateur surpuissant, dans le film *H2G2 : le guide du voyageur galactique*, qui répond à « la question de la vie, de l'univers et de tout le reste[9] » par

[4] M. Ribeiro, S. Singh et C. Guestrin, *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*, in *arXiv*, 9 août 2016, pp. 9-10.

[5] A. Mordvintsev, C. Olah, M. Tyka, *Inceptionism: Going Deeper into Neural Networks*, « ai.googleblog.com », 18 juin 2015, consulté le 7 novembre 2022.

[6] M. Broussard, *More than a Glitch: Confronting Race, Gender, and Ability Bias in Tech*, The MIT Press, Cambridge MA 2023 ; J. Buolamwini, *Unmasking AI: My Mission to Protect What Is Human in a World of Machines*, Random House, New York 2023 ; R. Demichelis, *L'Intelligence artificielle, ses biais et les nôtre. Pourquoi la machine réveille nos démons*, Faubourg, Paris 2024.

[7] Union Européenne, *Règlement (UE) 2024/1689 du Parlement européen et du Conseil du 13 juin 2024 établissant des règles harmonisées concernant l'intelligence artificielle*, 2024, art. 10-2, f.

[8] ivi, art. 10-2, g.

[9] G. Jennings, *H2G2 : le guide du voyageur galactique*, Spyglass Entertainment, Touchstone Pictures, Hammer & Tongs, 2005, 40:35:00. Exemple mentionné dans R. Demichelis, *L'IA, ses biais et les nôtre*, cit., pp. 12-13.

« 42 ». Cela n'a pas de sens et nous laisse dans le plus grand embarras ; nous ne savons que faire de cette information. Le problème est d'autant plus grand que, dès la phase de vectorisation (lorsque le symbole est transformé en valeurs numériques), les symboles de notre langage naturel disparaissent souvent : en traitement du langage, le logiciel considère rarement un mot isolément lorsqu'il l'intègre dans ses calculs, mais compris dans un groupe de plusieurs mots, selon un certain contexte. Dès lors, les vecteurs sur lesquels sont appliqués des opérations ne représentent qu'un ensemble de mots ou de lettres qui perdent leur signification pour nous.

3. Les solutions et leurs limites

S'il est difficile d'identifier exactement quels paramètres, quels vecteurs ou quelles données ont été déterminantes dans la décision d'un système d'IA, cela n'a pas empêché de nombreux chercheur·es d'essayer de relever le défi. L'explicabilité des algorithmes est ainsi devenue une branche à part entière de la recherche en informatique connue sous l'anglicisme d'Explainable AI ou l'acronyme XAI[10]. L'objectif est donc d'arriver à une *explicabilité*, mais nous soutenons toutefois, comme des professionnels nous l'ont partagé lors de nos enquêtes de terrain, qu'il s'agit davantage d'*interprétabilité*, car il est impossible d'arriver à une absence d'ambiguïté mathématique (ce que désigne le terme d'explicabilité) avec des systèmes statistiques opaques.

Deux approches sont souvent évoquées dans la littérature pour parler des différents types d'interprétation : l'approche *locale* et l'approche *globale*[11]. L'approche locale consiste à identifier les déterminants pour tel individu : pourquoi telle personne a vu sa demande de crédit refusée ? Pourquoi tel chien a été catégorisé correctement comme un chien ? Etc. L'approche globale, cherche à détailler comment le modèle fonctionne en général, quels que soient les individus : quels sont les facteurs déterminants dans la classification des animaux ? Quelles sont les informations cruciales pour obtenir un crédit ? Etc. L'une et l'autre méthode ne sont pas exclusives et peuvent tout à fait être sollicitées de façon complémentaire. L'approche locale aurait toutefois plutôt tendance à s'appliquer à des cas d'usage problématiques, où l'on cherche à savoir pourquoi un individu a essuyé un refus ou a été incorrectement catégorisé.

Il existe pléthore de solutions techniques pour interpréter un algorithme, mais deux d'entre elles sont régulièrement évoquées : LIME[12] et SHAP[13]. Elles visent toutes les deux à dépasser l'écueil de ne se reposer que sur des tests standardisés pour

[10] W. Saeed, C. Omlin, *Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities*, in « Knowledge-Based Systems », 263, 5 mars 2023, art. 110273.
[11] J.-M. John-Mathews, *Interprétabilité en Machine Learning, revue de littérature et perspectives*, Telecom Paris, HAL, 2019.
[12] M. Ribeiro, S. Singh, C. Guestrin, *"Why Should I Trust You?"*, cit.
[13] S. M. Lundberg, S.-I. Lee, *A Unified Approach to Interpreting Model Predictions*, in « Advances in Neural Information Processing Systems », 30, 2017.

mesurer la pertinence [*accuracy*] des modèles. Car un algorithme peut s'avérer très pertinent en conditions de laboratoire (tests standardisés), mais pas dans la nature (face à des exemples ou des combinaisons jamais rencontrés et plus proches de la réalité).

LIME a vocation à créer des interprétations locales de tout type d'algorithme. Sa méthode consiste à mesurer le poids de certaines informations en entrée : quels mots ont permis de dire que ce texte parlait de christianisme ou d'athéisme ? Quels pixels ont permis de dire qu'il s'agissait d'un husky plutôt que d'un loup ? LIME ne propose pas d'interprétation globale, mais plusieurs interprétations locales portant sur divers résultats afin d'offrir une meilleure compréhension du modèle.

SHAP, de son côté, vise à estimer le poids de chaque information en entrée (comme LIME), cependant non plus de façon binaire (selon leur présence ou leur absence), mais selon des valeurs (de Shapley). L'enjeu est de mesurer la résistance à la variabilité du contexte et d'assurer la cohérence de l'interprétation, notamment si une information déterminante voit son poids augmenter ou rester identique [*consistency*].

Il est intéressant de remarquer que les articles sur LIME et SHAP font tous les deux appels à des humains pour évaluer l'intelligibilité des interprétations. « Les explications devraient être faciles à comprendre » et visent « une compréhension qualitative », est-il écrit dans le premier tandis que le second mentionne « l'intuition humaine » pour valider sa méthodologie. La compréhension relève d'un aspect qualitatif qui n'est pas formulable et qui requière un échange avec autrui pour savoir s'il a compris. Seul son ressenti peut faire foi.

C'est la raison pour laquelle il est souvent suggéré d'impliquer les utilisateurs et les personnes visées par une technologie aussi bien dans l'audit[14] que dans la création des modèles[15]. Ils est elles sont parfois les mieux placé·es pour savoir quels sont les écueils, les incohérences, affiner l'interprétation ou souligner un manque d'interprétabilité de l'outil, toujours selon certaines situations définies. Même si un modèle est explicable, il ne le sera peut-être pas pour une communauté particulière : on ne s'adresse pas à une scientifique comme à une littéraire. Il convient ainsi de mettre en place une approche herméneutique de l'interprétation. Derrière ce pléonasme se cache une référence à l'herméneutique en tant que critique sociale : c'est-à-dire une interprétation d'après les normes et connaissances[16] propres à une communauté et au plus près d'elle[17], non pas d'un point de vue externe qui ferait descendre l'explication de façon verticale comme un juge impérial.

Olya Kudina parle ainsi d'« analyse interprétative phénoménologique »

[14] M. Broussard, *More than a Glitch*, cit., p. 163.

[15] T. Achiume, UN. Human Rights Council, « Racial discrimination and emerging digital technologies: a human rights analysis », Organisation des Nations Unies, 2020.

[16] Union Européenne, *Règlement (UE) 2024/1689*, cit., art. 5, c ; M. Ribeiro, S. Singh, C. Guestrin, « "Why Should I Trust You?" », cit.

[17] M. Graziani, L. Dutkiewicz, D. Calvaresi, et al., *A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences*, in « Artificial Intelligence Review », 56, 2022, pp. 3473-3504.

[*Interpretative Phenomenological Analysis*][18], en tant que cette méthode se concentre sur des « micro-perspectives situées et des principes philosophiques d'interprétation circulaire[19] » ; la technologie réveille des valeurs propres à une société, mais les influence aussi, provoquant leur réactualisation dans un mouvement de « dynamisme de la valeur[20] » [*value dynamism*], ou dialectique. Il convient ainsi de mettre en place un travail itératif capable de prendre en compte les retours utilisateurs pour obtenir une meilleure compréhension du modèle.

Fabio Paglieri propose, quant à lui, de « suivre l'argent » pour mieux comprendre comment les outils sont orientés[21] :

> Les explications recherchées par l'XAI […] concernent toujours le fonctionnement interne des systèmes d'IA, « comment la magie opère », ce qui est exactement la raison pour laquelle cet enjeu est insaisissable pour l'IA générative et [l'apprentissage automatique]. Il est quelque peu surprenant cependant, que peu d'attention – voire aucune – ne soit portée sur les autres types d'explication, focalisés non sur comment ces système marchent, mais plutôt sur qui en tire profit (ou y perd) du fait qu'ils fonctionnent. « *Cui prodest* ? »

Il y a dans cette idée une certaine herméneutique du soupçon[22], c'est-à-dire une interprétation qui n'hésite pas à spéculer sur les raisons peu avouables et parfois inconscientes d'un locuteur. L'herméneutique du soupçon trouve son incarnation dans la psychanalyse ou le marxisme en tant qu'il y aurait souvent plus à lire dans les propos d'un patient ou d'un adversaire politique que ce qui est explicitement exprimé. Il y a donc l'idée que quelque chose est caché – intentionnellement ou non – qu'il faudrait faire apparaître. C'est un art délicat qui peut basculer dans une certaine folie, dans des surinterprétations excentriques, des pathologies de l'interprétation. Dans la sphère politique, économique et technologique cependant, l'herméneutique du soupçon est aussi une manière de ne pas prendre pour argent comptant les propos des grandes entreprises du numérique et de déjouer des stratégies d'enfumage. L'interprétation devient démystificatrice. « Suivre l'argent » permet de mettre en évidence les raisons qui poussent à la production des outils et comprendre pourquoi ils sont paramétrés d'une telle manière plutôt que d'une autre. Quand il y a des coupes budgétaires pour les expérimentations, l'éthique ou la conformité, nous comprenons mieux pourquoi certains systèmes enfreignent les lois ou la morale. Quand il n'y a pour seul objectif que la rétention des internautes sur le site web, afin de satisfaire les annonceurs et un modèle économique fondé sur la publicité, nous comprenons mieux pourquoi certains contenus sont mis en avant plutôt que d'autres. Les exemples peuvent ainsi se multiplier.

Il est intéressant de remarquer que toutes ces méthodes d'interprétabilité

---

[18] O. Kudina, *Moral Hermeneutics and Technology: Making Moral Sense through Human-Technology-World Relations*, The Rowman & Littlefield Publishing Group, Lanham MD 2023, 182 p.
[19] Ivi, p. 12.
[20] Ivi, p. 3.
[21] F. Paglieri, *Expropriated Minds: On Some Practical Problems of Generative AI, Beyond Our Cognitive Illusions*, in « Philosophy & Technology », 37, n. 2, 2024, p. 55.
[22] J. Michel, *Homo interpretans*, Hermann, Paris 2017, p. 158.

tiennent quasiment pour acquis que le modèle original ne pourra pas être parfaitement expliqué. Comme écrivent Marco Ribeiro et al. (article sur LIME) : « Il est souvent impossible pour une explication d'être complètement fiable [*faithful*] à moins qu'elle soit la description complète du modèle lui-même[23]. » Les logiciels et différentes méthodes évoquées dans cette partie n'ont donc pas véritablement pour ambition l'*explicabilité*, mais l'*interprétabilité*. Il demeurera toujours une incertitude sur les relations de cause à effet dans le modèle original. L'abandon de l'ambition d'explicabilité, c'est aussi le deuil de la transparence totale. Nous ne pourrons pas lire dans un réseaux de neurones artificiels comme dans un livre ouvert (et même le livre nous demande d'interpréter). Seules des méthodes d'*interprétabilité* seront de ce point de vue *satisfaisantes*, mais elles seront *insatisfaisantes* si l'objectif est la transparence diaphane de l'*explicabilité*.

4. Aspects philosophiques : le nom, la carte et la vitre

La difficulté d'identifier clairement ce qui dans une image définit un chat, un loup ou tout autre objet est déjà exprimée dans l'Antiquité d'une autre manière. Pour moquer Platon, qui avait défini l'homme comme un « bipède sans plumes[24] », Diogène de Sinope, dit le Cynique, lui apporta un coq plumé et s'exclama : « Voilà l'homme de Platon ! » Cette anecdote traduit la difficulté de toute définition qui risque toujours d'omettre certains aspects d'un objet. Que dire aussi des accidents ? Un homme sans jambes n'est plus un bipède et n'est pourtant pas moins homme. Dès lors énumérer des critères pour catégoriser des objets risque de négliger certains aspects, d'oublier quelques exceptions.

Cette idée que le modèle d'explication (la définition) *n'est pas* le modèle original (l'objet) se retrouve déjà dans la littérature scientifique lorsqu'il s'agit de différencier la carte et le territoire. « La carte n'est pas le territoire[25] », disait Alfred Korzybski en 1931. La légende raconte que ce mot lui a été inspiré d'une triste expérience : durant la Première Guerre Mondiale, des soldats sous son commandement auraient été abattus par une mitrailleuse prussienne qui n'était pas mentionnée sur la carte[26].

Mais plus que cela : le modèle d'explication *ne doit pas* être le modèle original. Jorge Luis Borges écrivit une nouvelle absurde en 1946, *De la rigueur de la science*[27], dans laquelle des géographes créent une carte qui représente exactement tout le territoire d'un Empire, c'est-à-dire à échelle 1:1. Seulement, les générations suivantes la jugent bien évidemment « inutile » [*Inútil*]. Cela signifie que représenter l'Empire de façon

---

[23] M. Ribeiro, S. Singh, C. Guestrin, *"Why Should I Trust You?"*, cit.
[24] D. Laërce, *Vies et doctrines des philosophes illustres*, tr. fr. de T. Dorandis, Le Livre de Poche, Paris 1999, p. 718.
[25] A. Korzybski, *A Non-Aristotelian System and its Necessity for Rigour in Mathematics and Physics* (1931), in *Semanticscholar.org*, 2004, p. 750.
[26] H. Krivine, *Comprendre sans prévoir, prévoir sans comprendre*, Cassini, Paris 2018, p. 98.
[27] J. L. Borges, *Histoire universelle de l'infamie / Histoire de l'éternité* (1935), tr. fr. de R. Caillois L. Guille, 10/18, Paris 1994, p. 107.

symétrique « point par point » ne sert à rien et que l'utilité d'une carte est justement de synthétiser l'information, dans le sens de résumer. Il faut que la carte contienne moins d'informations que le territoire qu'elle représente pour servir à quoi que ce soit. Avec l'IA, il faut que le modèle d'explication contienne moins d'informations que le modèle original, et cela même si le modèle original est explicable, c'est-à-dire sans ambiguïté. Comme l'écrivent Ribeiro et al.[28] :

> Si des centaines ou des milliers de caractéristiques contribuent significativement à la prédiction, il n'est pas raisonnable d'attendre de la part de n'importe quel utilisateur de comprendre pourquoi la prédiction est faite, même si chaque poids peut être inspecté.

La métaphore de l'inutilité de la *carte parfaite* traduit aussi l'idée qu'il n'est pas de *terme parfait* pour correspondre à la chose à laquelle il fait référence. Si nous pensons à un mot comme à une carte, alors il ne peut pas la recouvrir parfaitement. La carte comme les concepts impliquent une médiation, une traduction et presque une trahison de la chose visée. Bref, une interprétation et cela va dans le sens de ce que nous disions sur la vanité de toute entreprise d'explicabilité et sur le deuil nécessaire de la transparence totale. Emmanuel Alloa remarque qu'une « vitre qui est vraiment transparente finit par nier sa propre existence matérielle[29] » ; il n'y a de transparence que parce qu'il y a d'abord un obstacle. Il ajoute : « La promesse d'une circulation libre et informée […] finit par confiner le mouvement à un schéma méticuleusement prédéfini. »

Appliquons ce propos à l'IA : assigner à un réseau de neurones artificiels un chemin préétabli reviendrait à limiter ses potentialités. L'apprentissage profond profite de sa souplesse pour s'adapter à des situations variées pour lesquelles les règles rigides sont insuffisantes. S'il fallait imposer des règles rigides et les figer dans l'algorithme en pensant le rendre ainsi explicable, ce serait lui ôter sa capacité d'adaptation et son utilité initiale ; une *bonne vieille IA* pourrait faire aussi bien, donc les réseaux de neurones artificiels deviendraient caducs (même les IA dites « hybrides » [*neuro-symbolic*][30] ou « raisonnantes[31] » conservent en grande partie la souplesse que leur accorde l'inférence statistique).

Diogène, Borges, Korzybski ou Alloa nous montrent que le concept, la carte ou la modélisation, n'atteignent jamais l'objectif de correspondance parfaite à la chose censée être représentée. Nous saisissons qu'il est dans la nature même du medium

---

[28] M. Ribeiro, S. Singh et C. Guestrin, « "Why Should I Trust You?" », cit.

[29] E. Alloa, *Seeing Through a Glass, Darkly. The Transparency Paradox*, in E. Alloa (sous la dir. de), *This Obscure Thing Called Transparency: Politics and Aesthetics of a Contemporary Metaphor*, Leuven University Press, Leuven 2022, pp. 9-25.

[30] T. H. Trinh, Y. Wu, Q. V. Le, H. He, T. Luong, *Solving olympiad geometry without human demonstrations*, in « Nature », 625, n. 7995, 2024, pp. 476-482 ; B. Braunschweig, *IA : de l'intérêt d'un système hybride*, in « Les Echos », n. 24167, 11 mars 2024.

[31] T. Brown, B. Mann, N. Ryder, et al., *Language Models are Few-Shot Learners*, in *arXiv*. 2020 ; A. Leveau-Vallier, *Que comprend-on de ce que « comprend » ChatGPT ?*, in « Multitudes », 96, n. 3, 2024, pp. 160-166.

utilisé de receler une dimension pratique, portative parfois, afin de trouver une applicabilité. En IA, cette application recherchée est la compréhension, mais elle ne saurait donc passer par une tentative de description « point par point ». Dès lors, l'IA apporte sa contribution – avec quelques siècles de retards – à la querelle médiévale des universaux[32] : d'un côté, les réalistes croyaient en un isomorphisme entre le monde, les choses et les mots, de l'autre les nominalistes ne voyaient dans nos catégories que des objets sémantiques et non des essences réelles. L'IA connexionniste semble plaider pour ces derniers et ajoute une certaine dose de scepticisme. Il n'y aura pas de modèle symbolique définitif des essences, il n'y aura que des estimations grossières (sans que cela soit péjoratif) qui restent intimement dépendantes des modalités d'accès. Cependant, est-ce la seule limite à notre compréhension des modèles informatiques ? Nous avons jusqu'à présent abordé la question de l'interprétabilité sous l'angle de ce qui est décrit activement. Mais autre chose se joue dans notre connaissance. Il s'agit maintenant d'aborder la question à travers le processus d'apprentissage de façon située.

Car, un concept, contrairement à une carte, s'applique rarement à un seul cas particulier, à un seul territoire, mais à plusieurs dont les spécificités nourrissent la signification par rétroaction. La signification des mots dépend ainsi intimement des usages, selon différents contextes, différentes situations. Autrement dit, toute application enrichit la signification et donne donc lieu à un apprentissage généralisant à partir de configurations situées.

5. Le problème linguistique profond

La difficulté à laquelle les ingénieur·es sont confronté·es pour comprendre ce que désigne véritablement la machine est similaire à celle rencontrée par le « linguiste » décrit par Willard Van Orman Quine[33]. Le philosophe imagine une situation de « traduction radicale », c'est-à-dire une situation dans laquelle deux personnes parlant chacune une langue différente se rencontrent et essayent d'échanger verbalement pour la première fois. Avec un ethnocentrisme propre à son époque, Quine envisage ainsi un dialogue entre un « indigène » [*native*], d'une contrée inconnue, et ledit linguiste, qui se trouve évidemment être locuteur anglophone. Si l'indigène montre du doigt un lapin et prononce le terme de « *gavagai* », que doit en tirer comme conclusion le linguiste ? *Gavagai* peut aussi bien désigner le lapin, ses oreilles, ses moustaches, sa tête entière ou encore l'animal dans telle position. Il y a ainsi une « inscrutabilité de la référence[34] », comme l'histoire de la philosophie appelle ce problème, ou une « indétermination de la traduction[35] », comme l'écrit Quine. Sans pouvoir s'appuyer

[32] A. Conti, *Realism*, in R. Pasnau (sous la dir. de), *The Cambridge History of Medieval Philosophy*, 2 vol., Cambridge University Press, Cambridge 2009, vol. 2, pp. 647-660.
[33] W. V. O. Quine, *Word and object*, Technology Press of the Massachusetts Institute of Technology, Cambridge MA 1960, chap. 2.
[34] Ivi, p. 53. Juste [*inscrutability*].
[35] W. V. O. Quine, *Relativité de l'ontologie et quelques autres essais* (1977), Aubier, Paris 2008, p. 48.

sur d'autres concepts, le linguiste est dans l'embarras. Il doit passer par une inférence à la meilleure interprétation possible, selon le contexte, comme un logiciel d'apprentissage profond. Nous ne devrions ainsi pas être surpris de découvrir que le terme de *loup* désigne la neige pour un système de vision par ordinateur : il se retrouve face à l'inscrutabilité de la référence et tente de la dépasser comme il peut. Sans autres éléments pour lui indiquer le contraire, il peut continuer à estimer pendant longtemps que la neige sur l'image s'appelle *loup*.

L'inscrutabilité de la référence avait déjà été mise en évidence par Ludwig Wittgenstein quelques années auparavant. Lui ne disait pas *gavagai*, mais « *tove*[36] » (ce mot n'existe pas plus que celui de Quine) pour désigner soit *un crayon*, *rond*, *bois*, *un*, *dur* ou encore autre chose. Et il écrit que « c'est le travail de la définition ostensive de donner une signification[37] ».

Wittgenstein a ainsi défini la signification [*meaning*] d'un mot comme le fruit de « jeux de langage » [*language games*][38]. C'est-à-dire qu'elle se constitue au cours de situations durant lesquelles est fait usage d'un terme de façon ostensive [*ostensive*] ; lorsque quelqu'un pointe quelque chose qu'il désigne selon un certain contexte. Wittgenstein prend en exemple un ouvrier qui montrerait à un autre des matériaux ou des outils parmi d'autres objets. La signification des mots apparaît ainsi au fil de leurs usages. « Les jeux de langage sont les formes de langage avec lesquelles les enfants commencent à faire usage des mots. L'étude de jeux de langage est l'étude de formes primitives de langage ou de langages primitifs[39]. » Selon nous, les systèmes d'IA d'apprentissage profond, particulièrement – mais pas seulement – de vision par ordinateur, se retrouvent dans des situations similaires lors de leur entraînement. Ce qu'il faut comprendre en creux, et ce que Wittgenstein explique très bien, est qu'il n'est pas de règle pour l'application du langage en tant que règles ; *il n'est pas de règle de l'application de la règle*. Dès lors, chercher à savoir pourquoi une système d'IA appelle tel objet « chat » et tel autre « chien » semble voué à l'échec. Nous pouvons énumérer les caractéristiques de ces espèces, mais la façon dont nous appliquons les concept, ou la façon dont la machine le fait, relève d'un usage plutôt que d'une règle. Les usages sont variés et dépendent des contextes comme autant de « jeux » dans lesquels le langage est utilisé.

L'humain a le défaut, selon Wittgenstein, de mépriser le particulier au profit du général, mais c'est cette « soif de généralité[40] » qui nous fait perdre de vue comment se constitue la signification ; au lieu de regarder des exemples particuliers, car considérés comme « incomplets[41] », nous poursuivons la généralité d'une règle qui

---

[36] L. Wittgenstein, *The Blue and Brown Books* (1958), Harper Perennial, New York London Toronto Sydney 1965, p. 2.
[37] *Ibidem.*
[38] Ivi, p. 17 ; L. Wittgenstein, *Philosophical Investigations* (1953), tr. eng. de G. E. M. Anscombe, P.M.S. Hacker, J. Schulte, 4e éd., Wiley-Blackwell, Chichester 2009, paragr. 2 et 21 notamment.
[39] L. Wittgenstein, *The Blue and Brown Books*, cit., p. 17.
[40] Ivi, p. 18.
[41] Ivi, p. 19

n'existe pas ou qui sera toujours déceptive, insuffisante, et finalement incomplète. Surtout, elle entraînera une régression à l'infini, car une fois établie la règle, encore lui faut-il une « interprétation[42] » pour procéder à son application – interprétation qui ne peut être constituée que d'autres règles, etc. Des mots, « nous ne pouvons pas établir de règles strictes [*tabulate*] pour leur utilisation[43] ».

Nous demandons aux systèmes d'IA d'user du langage en s'émancipant des règles. C'est la raison pour laquelle l'apprentissage profond a été utilisé, parce que l'énumération des règles était impraticable pour ne pas dire impossible. Les informaticien·nes ont certainement eu recours à l'inférence statistique par sens pratique, parce qu'ils et elles se heurtaient à un plafond de verre avec l'IA symbolique (c'est-à-dire qu'ils et elles ne comprenaient pas forcément la difficulté à laquelle ils et elles avaient affaire), mais ils et elles ont ainsi illustré le propos de Wittgenstein et lui ont, d'une certaine manière, donné raison par la même occasion : *notre usage de la langue est une boîte noire*.

Les progrès de la technologie ont été rendus possibles par l'approche statistique et ostensive de la référence, qu'il s'agisse d'apprentissage supervisé, par renforcement ou non-supervisé. Si nous montrons une succession d'objets à une machine et que nous leur attribuons des étiquettes (apprentissage supervisé), nous sommes dans une approche ostensive par excellence. Si nous corrigeons la machine selon les réponses qu'elle nous fournit (apprentissage par renforcement), nous l'orientons implicitement vers l'étiquette que nous attendons. Si nous la laissons chercher par elle-même des catégories dans des données qui sont structurées de telle sorte que leur différenciation fasse sens pour nous (apprentissage non supervisé), alors nous orientons encore la machine implicitement, mais sans agir directement sur elle. Nous lui montrons ce que nous voulons signifier quel que soit le type d'apprentissage. Si la structure de ses réponses s'éloignait d'ailleurs trop de nos significations, la machine serait disqualifiée sur le champ comme aléatoire et/ou tenant des propos inintelligibles (à noter que les *hallucinations* des IA génératives signifient malgré tout quelque chose).

Dès lors, nous pouvons chercher à estimer le poids de chaque pixel ou de chaque mot dans le paramétrage de la machine, mais cela ne nous permettra pas de comprendre exactement la manière dont elle fait usage du mot. L'exactitude n'est pas de ce monde et c'est justement ce qui rend notre langage naturel si pratique pour exprimer ce que nous voulons signifier. C'est aussi ce manque de formalisme, cette porte ouverte à l'ambiguïté, qui offre aux réseaux de neurones artificiels des capacités jamais atteintes jusqu'alors.

Toutefois, chercher à comprendre avec des règles ce qui n'a pas lieu de l'être avec des règles n'a pas de sens. Les règles que nous recherchons ne sont pas des règles d'application ni de compréhension, mais du langage lui-même : *les modèles d'explication ne sont finalement que des modèles de langage. Il n'y a pas de règle de la règle.* Si nous voulons maintenant comprendre pourquoi le neurone *x* donne le résultat *y* et quel est son

[42] Ivi, p. 33.
[43] Ivi, p. 28.

influence sur la décision z, cette question n'a pas plus de sens. Si nous voulons savoir pourquoi le système fournit telle ou telle réponse, le secret est simplement que c'est ce que nous lui avons demandé de faire en lui montrant des exemples qui ne s'expriment pas en règles. A aucun moment nous ne lui avons demandé de choisir la route la plus élégante possible pour y arriver, mais au contraire la plus efficace sur le plan statistique. Pourquoi s'étonner alors du manque d'explicabilité ?

Il y a en fait trois problèmes dans celui de l'interprétabilité : (1) celui du poids des informations d'entrée, et cela exclut d'office toute explication exempte d'ambiguïté, puis (2) celui du fonctionnement du système et (3) celui de l'application. Nous pouvons estimer le poids de paramètres sans néanmoins parvenir à une explication complète ; ce n'est jamais l'objet de l'interprétabilité. Il y aura peut-être éclaircissement sur les raisons, même économiques ou sociétales, qui se lisent dans telle ou telle décision, mais il restera toujours une part d'obscurité. Nous pouvons à l'inverse avoir une vue complète des calculs sans néanmoins parvenir à une compréhension de ces opérations. Dans tous les cas, les interprétations ne nous éclairerons pas sur la règle de l'application qui disparaît et existe avec l'usage.

« La signification d'une phrase pour nous est caractérisée par l'usage que nous en faisons[44] », écrit Wittgenstein. Si nous voulons en savoir plus sur une phrase, nous pouvons soit étudier ses cas d'usage, lors de processus définitionnels ostensifs, mais cela évacue la possibilité de trouver la règle de l'application. Parcourir la base de données d'apprentissage pour en proposer une analyse qualitative se révèlera utile sans jamais être suffisant. Ou bien, nous pouvons analyser comment la phrase s'insère dans la langue, et c'est en fait cette recherche que proposent les systèmes d'interprétabilité. Autrement dit, si nous recherchons l'explicabilité, nous serons toujours déçus. L'interprétation est un pis-aller, mais cela ne veut pas dire qu'elle est inutile.

6. Conclusion

L'IA est aujourd'hui confrontée à un problème pour lequel des réponses pratiques deviennent nécessaires. Ce problème, c'est celui des algorithmes boîte noire ; il est devenu légalement contraignant de chercher à expliquer certains, ceux à « haut risque », avant leur déploiement. Nous avons vu qu'à cette fin plusieurs solutions existent, techniques ou herméneutiques. Il s'agit de pondérer les paramètres, de façon quantifiable sur le plan statistique, ou de chercher à comprendre les raisons profondes des résultats, qu'elles s'inscrivent dans une culture ou dans des incitations politiques et économiques.

Seulement ces méthodes ne sont jamais purement explicatives. Elles n'offrent jamais l'absence d'ambiguïté exigée par les mathématiques. Le seul espoir que nous pouvons avoir pour comprendre un peu mieux les systèmes d'IA, et particulièrement les algorithmes connexionnistes, est de recourir une *interprétation*. Il convient donc de

[44] Ivi, p. 65.

parler d'*interprétabilité* en lieu et place d'XAI. Cependant, nous ne nous sommes pas arrêtés à ce constat et nous avons essayé de délimiter les conditions d'impossibilité de l'*explicabilité*.

Nous avons exploré en premier lieu l'idée que *les systèmes d'IA connexionnistes n'ont à aucun moment l'ambition d'être explicables* et qu'ils s'affranchissent même de cette contrainte pour pouvoir être plus performants. En puisant dans la tradition philosophique, nous avons ensuite tâché d'apporter une réponse déjà esquissée par les chercheur·es en informatique et que nous pouvons résumer sous l'idée que *la modélisation soi-disant explicative ne saurait jamais correspondre parfaitement au modèle d'origine*. Nous espérons avoir apporté avec la troisième réponse une approche plus originale. Elle consiste à dire que *le processus d'apprentissage d'une langue repose sur des usages dont les règles nous échappent*. Ce n'est pas une impossibilité de comprendre la langue elle-même, c'est une impossibilité de formuler son usage au-delà de la sphère de la langue – nous comprenons, mais les symboles manipulables ne nous aident pas à comprendre pourquoi nous comprenons. Car il y a, via les réseaux de neurones artificiels ou via nos esprits, une inscrutabilité résiduelle de la référence lorsque nous parlons d'un objet quel qu'il soit.

Cela n'implique en rien qu'il y ait identité entre nos cerveaux et les systèmes informatiques. Bien entendu l'IA se fonde sur l'espoir de reproduire nos facultés cognitives et peut donc présenter des similitudes, mais il n'y a aucune raison de franchir le pas de l'anthropomorphisme. Même avec une architecture plus ou moins analogue, il se peut très bien que les logiciels se développent d'une manière différente et produisent les mêmes résultats. Si les ingénieur·es semblent donner raison à Wittgenstein sur le fonctionnement de l'apprentissage de la langue, ils ne prouvent rien. Dire que la machine *comprend* serait même aller au-delà de notre propos.

Cette critique de l'ambition de transparence absolue ne doit toutefois pas nous faire oublier qu'il existe des enjeux de conformité et moraux à essayer quand même d'esquisser une interprétation des outils. Il serait inacceptable dans une société libérale que cet effort ne soit pas fourni lorsque nous savons que ces systèmes amènent à des discriminations préjudiciables contre certaines catégories de la population, des personnes qui sont déjà susceptibles de subir des injustices aujourd'hui. Insister sur les limites de l'interprétabilité ne doit pas faire basculer dans l'immobilisme ; les discussions hautement théoriques servent trop souvent à dissimuler ce qui n'est que de la mauvaise foi, de la paresse, voire de la complicité.

Pour mieux comprendre les algorithmes, il faut littéralement penser « *outside the box* ». Il faut sortir du cœur de la machine et accepter que ses résultats ont certainement plus à nous dire que ses mécanismes, mais que ni les uns ni les autres ne nous diront tout. Nous ne sommes pas les premiers à avoir suggéré un pas de côté dans la méthodologie de l'interprétabilité, les approches statistiques ou herméneutiques en sont déjà de bons exemples.

The Vice of Transparency
A Virtue Ethics Account of Trust in Technology[a]

Francesco Striano[*]

Abstract

Questo articolo esplora il rapporto tra fiducia, trasparenza e tecnologia da una prospettiva di etica delle virtù. Mette in discussione l'assunto che la trasparenza sia essenziale per la fiducia, distinguendo tra fiducia, affidamento e confidenza. La trasparenza viene poi esaminata sia come disponibilità informativa sia come processo sociale di negoziazione. L'articolo sostiene che la trasparenza nel primo senso può portare a un sovraccarico di informazioni e a dinamiche di controllo, sostenendo invece un rapporto equilibrato e virtuoso con la tecnologia che enfatizzi le capacità interpretative dell'utente. Propone che la fiducia nella tecnologia dipenda sia dagli atteggiamenti individuali sia dall'affidabilità degli oggetti. Infine, l'articolo critica il "culto della trasparenza" contemporaneo, proponendo che l'onestà, come virtù tecno-morale, sostituisca la trasparenza quale obiettivo progettuale. Le tecnologie oneste medierebbero e negozierebbero l'accesso degli utenti alle informazioni, promuovendo una fiducia autentica e sostenendo la fioritura umana.

Parole chiave: confidenza, affidabilità, trasparenza, fiducia, etica delle virtù

This article explores the relationship between trust, transparency, and technology from a virtue ethics perspective. It challenges the assumption that transparency is essential for trust, distinguishing between trust, reliance, and confidence. Transparency is then examined as both informational openness and a social process involving negotiation. The article argues that transparency in the first sense can lead to information overload and control dynamics, advocating instead for a balanced, virtuous relationship with technology that emphasizes user interpretative skills. It proposes that trust in technology depends both on individual attitudes and objectual

a Questo saggio è un prodotto delle ricerche e delle collaborazioni condotte all'interno del progetto PRIN 2022 "Social Transformations & the Crisis of Expertise" (2022JR8Z8P) finanziato dall'Unione Europa – Next Generation EU. Saggio ricevuto in data 31/05/2025 e pubblicato in data 22/01/2025.

* Assegnista di ricercar, Università di Torino, email: francesco.striano@unito.it.

reliability. Finally, the article critiques the contemporary "cult of transparency," proposing that honesty, as a techno-moral virtue, should replace transparency as the design goal. Honest technologies would mediate and negotiate user access to information, fostering authentic trust and supporting human flourishing.

Keywords: confidence, reliance, transparency, trust, virtue ethics

1. Introduction

"Transparency builds trust," claim company slogans or leadership development courses. A presumption of trust—or at the very least, a reliance, a distinction that will be revisited subsequently—is contingent upon the assumption of reciprocal transparency between the involved parties. However, this assertion requires further examination. In a sense, when one chooses to place trust in another individual, it does not necessarily entail that one has complete information regarding how that individual will behave in a given situation. Alternatively, trust is based on the character traits of the individual in question, or it is founded upon the reliability of a professional or tool due to a prior history of satisfactory performance, which provides the basis for the assumption that the same level of reliability will be maintained in the future. In this sense, even in the absence of complete information about the situation, the premises, and so forth, we nevertheless decide to place trust or reliance.

The question thus arises as to how this applies to our relationship with technology. The European Union has repeatedly emphasised the importance of transparency, particularly in relation to algorithms and artificial intelligence[1]. Transparency appears to be a fundamental requirement of a trust relationship with technology. An opaque technology does not seem to be one that can be trusted. However, it is questionable whether it is even possible to "trust" technology in the first place. Furthermore, the relationship between trust and transparency in our interaction with technology is unclear. Does a virtuous relationship with technology, which promotes human flourishing, a good life with technology, and the use of technology towards a good life, require transparency?

This article will commence by examining the aforementioned inquiries, with the subsequent objective of determining the extent to which a methodology derived from virtue ethics—which appears to be optimally suited to the analysis of traits such

[1] As early as 26 January 2022, the *European Declaration on Digital Rights and Principles for the Digital Decade* already asserted the necessity of ensuring transparency about the use of algorithms and artificial intelligence, and that people are empowered and informed when interacting with them." Consequently, the AI Act (Regulation - EU - 2024/1689) incorporates transparency (along with human agency and oversight; technical robustness and safety; privacy and data governance; diversity, non-discrimination and fairness; societal and environmental well-being and accountability) among the seven principles that must "ensure that AI is *trustworthy* and ethically sound" (italic mine). The two documents may be accessed at the following web addresses: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52022DC0028 and https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689 (last accessed: 26 August 2024).

as trust, reliability, and transparency—can prove beneficial in resolving the identified issues.

The initial section will be dedicated to the development of a non-standard theory of trust, reliance, and confidence and to an examination of the applicability of these concepts to the relationship between humans and technological artefacts.

The second section will elucidate the existence of disparate accounts of transparency and illustrate the inherent ambiguity in defining a technology as "transparent."

The third section will examine the conditions under which transparency is necessary for a human being to trust or rely on a technology.

The fourth section will review virtue-oriented approaches to the relationship between trust and transparency in relation to technologies.

In the conclusion, I will argue, following Shannon Vallor, that honesty is the techno-moral virtue to be cultivated in relation to trust issues in the techno-social sphere. Furthermore, I will argue the need to design technologies *that are themselves honest*.

2. Trust or Reliance: How do we relate to technology?

"Imagine a society in which there is no trust in doctors, teachers, or drivers," writes Mariarosaria Taddeo: "This would require that all the members of the society to spend a significant portion of time and resources controlling the way others perform their tasks, at the expenses of their own tasks."[2]

A technologically advanced society necessitates that all its constituent elements, whether human, environmental, mechanical, or informational, operate with the greatest possible autonomy. However, the willingness to confer agential autonomy[3] upon a given component hinges upon the existence of trust.

The question thus arises as to how this concept of trust is to be defined. Furthermore, it is pertinent to inquire whether it is indeed feasible to repose trust in artificial agents. This issue is the subject of considerable debate, and there is no consensus in the literature. If we adhere to Taddeo's original definition, trust can be conceptualised as a second-order property pertaining to primarily binary and purpose-oriented relationships. In order to achieve a desired outcome, a trustor decides to do so through an ability possessed by a trustee, who is perceived as a trustworthy agent and therefore the relationship between them will have the property of being beneficial to the trustor. "Such a property is a second-order property that affects the first-order relations occurring between agents and is called trust."[4]

---

[2] M. Taddeo, *Trusting Digital Technologies Correctly*, in «Minds and Machines», n. 27, 2017, pp. 565-568: 566.

[3] For a discussion on the possibility of artificial agents with proper agency, I would direct your attention to my *Can Artificial Agents Act? Conceptual constellation for a de-humanised theory of action*, in «S&F_scienzaefilosofia.it», n. 31, 2024, pp. 224-244.

[4] M. Taddeo, *Modelling Trust in Artificial Agents, A First Step Toward the Analysis of e-Trust*, in «Minds and Machines», n. 20, 2010, pp. 243-257: 249.

This definition has the advantage of being readily applicable to trust in artificial agents; however, it may be inadequate in other respects. Firstly, due to its binary nature, it appears to present challenges in extending this definition to encompass more complex relationships or forms of trust, such as that which may be placed in institutions or decision-making processes. Furthermore, the definition may be perceived as somewhat circular: trust is defined as a quality inherent to relationships between individuals who are perceived as trustworthy. However, this raises the question of what factors contribute to an individual being regarded as trustworthy. What is this property that seems to be of first-order, in that it is possessed by the trustee, but must be defined by a second-order property?

Moreover, trust seems to be a more vague and undefined state. It can be understood as an attitude held by some individuals towards others or institutions that are perceived as trustworthy. The concepts of trust and trustworthiness appear to persist as attitudes and traits even when there is no relationship based on immediate utility at stake. For example, one may trust their partner both in the sense that they believe the partner will not lie to them and in the sense that they believe the partner will be able to provide support when needed. In such a case, the trustworthiness of the partner will be described as such even when there is no need to establish whether the partner is lying or when they are not providing support. Similarly, if the government of a State is described as untrustworthy, it is because, even if it is not currently causing harm or disseminating false information, there is a possibility that it may do so in the future. This distrust therefore translates into a state of heightened awareness and scrutiny of its actions.

The trust based on immediate usefulness described by Taddeo can be more accurately defined as "reliance." Some literature suggests that reliance is a broader, more neutral, and more general concept than trust. It posits that trust is a special case of reliance[5].

In line with this conception are de Fine Licht and Brülde, who propose the following definitions for reliance and trust.

They define reliance "as a three-place relation, where one agent (A) relies on an agent or some other object (B) to do something, to maintain some state, or the like (C)."[6] Reliance "can be both voluntary and involuntary"[7], so it can be applied to humans as well as objects: I rely on my bicycle to get me to the city centre because I judge that using it increases my likelihood of arriving on time for my appointment, because my bicycle is able to cover that distance without damage, because it is

[5] Cf. S. Blackburn, *Trust, cooperation, and human psychology*, in Id., *Practical Tortoise Raising and other philosophical essays*, Oxford Academic, Oxford 2010, pp. 90-108.
[6] K. de Fine Licht and B. Brülde, *On Defining "Reliance" and "Trust": Purposes, Conditions of Adequacy, and New Definitions*, in «Philosophia», n. 49, 2021, pp. 1981-2001: 1989.
[7] *Ibid*., p. 1990.

specially designed to move nimbly in city traffic, and because it has no misplaced parts and is at my immediate disposal at this moment[8].

Trust, on the other hand, is a special form of agential reliance in which "a trusting agent attributes a certain kind of moral motivation to the trustee, namely a perceived normative responsibility to care about something."[9] The attribution of this moral value is what makes it possible for trust to be *betrayed*, whereas a failed reliance can at best result in *disappointment*[10]. This is why, according to Deley and Dubois, a technology cannot be trusted, but, at best, can be relied upon: "technologies cannot possess a good will,"[11] and therefore cannot genuinely *betray* us.

Similarly, Christopher Thompson asserts that, in the absence of artefacts exhibiting mental states, it is not possible to place trust in technologies. He posits that when we appear to repose trust in the artefacts themselves, we are in fact confusing trust with reliance[12]. Furthermore, if a form of trust in artefacts exists, it is the trust that we accord to their designers or developers when we use them[13]. Likewise, Deley and Dubois maintain that "reliability is a *mediator* of trust toward the makers of a technology and that relationships of reliance mediate our relationships of trust."[14]

In contrast to Taddeo's conceptualisation of trust as a form of property, these alternative perspectives view it as a relationship, although sometimes it seems that trust is a state of mind of the trustor, at other times a judgement on trustworthiness. In light of the aforementioned considerations, it seems pertinent to direct attention to an alternative definition that merits consideration. This definition, proposed by Shionoya, characterises trust as an "evaluative act of one individual directed toward another with regard to whether that other person is or is not trustworthy in the relevant circumstances."[15] The definition presented is based on the concept of trust between two individuals. However, Shionoya also discusses the notion of trust as a

[8] These conditions correspond to what de Fine Licht and Brülde call *probability condition*, *competence condition*, *motivation condition*, *opportunity condition*. We could add the *value condition*, i.e., that I assign a positive value to my goal of getting to the city centre to arrive on time for my appointment and this leads me to judge my act of reliance positively as well. Cf. *ibid.*

[9] *Ibid.*, p. 1991. In order to attempt to decouple the notions of intentionality and motivation on the one hand and responsibility on the other, I will refer once more to F. Striano, *Can Artificial Agents Act?*, cit., pp. 241-244, in which I attempted to extend the concept of responsibility (distinguishing it, however, from accountability) to artificial agents as well, and to broaden the concept of distributed responsibility.

[10] Cf. A. Baier, *Trust and Antitrust*, in «Ethics», vol. 96, n. 2, 1986, pp. 231-260: 235.

[11] T. Deley and E. Dubois, *Assessing Trust Versus Reliance for Technology Platforms by Systematic Literature Review*, in «Social Media + Society», April-June 2020, pp. 1-8: 2.

[12] Cf. C. Thompson, *Faire confiance aux artéfacts – Faire confiance à distance*, in M. Doueihi and J. Domenicucci, *La confiance à l'ère numérique*, Éditions rue d'Ulm, Paris 2018, pp. 97-111: 108.

[13] Cf. *ibid.*, pp. 105-107.

[14] T. Deley and E. Dubois, *Assessing Trust Versus Reliance for Technology Platforms by Systematic Literature Review*, cit., p. 2, italic mine.

[15] Y. Shionoya, *Trust as a Virtue*, in Y. Shionoya and K. Yagi, *Competition, Trust, and Cooperation. A Comparative Study*, Springer, Berlin-Heidelberg 2001, pp. 3-19: 10.

shared virtue, as well as its community-based aspects within the context of trust networks.

Furthermore, Shionoya posits that the evaluative act of "trust" is contingent upon a willingness to trust and thus upon a disposition to trust or *confide*. Shionoya employs the term "confidence" as a synonym for trust, whereas de Fine Licht and Brülde understand it as a synonym for reliance.[16] I propose, however, that it be used to denote precisely that disposition which, according to Shionoya, underlies an act of trust. The result of my re-interpretation of Shionoya non-standard definition could be represented by a scheme (fig. 1) that sees trust as a disposition of the trustor leading to trust, understood as an evaluative act on being trustworthy, the latter being understood as a characteristic disposition of the trustee.

The latter disposition of the trustee, however, must have been earned in the past, probably on the basis that the trustee proved to be *reliable*. In this sense, we can describe reliance through an inverse scheme (fig. 2) to the confidence-trust-trustworthiness scheme derived from Shionoya. We can posit that reliability is a characteristic disposition of an individual or object, which inspires in a subject an evaluative act of reliance, which in turn leads to the evaluative subject's disposition of confidence.

In this sense, even an artefact, given proof of reliability, could become the object of an evaluative act of trust, not because it is endowed with a will or the possibility of betrayal, but because it is invested with trustworthiness by the confiding subject.

The argument put forth is that a person may confide in another person as well as in a technology on the basis of an evaluative act of trust – eventually based on

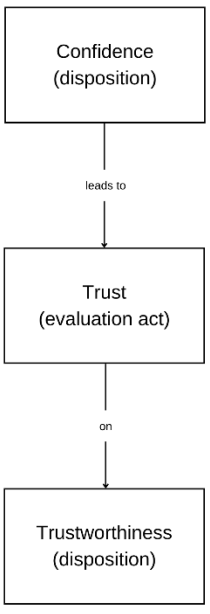

Fig.1

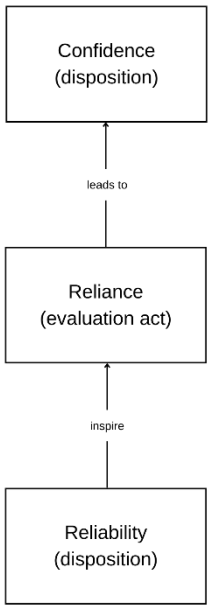

Fig. 2

[16] Cf. K. de Fine Licht and B. Brülde, *On Defining "Reliance" and "Trust"*, cit., p. 1981.

precedent or simultaneous evaluative acts of reliance. This evaluation may be based on certain conditions that must be fulfilled by the subject of the evaluative act, its object, or the act itself[17]. At this juncture, the research question of this paper can be rephrased as follows: do these conditions of confidence include the transparency of technologies?

3. What does transparency mean?

Prior to responding to this question, it is necessary to pose a preliminary inquiry: what is the precise meaning of the term "transparency?" If, as we have seen, the notion of trust is not entirely unambiguous, the notion of transparency also lends itself to ambiguity or different interpretations. Further difficulties arise when transparency is to be applied to the technological sphere. In order for transparency to be achieved with regard to technology, what, exactly, must be transparent in a technology?

Bridget O'Brien notes that there are at least two ways of framing the question of transparency, two different accounts that, as we shall see, can have different reflections on the question of trust[18].

The first account is that of transparency as informational openness. In this view, transparency is to be understood as comprehensive disclosure, as a free, continuous, and total flow of information, without restriction or censorship of any kind. In the context of medicine, for instance, this account entails that patients are granted comprehensive and unimpeded access to information pertaining to their care. This encompasses details about the treatments they are receiving, the diagnostic procedures scheduled, and the associated financial and social costs. In the context of the relationship between institutions and citizens, it is imperative that no information be withheld. In the case of a decision-making support algorithm the source code must be open and searchable, and each step must be explainable.

A second account is that which views transparency as a social process, specifically as a communicative act that is not without tensions and negotiations. This account challenges the assumption that access to comprehensive information is inherently beneficial. It posits that the mere availability of information does not guarantee its objectivity or completeness, and that individuals may not always possess the capacity to receive and interpret it accurately. This is because, according to this account, transparency is dependent on three interrelated elements: the *content*, the *receiver* ("the viewer," as defined by O'Brien) and the *medium*. In her analysis, O'Brien employs the metaphor of the window to illustrate the three components of transparency in relation to information. She posits that the scene outside the window, which represents the content, may fluctuate due to external factors beyond human control, such as weather, season, and the mobility of objects. The individual who

[17] These conditions may be consistent with those proposed by de Fine Licht and Brülde. Please refer to footnote 9 for further details.

[18] See B. C. O'Brien, *Do You See What I See? Reflections on the Relationship Between Transparency and Trust*, in «Academic Medicine», vol. 94, n. 6, 2019, pp. 757-759.

receives the information, on the other hand, can be conceptualised as the viewer, situated on the internal side of the window, influences the scene observed through the lens of their senses, prior knowledge, emotions, and interpretative abilities. The medium, in this case the window itself, shapes the scene in a particular manner, influencing the viewer's gaze and potentially distorting or deforming it.

The second account does not directly contradict the first. It may be the case that, under ideal conditions, a direct and total flow of information is possible. This will be discussed further in the next session, along with the question of whether it is also desirable to build a relationship of confidence with the technology. However, the social account appears to provide a more accurate representation of the intricacies associated with information transmission mechanisms. This approach facilitates comprehension of the fact that transparency can be conceptualised as a negotiation between these elements. The concept of transparency, in this context, refers to the optimal access to information, which may not be comprehensive but is tailored to the specific needs of the moment. This represents a goal, and the effort of negotiation should be directed towards its realisation. Furthermore, it seems desirable that the process of selecting the portion of information of interest to the receiver should also be transparent. Furthermore, the degree of deformation that a medium implements on the content should also be transparent and known. It is also essential that the receiver is equipped with the requisite knowledge and skills to interpret the information in an appropriate manner. The aforementioned measures must be implemented in order to achieve transparency, according to the social account.

It would appear that the focus of this paper should be on interventions on the medium. The question thus arises as to how one might intervene in order to make technologies (media) more transparent, thereby increasing user trust. Nevertheless, this raises further questions. Firstly, it is necessary to define what is meant by the term "transparent technology." It is necessary to determine whether the discussion is focused on support transparency or interface transparency.

The term "transparency of the medium" is used to describe a situation in which the medium presents itself as transparent and accessible. In the context of computer-based media, this could include open-source technologies or free software. It is not my intention to provide a detailed analysis of the differences between the two movements. However, it is important to note that a key aspect that unites them is the accessibility of the source code (and, in the case of free software, the option of modifying it according to one's own requirements). This could be regarded as an instance of transparency of the support, but it is also important to consider that, according to this account, the support can only be genuinely transparent to those individuals (or other machines) who are able to read, interpret, and potentially modify the code. For the average user, this form of transparency (which, according to the informational account of transparency, would appear to be transparent) may not satisfy the social requirements of transparency.

To address this challenge, significant efforts have been made throughout the history of personal computing to develop interfaces that are perceived as

"transparent" in terms of user-friendliness. Nevertheless, this transparency assumes an illusory quality. The interface offers a selective view, translating computational elements into human-experienced forms while simultaneously shaping perception by obscuring functional and operational aspects. Despite this partial vision, digital interfaces present themselves as *transparent totalities*, claiming to reveal all of being and provide unprecedented access to the world[19]. This creates the *illusion* of full access to information and a virtual world at the user's disposal. The epistemological and ethical consequences are significant: users may believe they fully understand the real world while feeling detached from responsibility, acting as if in a fictional realm[20]. The crucial mediation process is overlooked, as the design promotes the illusion of immediate access to both the real and virtual worlds.

Two additional conceptualisations of transparency in technology are transparency in operations and transparency in design[21]. The former concerns the utilisation of technologies and the notion of information sharing and accountability on the part of the entity employing specific technologies or processes, particularly in relation to user data. The latter emphasises the necessity for public regulation of the explicability of technological processes and the importance of source code auditability. These two conceptualisations of transparency can be situated within the informational account.

4. What kind of transparency is necessary to have confidence?

As illustrated at the outset with reference to Taddeo, trust is a crucial element in intricate social interactions, as it enables the conservation of energy—both attentional and otherwise—that would elsewise be expended on the constant monitoring of others' behaviour. This also applies to the energy we would expend in the constant supervision of technologies. The development of artificial intelligence and the constant research into the design of autonomous agents that can assist us in activities,

[19] A critique of this simplistic notion of transparency, which fails to acknowledge the inherent limitations of the interface and the inevitable trade-off between visibility and obscurity, can be found in M. Carbone and G. Lingua, *Toward an Anthropology of Screens. Showing and Hiding, Exposing and Protecting*, Palgrave Macmillan, Cham 2023, pp. 107 ff.

[20] This discussion is not intended as a comprehensive development of this topic. However, I have previously discussed it in greater depth in a few publications, the most recent of which are F. Striano, *The dangerous liaison between rape culture and information technologies. Reality, virtuality, and responsibility in cyber-rapes*, in M. L. Edwards and O. Palermos (eds.), *Feminist Philosophy and Emerging Technologies*, Routledge, London 2024, pp. 74-94 and Id., *Violenza virtuale. Vita digitale e dolore reale*, il Saggiatore, Milano 2024.

[21] Cf. D. Kwan, L. M. Cysneiros and J. C. S. do Prado Leite, *Towards Achieving Trust Through Transparency and Ethics*, in «2021 IEEE 29th International Requirements Engineering Conference Proceedings», pp. 82-93: 88-90. These two conceptualisations of technological transparency can be linked to what David Heald in political theory calls "event transparency" (transparency of inputs, outputs, and outcomes) and "process transparency" (transparency of procedures and decision-making processes): cf. C. Hood, D. Heald (eds.), *Transparency: The Key to Better Governance?*, The British Academy, Oxford 2006, pp. 30-32.

decisions, and even artistic productions is precisely in the direction of relieving us of certain tasks and ceding control to these agents, which we must be able to trust.

It is imperative that any entity designated as an artificial agent, or indeed any person or object in which we place trust, is able to demonstrate reliability and earn our trust. The objective is to instil a disposition of confidence in the subject. Both institutionally and in the scientific literature[22], transparency is regarded as a means of reinforcing this disposition. In light of the preceding discussion on transparency, it is now necessary to consider which of the various accounts presented should be adopted and whether an increase in transparency truly corresponds to a strengthening of the confidence disposition.

It can be argued that an informational account is a fundamental requirement for fostering confidence in technologies. A technology that enables comprehensive access to data, performance, and metrics appears to be one that has no hidden aspects, a reliable and trustworthy technology. This level of access can potentially enhance the quality of performance and ethical use due to the transparency not only of the support but also of the entire process.

Nevertheless, it is not necessarily the case that mere equal access to information is in itself beneficial for the establishment of trust. The lack of guidance to transform data from disparate sources into meaningful information, or the overflow of information not accompanied by interpretative support, could potentially lead to a state of stress for the user, resulting in a loss of trust in a system that, while sharing access to its algorithm with any individual, employs a language that is opaque to those lacking the requisite expertise.

This is why the social account of transparency places emphasis on the involvement of a number of factors in order to guarantee effective transparency. The construction of meanings and the interpretation of information are collaborative processes that involve the interplay between content, viewer, and medium. Therefore, transparency can be conceptualised as a virtuous relationship between these three poles, which is a crucial aspect in the establishment of trust. However, even within this account, it is unclear whether the disposition of confidence within the human user (the viewer) is exclusively contingent on the establishment of a transparent relationship[23]. It is not necessarily the case that the medium must be transparent in order for it to be trusted. Indeed, it may have to conceal certain functions in order to make itself comprehensible. It is similarly unproven that the viewer desires transparency of content in order to be able to trust it. Sometimes, we elect to place our trust in a given entity precisely because this relieves us of the necessity of continually monitoring the process in all of its parts and of bearing the responsibility (however shared with the medium and other actors in the process) of interpreting and managing a substantial amount of information.

---

[22] See both the EU documents and the articles that have been previously cited.

[23] Transparency can be seen as the opposite of secrecy, but it does not necessarily limit deception, manipulation, or disinformation. See, in this regard, C. Birchall, *Radical Secrecy: The Ends of Transparency in Datafied America*, University of Minnesota Press, Minneapolis-London 2021, pp. 71-74.

We were confronted with a dilemma: on the one hand, transparency seems to be a crucial factor in establishing confidence. If a technology were entirely opaque and its designers were to refuse to share information about its development, we would likely be inclined to distrust it. This is evident in the case of AI and decision-support algorithms, where the fear of these systems replacing human decision-making stems from our lack of understanding of their functions and processes. On the other hand, complete and total transparency does not necessarily foster confidence. Constant monitoring of processes suggests a shift from trust to user control, undermining confidence. Moreover, the availability of an overwhelming amount of information that users are unable to interpret can lead to increased opacity and mistrust rather than clarity[24]. Conversely, there are instances when we appear to repose trust or reliance in it, specifically to avoid the burden of interpretation that transparency (at least as understood according to the social account) places on us.

The social account of transparency, however, appears to indicate that rather than seeking a prescriptive formula for achieving the optimal degree of transparency to foster confidence in the socio-technical system, the most promising approach is to strike a balance between attitudes. This entails cultivating in the human subject the capacity for interpretative engagement and in the medium a certain degree of transparency, which should be balanced with explicability and usability. The objective, therefore, should be to foster a *virtuous* relationship with technology.

5. Virtue-Oriented Approaches

A core tenet of virtue ethics is the assertion that a virtuous individual is capable of making decisions that are morally sound. The question thus becomes not "how should I behave?" but "what kind of person do I want to be?" The cultivation of virtues—being a virtue defined as "an acquired human quality the possession and exercise of which tends to enable us to achieve those goods which are internal to practice and the lack of which effectively prevents us from achieving such goods"[25]—and the knowledge of when to exercise them through *phronesis* (practical wisdom) enables the betterment of the individual. It is not the case that we are good people if we perform just and moral actions; rather, it is the case that if we are virtuous, we will behave correctly. This perspective also extends to the relationship between humans and technologies: to use the technologies at our disposal virtuously (i.e. to be people who use them to behave virtuously, but also to use them to make the most of their potential) is to achieve the goods which are internal to techno-social practices.

[24] Cf. M. Ananny, K. Crawford, *Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability*, in «New Media & Society», vol. 20, n. 3, 2019, pp. 973-989: 979. This is the so-called "data smog" phenomenon. See also C. Birchall, *Radical Secrecy*, cit., p. 71.

[25] A. MacIntyre, *After Virtue: A Study of Moral Theory*, University of Notre Dame Press, Notre Dame 1984, p. 191.

Nevertheless, often "the technologies' users lack such dispositions and habits of a virtuous person."[26] Furthermore, the technologies themselves do not always appear to be designed with the intention of fostering or cultivating virtues. Firstly, it is necessary to consider which virtues should be cultivated in order to establish a more positive relationship with the contemporary socio-technical system.

In certain theoretical frameworks, trust is conceptualised as a virtue[27] that can inform our engagement with technology. In accordance with a virtue-oriented approach, a trust evaluative act is not contingent upon objective and unambiguous data; rather, it is shaped by an individual's attitude towards technology and a number of factors such as "our beliefs, values, the nature of the problem, how we perceive the problem, prior decisions, availability of alternative options, etc."[28] Nevertheless, this attitude must be founded upon the reliability of the technology in question[29]. This perspective aligns with the account I previously outlined in section 1, which posits that confidence is a disposition shaped by reliability, and that it motivates the expression of evaluative trust based on the disposition itself, rather than on rational arguments.

Accordingly, a virtue-oriented approach should encourage the utilisation of technologies by prioritising the cultivation of virtue traits and fostering a robust disposition towards confidence. Confidence should, in turn, be nurtured by trustworthy technologies, as they too are, in a sense, virtue-oriented. Value Sensitive Design (VSD), for instance, could be taken as a design model, as its tripartite model reserves a prominent place for the incorporation of values useful for human flourishing within the technology.

VSD is a design theory that views techno-social processes as dynamic and iterative, engaging researchers, designers, engineers, and policy-makers in the design

---

[26] A. Bilal, S. Wingreen, R. Sharma, *Virtue Ethics as a Solution to the Privacy Paradox and Trust in Emerging Technologies*, in «Proceedings of the 3rd International Conference on Information Science and Systems (ICISS 2020)», pp. 224-228: 225.

[27] McCraw argues that epistemic trust, understood as an attitude of dependence and confidence in another subject considered trustworthy and authoritative, is a virtue that enables knowledge to be gained through witnessing (cf. B. W. McCraw, *Virtue Epistemology, Testimony, and Trust*, in «Logos & Episteme», vol. V, n. 1, 2014, pp. 95-102 and Id., Proper Epistemic Trust as a Responsibilist Virtue, in K. Dormandy (ed.), *Trust in Epistemology*, Routledge, London 2019, pp. 189-217). Hills, instead, argues that *trustworthiness* is a moral virtue in that it is a disposition to take responsibility for what is entrusted, requiring correct values, motives, and judgements, and contributing to a good life and a well-functioning society (cf. A. Hills, *Trustworthiness, Responsibility and Virtue*, in «The Philosophical Quarterly», vol. 73, n. 3, 2023, pp. 743-761). On trustworthiness as a virtue see also N. N. Potter, *How Can I Be Trusted?: A Virtue Theory of Trustworthiness*, Rowman & Littlefield, Lanham 2002. For a conception of *reliability* as an intellectual virtue – insofar as it implies a combination of cognitive ability, stability of character, and a response to reasons and foundations of justification, not merely an ability to arrive at true beliefs – see R. Audi, *Rational Belief: Structure, Grounds, and Intellectual Virtue*, Oxford University Press, Oxford 2015, pp. 85-97.

[28] A. Bilal, S. Wingreen, R. Sharma, *Virtue Ethics as a Solution to the Privacy Paradox and Trust in Emerging Technologies*, cit., p. 225.

[29] *Ibid.*

process. VSD follows a threefold methodology: conceptual investigation to identify stakeholders, their values, and potential conflicts; empirical investigation using social science methods to study stakeholders' values and motivations; and technical investigation to assess how technologies support or hinder these values. These investigations are interrelated, often conducted together, and iterated throughout the design and testing phases[30].

The concept of transparency also appears to be pertinent in this context. It is noteworthy that transparency is regarded as a socially oriented value in VSD, particularly in the development of responsible artificial intelligence[31].

In this sense, particularly when considered within the context of the social account, transparency can be conceptualised as a virtue, or at the very least, as a means of exercising it, or of gaining trust and strengthening the disposition of confidence. It seems reasonable to posit that a certain degree of transparency on the part of the medium and an effort to increase interpretative capacities are necessary for the establishment of a virtuous and trusting relationship between humans and technology. However, there is also an alternative approach to transparency from the perspective of virtue ethics.

Shannon Vallor, for instance, links the new cult of transparency to *sousveillance*, understood as "contemporary culture of expanding, reflexive, and manifold forms of watching and being watched."[32] In this context, transparency is understood to encompass not only the transparency of institutions, instruments, and means, but also the transparency of individuals with regard to their own selves. The processes of datafication and self-tracking afford individuals access to information about themselves that they might otherwise be unaware of. In this way, proponents of the Quantified Self Movement might argue that these processes enable individuals to cultivate and enhance their own self.

While transparency can act as a deterrent to illegal or immoral behaviour, particularly if this transparency is imposed on institutions or companies that develop, for example, decision-making technologies or artificial intelligence models (increasing trust in such institutions or companies), it is also true that the demand for absolute transparency can lead to a loss of space for free moral and cultural play. Furthermore, the utilisation of technologies to perpetuate and reinforce moral and political practices that are perceived as virtuous can impede and obstruct moral and political experimentation. This, in turn, may contribute to a decline in confidence in the capacity to adapt[33].

---

[30] Cf. B. Friedman and D. G. Hendry, *Value Sensitive Design: Shaping Technology with Moral Imagination*, The MIT Press, Cambridge (MA) 2019.

[31] Cf. J. Dexe, U. Franke, A. A. Nöu, A. Rad, *Towards Increased Transparency with Value Sensitive Design*, in H. Degen, L. Reinerman-Jones, (eds.), *Artificial Intelligence in HCI. HCII 2020. Lecture Notes in Computer Science*, Springer, Cham 2020, pp. 3-15.

[32] S. Vallor, *Technology and the Virtue: A Philosophical Guide to a Future Worth Wanting*, Oxford University Press, Oxford 2016, p. 190.

[33] Cf. *ibid.*, p. 191.

The contemporary cult of transparency, driven by the ideas of the neutrality of technology and a certain degree of techno-fatalism[34], has the potential to become a significant vice, which could limit our ability to flourish. This is because it is based on a quantitative and supposedly universal concept of the self, which is assumed to be objective and devoid of the particular characteristics of individuals. Vallor does not oppose the examined life or the cultivated self, which are fundamental and preparatory to securing a good life, providing it with the possibility of enrichment and enlargement. Nevertheless, she points out that although every attempt to cultivate one's self is also shaped by social and cultural forces, this does not mean that a single universal model valid for all and a single technique for achieving the goal of a good life can be imposed on cultivation[35]. The Quantified Self Movement[36], however, appears to pursue this objective through the means of a supposed objectivity and a quantitative model.

In this sense, transparency is not a virtue; rather, it is a vice to be avoided on the road to developing a virtuous relationship with technologies. Furthermore, the notion of transparency, particularly when interpreted through the lens of mere information disclosure, has the potential to disempower users. This is because it subjects them to an overwhelming influx of data, which may nudge their actions in ways that align with the agendas of policymakers or technology companies, without prompting them to reflect on their own intentions and goals[37]. Such circumstances may result in a reduction of trust and an increase in suspicion of manipulation.

A further, more radical perspective on the relationship between trust and transparency is presented by C. Thi Nguyen[38]. The author posits that, in fact, transparency and trust are radically alternative concepts. Vallor's concept of the "cult of transparency" (which can be seen as a cult of surveillance) impose to experts to effectively communicate their reasoning to non-experts, as Nguyen argues. However, it is inherent to the nature of expertise that the reasoning employed by experts is not always accessible to non-experts. This would result in experts communicating only that which can be publicly justified, undermining the practical application of expertise itself (what Nguyen calls the *epistemic intrusion argument*)[39].

Similarly, the imposition of transparency on individuals or communities will, paradoxically, result in an increase in the prevalence of untruths. This is because some individual or community deliberations are based on intimate reasons and discourses that are challenging to articulate outside the group to which they belong. Consequently, there is a tendency to fabricate reasons that are not entirely accurate in

[34] Cf. *ibid*., p. 193.
[35] Cf. *ibid*., p. 198.
[36] Cf. *ibid*., pp. 198-202.
[37] Cf. *ibid*., pp. 202-204.
[38] Cf. C. T. Nguyen, *Transparency is Surveillance*, in «Philosophy and Phenomenological Research», vol. 105, n. 2, 2022, pp. 331-361.
[39] This argument is similar to Birchall's argument on self-censorship: cf. C. Birchall, *Radical Secrecy*, cit., p. 71.

order to justify certain decisions in the name of transparency. This suggests that it is not possible to place trust in any individual or deliberating community that claims to be transparent.

When applied to technology, this can be illustrated as follows: (1) experts will only provide an explanation that is comprehensible to the public, therefore a complex technology can never be truly explicable; (2) despite the metaphors associated with AI, the technology itself "reasons" differently from a human and cannot therefore be totally transparent to the human, instead, it can at best deceive the human through supposedly transparent interfaces.

According to Nguyen, the inherent tension between transparency and trust is a genuine moral dilemma that cannot be easily resolved. Vallor presents a less pessimistic view, suggesting that flourishing is possible despite the challenges of transparency/surveillance[40]. She proposes that this is contingent upon the cultivation of *technomoral virtues*[41]. In particular, the virtue that Vallor links to trust and reliability is *honesty*.

6. Conclusion: Towards Honest Technologies

Reliance on technology, while often measurable through performance outcomes, presents significant limitations. Its focus on results can detract from the importance of the processes involved, which are critical in establishing a trustworthy relationship with technology. Both trust and reliance place substantial expectations on the success of an interaction, but they do not guarantee it, which can lead to frustration. Moreover, the letter concept appears to depend heavily on transparency, while transparency seems to have at least an ambiguous relationship with transparency. This relationship, moreover, raises the unresolved issue of what conception of transparency should be applied. The specific understanding of transparency can significantly alter—or even undermine—the type of reliance or trust that is developed. Transparency, in this sense, can even negatively impact our relationship with technologies and those interactions mediated by them.

Instead, technology should aspire to be honest—maintaining and making evident the sense of interaction and encouraging hermeneutical attention[42]. This aligns more closely with a social account of transparency, where the relationship among content, viewer, and medium must be based on honesty. I propose that technologies fostering such honest interactions be termed "honest technologies."

By this terminology I mean to refer to technologies that are not only (or perhaps not so much) transparent, but that do not conceal mediation or interaction.

---

[40] Cf. S. Vallor, *Technology and the Virtue*, cit., pp. 204-207.

[41] In her taxonomy of technomoral virtues, Vallor identifies a total of 12 virtues: Honesty, Self-Control, Humility, Justice, Courage, Empathy, Care, Civility, Flexibility, Perspective, Magnanimity, Technomoral Wisdom (cf. *ibid*., p. 120).

[42] Cf. F. Striano, *Towards "Post-Digital". A Media Theory to Re-Think the Digital Revolution*, in «Ethics in Progress», vol. 10, n. 1, 2019, pp. 83-93: 90.

The kind of honest technologies I envisage do not seek explicability at any cost, or at least not to the public. They must, however, make the processes of mediation and interaction evident and negotiate with the user the kind of information the user has an interest in accessing. Technologies designed to be honest, in short, must not sacrifice mediation awareness in the name of usability, and their honesty, combined of course with reliability, will be a far more powerful confidence booster than any ambiguous call for transparency.

Furthermore, when considering human relationships *mediated by technology*, honesty is once again central. The balance between community building and the right to privacy, including the right to be forgotten, is delicate. A blanket call for transparency may not be the solution; in fact, it may lead to an overexposure that contradicts an individual's desire for privacy. Here again, the virtue of honesty, as a technomoral and social virtue, becomes crucial. Drawing from Vallor's insights, honesty requires "*an exemplary respect for truth*" (aligned with the Buddhist concept of "Right View") and "*the practical expertise to express that respect appropriately in technosocial contexts*"[43] (akin to "Right Conduct"). The invitation to behave honestly thus provides a superior resolution to the tension between community ties or expert reasoning on the one hand and the demand for truthful information on the other, as highlighted by Nguyen, than the call for transparency.

This is not to suggest that there are no circumstances in which transparency should be demanded or that it is invariably a vice of the contemporary communication model. The practice of full disclosure can be perceived as a form of viciousness when it involves the indiscriminate sharing of information "without discernment or contextual sensitivity:"[44]

> an honest scientist must know her audiences and venues, and there are times when excessive precision or transparency will obscure the truth rather than reveal it. […] Still there are circumstances in the life of government, academia, business, and other institutions—even the private lives of citizens—that justify opening the books for inspection, so to speak. We need the technomoral wisdom to make more intelligent and practically discerning use of the new technologies that can make opening the books in such circumstances easier.[45]

Vallor invites us to reflect also on who our models of technomoral honesty might be[46]. I extend this reflection by considering the influence that technology itself can have on our traits, conduct, and habits. Without succumbing to radical technodeterminism and admitting that the reasons for an act must be internal[47], we can acknowledge that, just as culture and examples shape our cultivation of virtues

[43] Cf. S. Vallor, *Technology and the Virtue*, cit., p. 122.
[44] *Ibid.*, p. 205.
[45] *Ibid.*, pp. 205-206.
[46] Cf. *ibid.*, pp. 122-123.
[47] Cf. S. van Hooft, *Caring: An Essay in the Philosophy of Ethics*, University Press of Colorado, Niwot 1995, pp. 140 and ff.

and internalization of motivations, so too can our habitual use of technology. If honest individuals design honest technologies, it is also true that habitual interaction with honest technologies can cultivate honesty in individuals.

Go Hack Yourself! Transparency Through the Lens of Biohacking[a]

Giustina Benedetta Baron[*], Accursio Graffeo[†]

Abstract

L'antropologia e gli studi sociali hanno ampiamente studiato le culture del self-tracking, ma i potenziali risultati del "framework del biohacking" rimangono relativamente poco esplorati. Il biohacking incarna una forma distintiva di tecno-ascetismo moderno con le sue norme uniche di autoregolazione del corpo. Come verrà chiarito, questo paradigma stabilisce nuovi "spazi di visibilità" in cui le informazioni relative al corpo e alle sue funzioni interne sono rese trasparenti, organizzate e condivise. Tuttavia, l'intricata politica che circonda la scienza aperta trascende una dicotomia semplicistica tra trasparenza e chiusura. È necessaria un'esplorazione più approfondita delle attuali trasformazioni non solo all'interno della ricerca scientifica, ma anche dei quadri epistemologici ad essa associati. Partendo da queste basi, questo studio cerca di contestualizzare gli approcci antropologici contemporanei al corpo all'interno di un panorama più ampio, esplorando il loro allineamento con modelli distinti di elaborazione delle informazioni e culture sanitarie alternative che possono influenzare le risposte tipologiche al paradigma dominante stabilito dai discorsi sul biohacking che enfatizzano la trasparenza attraverso la raccolta dei dati.

Parole chiave: biohacking, trasparenza, antropologia, corpo, Itskov

The realm of anthropology and social studies has extensively investigated self-tracking cultures, yet the potential outcomes of the "biohacking framework" remain relatively underexplored. Biohacking embodies a distinctive form of modern techno-asceticism with its unique norms for self-regulation of the body. As will be elucidated, this

[a] The conception of the article is by both authors. However, the drafting of paragraphs 1 and 4 is due to Baron and Graffeo, that of paragraph 3 to Baron, that of paragraph 2 to Graffeo. Saggio ricevuto in data 31/05/2024 e pubblicato in data 22/01/2025.

[*] Dottoranda, Dottorato nazionale in Studi Religiosi, email: giustinabenedetta.baron@unimore.it.
[†] Dottorando, Dottorato nazionale in Studi Religiosi, email: accursio.graffeo@unito.it.

paradigm establishes novel "spaces of visibility" where information regarding the body and its internal functions is rendered transparent, organized, and shared. Nonetheless, the intricate politics surrounding open science transcend a simplistic dichotomy between transparency and closure. It necessitates a more profound exploration of current transformations not only within scientific research but also concerning its associated epistemological frameworks. Building upon these foundations, this study seeks to contextualize contemporary anthropological approaches to the body within a broader landscape, exploring their alignment with distinct models of information processing and alternative health cultures that may influence typological responses to the dominant paradigm set forth by biohacking discourses emphasizing transparency through data collection.

Keywords: biohacking, transparency, anthropology, body, Itskov

1. Introduction

Biohacking, a term that has gained popularity in recent times, can be traced back to a 1988 article by journalist Michael Schrage in the Washington Post[1]. In this article, Schrage observed a subculture of individuals conducting techno-biological experiments in their garages. He contemplated the implications of perceiving life as something to be mapped out, comparable to a "computer program" with its outcomes manifested not on "papers" but on protein chains:

> What happens, for example, if future generations begin to see life as something that's manipulable – just another computer program, but one in which the printout isn't on paper but in proteins? If children grow up believing that life is nothing more than organic chemistry?[2]

Over the past thirty-five years, biohacking has not only endured but also risen in prominence, as the questions surrounding the understanding and manipulation of biological life at the molecular level have become increasingly urgent. From its origins, as an underground and secretive subculture operating within garages, biohacking has evolved into a *normalized* concept, frequently discussed in public discourse regarding citizen access to biomedical science and personal health responsibility[3].
Currently, the biohacking community encompasses various subgroups with distinct objectives. One common characteristic is that biohackers conduct biological, biomedical, or biotechnological experiments outside institutional settings such as universities or private medical companies. Proponents argue that biohacking is closely associated with self-tracking, therapeutic, and enhancement technologies that aim to

[1] M. Shrage, *Playing God in your basement*, «The Washington Post», January 31, 1988.
[2] *Ibidem.*
[3] M. Grewe-Salfeld, *Biohacking, Bodies and Do-It-Yourself. The Cultural Politics of Hacking Life Itself*, transcript, Bielefeld 2022, pp. 17-33.

explore one's "hidden biology", using technology as a tool for mapping and making it *transparent* and accessible to everyone[4]. Andrew Pickering describes it as an adaptive worldview and a "performative ontology of the black box" as functionally indispensable for cybernetics[5]. This same black box is reframed by Dave Asprey[6] who promotes his brand of self-experimental performance science by viewing his body as a system amenable to manipulation through experimentation on its input-output correlations. Similarly, Minna Ruckenstein and Mika Pantzar identify *transparency*, *optimization*, and *feedback loops* as grounding metaphors guiding the practice of self-discovery through data collection, driven by the idea of "quantifying life" to manage its constitutive feedback loops and correlations[7].

Members of biohacking maintain direct connections with the hacker movement. Their organizational structures frequently mirror those of *hackerspaces* (community-operated facilities) where individuals gather to partake in hacking activities and discussions related to computing. Notably, the "hacker ethic" embodies a codified set of moral principles. Steven Levy[8] articulates this ethic through several foundational tenets: unrestricted access to computers; the belief that all information should be freely accessible; skepticism towards authority[9]; assessment based on hacking expertise rather than traditional metrics such as academic credentials;

[4] Ivi, p. 222.

[5] A. Pickering, *The Cybernetic Brain: Sketches of Another Future*, University of Chicago Press, Chicago 2010.

[6] D. Asprey, *Head Strong - The Bulletproof Plan to Activate Brain Energy to Work Smarter and Think Faster - In Just Two Weeks*, Harper Wave, New York 2017.

[7] M. Ruckenstein, M. Pantzar, *Beyond the quantified self: thematic exploration of a dataistic paradigm*, «New Media & Society», 19, n. 3, 2017, pp. 401-418.

[8] S. Levy, *Hackers: Heroes of the Computer Revolution*, O'Reilly, Cambridge 2010.

[9] In this context, it is essential to highlight that since 2016, there has been a growing discourse among scholars and commentators regarding the concerns surrounding the nexus between the proliferation of misinformation and crises within democratic processes. This relationship has been shown to correlate with an increasing skepticism towards scientific authority and formal political engagement (A.M. Enders, J.E. Uscinski, M.I. Seelig, et al., The relationship between social media use and beliefs in conspiracy theories and misinformation, «Polit Behav», 45, 2021, pp. 781-808; M. Kim, X. Cao, The impact of exposure to media messages promoting government conspiracy theories on distrust in the government: Evidence from a two-stage randomized experiment, «International Journal of Communication», 10, 2016, pp. 3808–3827; E.C. Tandoc, D. Lim, R. Ling, Diffusion of disinformation: How social media users respond to fake news and why, «Journalism», 31, n. 3, 2020, pp. 381–398; S. Valenzuela, D. Halpern, J.E. Katz, J.P. Miranda, The paradox of participation versus misinformation: Social media, political engagement, and the spread of misinformation, «Digital Journalism», 7, n. 6, 2019, pp. 802–823). Within this context, the DIY biology movement arises as a proactive countermeasure to the proliferation of conspiracy theories and misinformation, thereby championing an alternative paradigm of scientific inquiry. Indeed, with the Internet evolving into a genuine mass medium, postmodern skepticism has catalyzed the formation of a novel scientific framework that skillfully utilizes all available codes offered by transmission media. This development is facilitated by the burgeoning "democracy" of cyberspace, which fosters narratives in which the individual subject is made visible and heard (M. Lock, Cultivating the Body: Anthropology and Epistemologies of Bodily Practice and Knowledge, «Annual Review of Anthropology», 22, 1993, pp. 133–155).

acknowledgment that artistic expression can thrive through computing; and the conviction that computers hold transformative potential for individual lives[10]. Notably, the practice of "hacking" as a means of generating foundational knowledge has subsequently evolved into a vast array of methodologies for managing the increasing production of data, thereby facilitating instantaneous exchange of information and its re-materialization through different coding operations, which result in a hybridization of different interaction strategies within the *digital milieu*[11].

Moreover, what distinguishes biohacking as a contemporary form of techno-asceticism focused on "self-improvement" is its simultaneous idealization of *nature* and the *machinic body* as metonyms. This idealization extends beyond technological innovation and incorporates a *biomimetic* approach, drawing inspiration from natural processes, emphasizing their innate capacities for healing, adaptation, and regeneration[12]. Remarkably, biohackers often seek to align their interventions with principles found in nature, in order to replicate or enhance its complexity, efficiency, and resilience within their own biological systems. However, this perspective tends to overlook the complexities inherent in the concept of *nature* itself, thus neglecting to recognize it as a space for contestation and politics[13].

A noteworthy precursor to this biomimetic approach is evident in an article that reported on the International Human Genome Sequencing Consortium's initial sequencing and analysis of the human genome. The authors of this comprehensive survey reached the conclusion that:

> Finally, it has not escaped our notice that the more we learn about the human genome, the more there is to explore: We shall not cease from exploration and the end of all our exploring will be to arrive where we started. And know the place for the first time[14].

As expected, the initial creation of the map did not represent the culmination of exploratory efforts. Instead, the following fifteen years have seen a remarkable proliferation in mapping activities, which shaped biohackers' ethos regarding *transparency*. This concept was envisioned by Donna Haraway, who articulated that "maps are models of worlds crafted through and for specific practices of intervening and ways of life"[15]. In this regard, the exploration envisioned by the authors

[10] A. Delfanti, *Tweaking Genes in Your Garage: Biohacking between Activism and Entrepreneurship*, in W. Sützl, T. Hug (eds.), *Activist Media and Biopolitics. Critical Media Interventions in the Age of Biopower*, Innsbruck University Press, Innsbruck 2012, pp. 163-178.

[11] M.G. Sindoni, *Spoken and written discourse in online interactions: a multimodal approach*, Routledge, New York 2013.

[12] A. Lindfors, *Between Self-Tracking and Alternative Medicine: Biomimetic Imaginary in Contemporary Biohacking*, «Body & Society», 30, n. 1, 2024, pp. 84-85.

[13] H. Helen, *Xenofeminism*, Polity, Cambridge 2018.

[14] International Human Genome Sequencing Consortium, *Initial sequencing and analysis of the human genome*, «Nature», 409, February 15th, 2001, pp. 860–921.

[15] D. Haraway, *Modest_Witness@Second_Millennium. FemaleMan_Meets_OncoMouse. Feminism and Technoscience*, Routledge, New York 1997, p. 135.

concerning the Human Genome Project, whose objective is "to arrive where we started" implies that once "life's inner workings are mapped and revealed by science" biology will have undergone a deep transformation, interrogation, and re-articulation. In fact, biological data that not only contain biological instructions that constitute our identity but are inscribed and shared in ways that are strategically intertwined with an extensive array of "institutions, procedures, instruments, practices, and forms of capitalization"[16].

The significance of this perspective lies in its elucidation of a new scientific epistemology, which ultimately contemplates the notion of an impending post-humanity or post-nature, arising from the profoundly transformative ethos of the "hackers culture" that is both facilitated by and facilitates "particular forms of institutional structures"[17]. Sunder Rajan articulates this change as "a shifting grammar of life, towards a future tense"[18]. Such an interpretation is inherently *linguistic* as it underscores how coding and programming are integral to processes of scientific expression and regulation that can be transcribed in various ways. Occasionally, this type of approach regarding exploration of "life" is framed as an *origin myth*: a return to primordial times when all was inherently "bio-based" rather than reliant on fossil resources underpinning contemporary industrial civilization.

The movement experienced significant growth in the United States around 2005 before spreading globally. However, it was not until 2008 that biohacking became organized at a societal level under the name DIY-biology in Boston. Remarkably, in Russia, biohacking became widely known in 2017 when an article by Sergey Faguet was published, detailing how he managed to become more energetic and reduce his biological age through detailed study of vital signs[19]. As he wrote one year later in his personal Facebook account[20]:

> You're a biorobot. Observe your programs. Rewrite ones that you don't like […] My aim was to enhance my vitality, well-being, happiness, self-assurance, determination, and intellect, while also enhancing my emotional state and focus, and

[16] N. Rose, *The politics of life itself*, «Theory, Culture & Society», 18, n. 6, 2001, pp. 1–30: 13-15.
[17] K.S. Rajan, *Biocapital: The Constitution of Postgenomic Life*, Duke University Press, Durham 2006, p. 14.
[18] The advancement of life sciences is transitioning from an ahistorical past toward a perpetually attainable horizon, striving towards fulfilling the promise of a future wherein science subjugates life and nature (S. Tamminen, E. Deibel, *Recoding life: information and the biopolitical*, Routledge, London, 2018, pp. 4-5).
[19] S. Faguet, *Мне 32 года, и я потратил $200 тысяч на "биохакинг"*, vc.ru, 2017. https://vc.ru/future/26886-personal-biohacking.
[20] Original source in Russian: Ты биоробот. Наблюдай свои программы. Перепиши те, которые тебе не нравятся. (…) Мне хотелось стать более энергичным, здоровым, счастливым, уверенным, волевым и умным, улучшить настроение и концентрацию, а также продлить свою жизнь. Последние 4-5 лет я занимаюсь биохакингом тела и разума с помощью логики и научного подхода. Для этого я оптимизировал сон, питание и тренировки, прошёл через тысячи тестов, принял десятки разных препаратов и сотни добавок (…) работал вместе с великолепными врачами, медитировал более тысячи раз, ходил к психотерапевту — и потратил на всё это примерно двести тысяч долларов. Translation by the authors.

> prolonging my lifespan. Over the past 4-5 years, I have been involved in the optimization of both physical and mental aspects through a methodical and scientific approach. This entailed refining my sleep patterns, dietary habits, and exercise routines; undergoing numerous medical examinations; […] engaging in meditation extensively; seeking therapeutic assistance - all resulting in an expenditure of approximately two hundred thousand dollars[21].

As such, whether in the guise of epistemological shift, enhancement, or information, the narrative of biohacking has two prominent consequences. On the one hand, it creates a new form of *transparency*. The "transparent body"[22] in the sense of a manageable entity, mapped through scientific tools can be seen as a direct outcome of development of sophisticated medical imaging and information technologies. This transparency now extends even to those spaces and structures in the sub-microscopic regions and is facilitated to a large extent by information technologies that allow for elaborate simulations. This instance underscores the ongoing emphasis within the field of biology and bioresearch on comprehensively investigating both the physical properties and the informational foundation underlying genetic data that create, to use Thomas Lemke's words, "spaces of visibility"[23] in which information about the body and its inner workings is made *transparent*, *intelligible*, *imaginable*. These spaces of visibility, as Lemke argues, do not just concern individuals: rather, genetic diagnostics offer predictive information about individuals but also their descendants, creating a "new, transgenerational transparency of the body"[24].

However, this transparency, as of now, is being interpreted: without this step of *translation*[25], the images and narratives circulating in (popular) culture would not be accessible for the lay public. In fact, as we will discuss later, transparency operates as an *ideology*, a syntagmatic organization of values[26] that substantiates technological advancements taking precedence over human existence. The distinctive features of this intricate mode of interaction, manifested through conglomerates of linguistic expressions, visual representations, and other non-exclusively linguistic codes, are manifold and span multiple fields of inquiry. The establishment of a new communicative reality that is increasingly recognized as *normative* necessitates a reassessment of the interpretive frameworks traditionally employed in scientific discourse analysis. It implies the need for methodologies that inevitably traverse theoretical domains from various disciplines: areas of study that may be partially

[21] S. Faguet, via Facebook, 2018. https://www.facebook.com/sergef/posts/10104132396121843.
[22] M. Chrysanthou, *Transparency and Selfhood: Utopia and the Informed Body*, «Social Science & Medicine», 54, 2002, pp. 469-479.
[23] T. Lemke, *Disposition and Determinism. Genetic Diagnostics in Risk Society*, «The Sociological Review», 52, n. 4, 2004, p. 555.
[24] *Ibidem*.
[25] D. Monticelli, *Borders and translation: Revisiting Juri Lotman's semiosphere*, «Semiotica», 230, 2019.
[26] A.G. Greimas, J. Courtés, *Sémiotique: dictionnaire raisonné de la théorie du langage*, Hachette, Paris 1979.

related or even distantly connected will find themselves interconnected in transdisciplinary ways.

Therefore, it becomes imperative to explore how transparency intersects with specific idiosyncratic reactions to prevailing societal norms, identifying the anthropo-semiotic elements that might shape specific *typological responses*[27] to this "ideology of transparency". Consequently, an inquiry arises: What is the precise relationship between transparency and cultural patterns (i.e. typology)? What prompts their current interconnection?

We posit that existing studies have not thoroughly explored the political and ideological implications of framing contemporary biohacking as a fusion of self-tracking with alternative or local cultures. As we will discuss, when data-driven self-monitoring is melded with a focus on a specific intellectual heritage, it gives rise to a biomimetic inclination that harmonize the normativity linked with techno-scientific orientations with cultural codes grounded in the typological features of a given culture[28].

To illustrate the structure of the text, we begin with an overview. The initial paragraph serves as an introductory section that addresses biohacking, methodologies, and objectives, while the final segment provides a conclusion or synthesis. The body of the text comprises two distinct sections: one characterized by an anthropological perspective (paragraph 2), aimed at examining the processes involved in "fabricating" human identity; and another section (paragraph 3) adopting an interpretative semiotic framework. This structural design is predicated on the notion that both perspectives can offer significant insights into the phenomenon under consideration, whether approached through a cultural lens pertaining to health-related dimensions or directed towards metaphysical considerations, as exemplified in paragraph 3.

In order to analyze the modalities of the "fabrication" of the human being, we will draw on hermeneutics in the anthropological field in order to outline some points of contact, especially with those aspects of "enhancement" that are central when talking about biohacking, and to open up those spaces of reflection that project us towards the dimension of the possible. The human body, in its broadest sense, may be considered a cultural construct, and thus a potential site for the examination of the various processes of implementation, trimming, and enhancement associated with the shaping of the human.

In this outlook, Itskov's theoretical framework is crucial for understanding the evolution of the normative aspects of openness and transparency that have emerged from biohacking. In fact, we argue that this phenomenon has catalyzed various idiosyncratic *responses*, necessitating a thorough examination of the *encyclopedic* dimension of "transparency", which constitutes the foundational methodological framework for our research. By utilizing this theoretical lens, we investigate how the

[27] V. Strada, *La questione russa. Identità e destino*, Marsilio, Venezia 1991.

[28] To deepen the concept of *typology* in culture, see: J. Lotman, B. Uspensky, G. Mihaychuk, *On the Semiotic Mechanism of Culture*, «New Literary History», 9, 1978.

phenomenon of transparency, as articulated by DIY biology, has engendered various yet interconnected *typological* responses concerning bodily transparency.

2. Anthropological considerations on the "fabrication" of human being

The term *biohacking* has been coined to describe a process of *reprogramming* that affects both the mind and the body of human beings. This process is intended to enable individuals to manage and alter their environment, thereby exerting control over their own biology. The objective is to optimize and update the biology in question.

The capacity to act upon the human body, including its biology, in order to modify, improve or create structural and functional changes at different levels, has been a constant feature of human cultures throughout history.

Many approaches in the field of (cultural and philosophical) anthropology have made interpretative contributions to this phenomenon. In addition to the concept of "anthropo-poiesis" (which will be discussed in greater detail later in this section), formulations such as Michel Foucault's "technologies of the self" or Marcel Mauss's "techniques of the body" find full legitimacy, not forgetting to mention the significant contribution of Peter Sloterdijk, who straddles the fields of philosophical anthropology and historical anthropology with his theorisation of "anthropotechnics". Foucault, for example, considers the human being to be a "recent invention"[29] and introduces the concept of the "technologies of the self":

> [...] which permit individuals to effect by their own means or with the help of others a certain number of operations on their own bodies and souls, thoughts, conduct, and way of being, so as to transform themselves in order to attain a certain state of happiness, purity, wisdom, perfection, or immortality[30].

In Sloterdijk's view, human beings are the result of evolutionary processes that occurred in the context of prehistoric hordes. These processes were characterized by a dynamic interplay between natural and cultural forces, whereby the evolution of human characteristics was shaped by a complex interplay of natural and cultural selection pressures. He is the proponent of the term "anthropotechnics", which denotes the techniques of "production" of the human being. This concept represents a focal point where individual and collective history converge, encompassing the process of hominization and the emergence of culture, the history of existence and the history of civilization. A further step in Sloterdijk's concept of anthropotechnics will situate the link between anthropotechnics and exercise at the center of his analyses. The author's attention will therefore be directed to individuals who, through the practice of exercise, are capable of acting upon themselves and the contexts in

[29] M. Foucault (1966), *The Order of Things: An archaeology of the human sciences*, Routledge, London-New York 2002 (2005).
[30] M. Foucault, *Technologies of the Self. A Seminar with Michel Foucault*, L.H. Martin, H. Gutman, P.H. Hutton (eds.), The University of Massachusetts Press, Amherst 1988, p. 18.

which they are situated, thereby modifying them and themselves[31]. The introduction of athletic-oriented anthropotechnics enables the subject to engage with cultural practices that have been deeply embedded in the evolutionary history of the species. It can be further postulated that biohacking may be regarded as a form of action directed towards established practices, thereby facilitating the autonomization and self-determination of the individual.

The notion of the human being as a product necessitates an analytical approach to understanding, whereby the production process is examined in its constituent parts. It is important to note that the human being is not a finished product, but rather a work in progress. The production of the human being is not a process that can be attributed to human agency. Furthermore, it was never the intention of humankind to create itself in this way. Prior to attaining the status of a fully formed human being, he existed in a state of becoming. The process of human development is seen to be closely linked to the concept of "home" and can be understood as a significant narrative of domestication[32].

The metaphor of the house presents a place that serves to stabilize the gap between the interior and the surrounding environment, thereby amplifying the contrast with the non-interior space. It is a space that is capable of securing its inhabitants, offering them a place in which to reproduce themselves[33]. The inherent vulnerability of the human condition compels individuals to confront the dual realities of their physical and emotional fragility, as well as their inherent restlessness and fluctuating motivations.

The conditions that lead to the formation of man, the entry into the *Lichtung* – the bearer of something prehuman that opens up towards man – can only be created through the action of four synergetic and interrelated mechanisms: insulation, liberation from bodily limitations, neoteny and transposition[34].

At the cultural level, there are also human shaping techniques through which human groups have been able to take care of their own symbolic and disciplinary shaping. These include the shaping of orders and forces (e.g. rituals, habits, codifications and social rules, etc.), which are appropriately referred to as "anthropotechnics". Such techniques are thus designated as such due to their indication of the direct shaping of humanity through a civilizing process[35].

Sloterdijk's suggestions and the concept of "anthropotechnics" certainly provide important stimuli for the study of the concept of "human fabrication" and for the analysis of the fundamental issues surrounding biohacking. In relation to the object of analysis of this paper, a recurring aspect seems to emerge that can be ascribed to biohacking practices and can be traced back to the communicative

[31] A. Sloterdijk (2009), *You Must Change Your Life: On Anthropotechnics*, Polity, Cambridge 2013, pp. 109-110.
[32] A. Sloterdijk (2001), *Not saved: essay after Heidegger*, Polity, Cambridge 2017, pp. 105-108.
[33] Ivi, p. 110.
[34] Ivi, p. 111.
[35] Ivi, pp. 126-127.

subtracks of various biohackers: the need to biohack one's own body and person arises from a deficit, a "lack of", and a simultaneous desire for "enhancement". This deficit (in the subtrack) is declined in the form of a lack of... information[36], health, health insurance[37], treatments, technology, and so on[38]:

> The general desire for – and cultural obsession with – a hacked, healthier or more efficient brain carries this promise [a happy, productive, extended lifespan]. This promise not only positions the human body as inherently deficient but also offers a solution to this lack: to use the tools of biotech to optimize the body[39].

In light of the aforementioned reasons, we will now direct our attention predominantly to a specific strand of anthropological thought that posits that the human inclination to act upon the human body, including its biology, with the intention of modifying, improving or creating structural and functional changes, is a consequence of the incompleteness of the human species. The human being is an "unfinished" animal, capable of completing and refining itself through culture, which is able to bridge the gap between what our body communicates and what we need to know in order to function[40]. This theme of man as an unfinished animal re-emerges in the second half of the 19th century in the philosophy of Friedrich Nietzsche and in the 20th century in the philosophical anthropology of Arnold Gehlen, who reconstructs this genealogy of thought from Herder. For Gehlen, the non-specificity of the human organism is the starting point of everything that is human, a living being biologically endowed with a "deficient equipment". Humans as "beings to be disciplined" are continually threatened by the inherent possibility of failure and, because of their incompleteness, are forced to structure themselves. This theme introduces the necessity of action in order to process - in and outside the self - nature and construct the specific sphere of his life, thereby showing openness to the world[41].

Human beings' innate constitution is as structurally incomplete as it is inefficient. Tools, hunting, social and family organization, art and religion, and even science have all contributed to the modification of the human body in a somatic sense, becoming fundamental elements for survival and, more importantly, for the fulfilment of the human condition[42]. The concept of incompleteness is not limited to the

---

36 A. Wiggins, J. Wilbanks, *The Rise of Citizen Science in Health and Biomedical Research*, «The American Journal of Bioethics», 19, n. 8, 2019, pp. 3-14.

37 J. Keulartz, H. van den Belt, *DIY-Bio - Economic, Epistemological and Ethical Implications and Ambivalences*, «Life Sciences, Society and Policy», 12, n. 7, 2016, pp. 1-19.

38 J. Lee, *The Biohacking Manifesto: The Scientific Blueprint for a Long, Healthy and Happy Life Using Cutting Edge Anti-Aging and Neuroscience Based Hacks,* CreateSpace, South Carolina 2015.

39 M. Grewe-Salfeld, *Biohacking, Bodies and Do-It-Yourself. The Cultural Politics of Hacking Life Itself*, transcript, Bielefeld 2022, p. 144.

40 C. Geertz, *The Interpretation of Cultures*, Basic Books, New York 1973, pp. 48-50.

41 A. Gehlen (1940), *Man. His Nature, and Place in the World*, Columbia University Press, New York 1988.

42 C. Geertz, *The Interpretation of Cultures*, cit., pp. 82-83.

biological aspects of the human condition; it also encompasses the cultural component[43].

This different perspective attributes to culture an emptying operation, which has involved various aspects of the biology of the human being. The cultural element has been posited as a contributing factor in the fading of certain biologically regulated mechanisms, such as walking for food research. This has led to the emergence of a new form of incompleteness, possibly reproposed on another level[44]. The shaping role of culture in relation to human biology is characterized as a process of "pruning" and selection[45], whereby some options are chosen and others are discarded[46].

The concept of the "void" to be filled and the selection of certain options to the detriment of others inevitably leads us to consider human nature, both cultural and biological, as endowed with great plasticity[47]. This plasticity opens to the possibility of continuous redefinition and shaping of the constitutive form of human nature. From the perspective of the human brain, the processes of selection and deselection do not represent a deficit or something hostile; rather, they determine the conditions necessary to enable the brain to express its potential effectively.

In the context of *biohacking*, the concept of plasticity is of significant importance. It constitutes the material potential – of the brain, but also of the entire body – of living organisms, enabling bodies, whether human or animal, to gain a constant capacity for negotiation and exposing the contiguous aspects related to biology and embodiment. The plasticity of the brain is able to challenge the old separations between mind and body, culture and nature, thus blurring the distinction between brain and mind[48]. The vision of a plastic body would therefore seem to indicate the natural and biological human predisposition to operations of recalibration, reconfiguration and extension through the technological mediation, towards a variety of forms[49].

---

43 F. Remotti, *Cultura. Dalla complessità all'impoverimento*, Laterza, Roma-Bari 2011.

44 Ivi, pp. 51-76.

45 The word hacker – from which hacking, present in the term biohacking – comes from the English "to hack" which means "to tear to pieces" or "to break", but it also means "to cut", "to reduce", "to trim", "to open a passage" (Cambridge Dictionary, Cambridge University Press, https://dictionary.cambridge.org), precisely between the lines of code that instruct software programs. The word describes the activity of assembling programs, with little regard for "official" methods, to improve the efficiency and speed of existing softwares.
It is intriguing to observe how the element of reduction, cutting, and trimming somehow evokes the shaping role of culture in relation to human biology. This can be viewed as a process of "pruning" and selection, whereby specific options are chosen for further consideration while others are rejected.

46 A. Favole, S. Allovio, *Plasticità e incompletezza tra etnografie e neuroscienze*, in F. Remotti (a cura di), *Forme di umanità*, Bruno Mondadori, Milano 2002, pp. 167-205: 199.

47 F. Remotti (1996-2011), *Fare umanità. I drammi dell'antropo-poiesi*, Laterza, Roma-Bari 2013, pp. 14-19.

48 V. Pitts-Taylor, *The Brain's Body: Neuroscience and Corporeal Politics*, Duke University Press, Durhman, 2016, pp. 4-5, 24.

49 A. Clark, *Re-Inventing Ourselves: The Plasticity of Embodiment, Sensing, and Mind*, «The Journal of Medicine and Philosophy», 32, n. 3, 2007, pp. 263-282: 278.

The notion that human beings can be shaped or modeled through such processes implies the consideration of the concept of *anthropo-poiesis*[50]. The principle of anthropo-poiesis, according to which, human beings are to be "modeled" and in a certain sense "constructed", is indeed a humanistic ideology, taken up by certain currents of cultural anthropology, but it seems to be an idea shared also by the natural sciences. This sharing primarily concerns the idea of *plasticity* and consequently that of *modeling*[51]. Human body is a bearer of potential; its use is not reduced to mere exploitation as a tool already given by nature, as Marcel Mauss suggests, but implies an act of modeling and training bodily functions and activities, also from an aesthetic point of view and in all the many forms that different societies and traditions have established. The notion of *techniques du corps*[52] dates back to 1936[53] and considers the body to be mankind's first technical object and technical means[54]. Mauss highlights how fundamental the intersection of the biological and the social is to the constitution of bodies as a set of "techniques"[55], e.g. examining how everyday activities such as walking, running or sleeping, appear contextualized, learned and taught, making the body adapt to its purpose in a social context[56]. Some of these simple activities are included in some proposed biohacking programs[57].

The term *anthropo-poiesis* comes under the category of terms that identify the idea of the formation and genesis of human beings. According to this theory, human beings are subjects to be constructed and shaped and, in a certain way, do not know only biological birth, but several births within society[58]. How many births do human beings know? One unavoidable birth is linked to childbirth, coinciding with the exit

[50] In Greek *poiesis*, from the verb *poiein* (to make), expresses the idea of modeling. The Greeks were the first to propose that human beings must be "shaped" (invented) in accordance with the (hidden, to be discovered) directives of "human nature".

[51] cfr. G.F. Azzone, *Perché si nasce simili e si diventa diversi? La duplice nascita genetica e culturale*, Bruno Mondadori, Milano 2010; cfr. L. Maffei, *La libertà di essere diversi. Natura e cultura alla prova delle neuroscienze*, il Mulino, Bologna 2011.

[52] M. Mauss (1936), *Les techniques du corps*, in *Sociologie et anthropologie*, Presses Universitaires de Paris, Paris 1950, pp. 365-386.

[53] This is the year of publication. The paper was presented to the Société de Psychologie on May 17, 1934.

[54] From this perspective, we could also interpret the Maussian body as the first and most immediate aesthetic object and means, capable of accepting aesthetic interventions or modeling, whether they are in the form of completion, ritualizable, or oriented towards the improvement and growth of the bearer.

[55] Each of the body techniques presents a bio-sociological phenomenon. Body education primarily entails training in the regulation and inhibition of disordered body movements, as well as the practice of consistently responding to external stimuli or one's own emotions. Concepts like "cold blood," presence of mind, and dignity for one's body, are elements that unite traditional initiations with the growth processes observed in industrial societies.

[56] E. Thacker, *What Is Biomedia?*, «Configurations», 11, n. 1, 2003, pp. 47-79: 56.

[57] A.R. Meisel, *Intro to Biohacking: How to Be Smarter, Stronger, and Happier*, CreateSpace, South Carolina 2014.

[58] S. Allovio, *Koino-poiesi. Progetti e costruzioni plurali fra i Medje-Mangbetu (Repubblica Democratica del Congo)*, in F. Remotti (a cura di), *Forme di umanità*, Bruno Mondadori, Milano 2002, pp. 129-147: 136.

from the womb. But this idea of birth (or rebirth) of human beings can be subject to significant transformations, to the point that they can be considered to be born several times. If the first birth takes on a physiological character, the second birth takes on a social connotation. A process that is literally invented and constructed within society, and in some ways "fake". The product of such births can change greatly depending on the purpose, giving rise to a process capable of generating something else[59].

Throughout history, this element of construction and fiction has been the subject of a number of rites. These rites – also fictional forms of representation – are capable of "fabricating" the human being, and are commonly referred to as "initiation rites" or "rites of passage"[60]. The concept of rite (of passage) in Van Gennep's traditional vision, implies an artificial subtraction from the context of everyday life, according to a three-phase division method: separation *(separation)*, transition *(marge)*, experience of liminality, and incorporation *(agregation)* reintegration into a new social position within the community[61]. It would be interesting to analyze the concept of liminality in relation to the theme of biohacking, which could perhaps be the subject of another future dissertation. In this context, we will limit ourselves to pointing out how the traditional tripartite conception of rite, according to Van Gennep, appears somewhat elusive in the context of biohacking as a contemporary expression, especially in relation to the instances of separation of individuals – raised in early modernity – from a wider social body and the concomitant process of differentiation that has taken place[62]. A process that is increasingly becoming more embedded in the individual and private sphere[63].

The advent of modernity has enabled men to dispense with such rituals, thanks to the discovery of their biological and scientific manhood[64]. Furthermore, we observe the diminishing significance of initiation rites, the loss of their public recognition, the dissolution of the individual's ontological transformation, and the capacity of these moments of transition to facilitate any form of spiritual and inner regeneration[65].

Following this idea of the second birth and "fabrication" of the human being, can we eventually reconsider the instances arising from *biohacking* with a perspective of reappropriating our own becoming human beings, although in a different form and with broader transformations than those encompassed by rites of passage?

The human condition is related to continuous reinvention. This task is frequently delegated to entities that are not human, such as other animals, machines,

---

[59] F. Remotti, *Fare umanità. I drammi dell'antropo-poiesi*, cit., pp. 35-36.
[60] E. Comba, *Cannibali e uomini-lupo: metamorfosi rituali dall'America indigena all'Europa antica*, Il Segnalibro, Torino 1992.
[61] A. Van Gennep (1909), *The Rites of Passage*, University of Chicago Press, Chicago 1960, pp. 1-15.
[62] N. Elias (1939), *The Civilizing Process: sociogenetic and psychogenetic investigations*, Blackwell, Malden-Oxford-Victoria 2000.
[63] M. Segalen (1998), *Rites et rituels contemporains (3e édition)*, Armand Colin, Malakoff 2017.
[64] M. Eliade, *Birth and Rebirth. Rites and Symbols of Initiation*, Harper & Row, New York 1958.
[65] M. Aime, G. Pietropolli Charmet, *La fatica di diventare grandi. La scomparsa dei riti di passaggio*, Einaudi, Torino 2014.

or even entities that are not physical at all, such as gods or spirits. It is a fallacy to believe that becoming a human being occurs in a neutral, peaceful, and natural manner. Rather, this process is always undertaken in a particular, conflictual, socially negotiated, and culturally conditioned way. The term *anthropo-poiesis* refers to the process by which humans shape their own being, also with modifications involving the body (e.g. tattoos, scarifications, mutilations, implantations, etc.). This process can be constructive or destructive, and encompasses a wide range of behaviors and outcomes.[66].

This results in a space of randomness, which provides the opportunity to experience different forms of humankind and a multiplicity of viable paths and alternative models. The recourse to rituality, as well as to entities (ancestors, gods, nature, etc.), implies an aspect of urgency and a decisive reduction of the multiplicity of models and viable routes, including the aspect of arbitrariness. Spaces and times dedicated to these urgencies and moments of passage, (re)generation and "crisis", are configured as true *spaces of reflection* in which, on the one hand, human beings are "modeled" and, on the other hand, the inevitability, necessity or sacredness of the model adopted is emphasized[67]. This constructed and ritualized space for reflection becomes an opportunity to become aware of the ways in which we "make humankind". The engagement of entities to whom the tasks of "making" humanity are delegated entails an opacification of the process. However, some ritual forms adopted have also the task of unveiling the underlying fictional aspects, showing the transparency of the solutions adopted[68].

The human being elaborates and utilizes forms of culture to explain and shape the self. It is almost inevitable that this process will result in the realization of the arbitrary nature of the forms that are invented, experimented and then put into shape and work, or abandoned. The source of inspiration for such models may be ancestors, deities, other human societies or species, so that a number of reactions are possible: (1) the acceptance of the role played by the model's precariousness, with the unveiling of the fictional aspect on which the social structure is built; or (2) the affirmation of a successful anthropo-poiesis – revealed, found in nature, or conquered thanks to technologies – over the condition of human precariousness, concealing the limits of its own project. The second option (the concealment of such a process) appears particularly intriguing when compared to those religious traditions that have entrusted their divinities with the power to resolve problems associated with anthropo-poietic practice. The concept of being made in *God's image and likeness*, or being the direct fruit of his creative work, responds to this demand. The principal Christian and Western religious formulations, as well as those from analogous traditions, also fall within this category. This similarity between human and God, in Christianity, is further

[66] J.G. Herder (1887-1909), *Ideas for the Philosophy of the History of Mankind*, Princeton University Press, Princeton 2024.
[67] F. Remotti, *Fare umanità. I drammi dell'antropo-poiesi*, cit., pp. 47-49.
[68] E. Comba, *Cannibali e uomini-lupo: metamorfosi rituali dall'America indigena all'Europa antica*, cit.

emphasized following the coming of the Son of God among mankind and – through his sacrifice – the transcending of death.

Furthermore, the philosopher Francis Bacon[69] proposed that modernity would result in the establishment of the reign of humans on earth, a human being who would resemble God in his dominion over nature. The concept of progressive resemblance to God is so extensive that it encompasses the desire to conquer earthly immortality through technology, thereby overcoming the existential precariousness of human nature. By technologically dominating nature, to the point of thinking of defeating or overcoming death on earth, on the specific plane of the corruptible body, humans believe that they have acquired anthropo-poietic powers that are increasingly similar to those once attributed only to God. This is a process whereby a human becomes God himself.

This aspect is also present among many transhumanist movements – referring to the thought of Theillard de Chardin – positing the possibility of *human transhumanization*[70] and the realization of an evolution of the human species guided by the human being (master of his own destiny) thanks to his capacity for invention. A vision that encompasses the preservation of the individual body, or, as will be seen subsequently, the transformation into another (even immaterial) form, with the objective of achieving true earthly immortality. A future humanity so immersed in technology that it will be able to transcend its physical and biological constraints, attaining mastery over its own destiny and the capacity to determine the length of its lifespan, so that the entirety of the universe will be saturated by human intelligence – a machine intelligence (not biological) – determining its fate[71]. The technological development of modernity appears to be accompanied by a profound religious afflatus.

What are the prospects for a human being who is capable of transcending his corporeal essence, capable of determining his own destiny, and who has been raised to the status of a divinity? Is humanity becoming increasingly self-absorbed and attempting to fill its perceived deficiencies, or is it embarking on a path towards a new light?

3. Dimitrj Itskov and 2045 Initiative

Itskov, the founder of Immortality, a corporate joint venture, has embarked on a mission to develop an artificial body. In 2011, Itskov partnered with Timour

[69] F. Bacon (1620), *The New Organon*, L. Jardine, M. Silverthorne (eds.), Cambridge University Press, Cambridge 2000 (2003).
[70] P. Theillard de Chardin (1959), *The Future of Man*, Image, New York-London 2004, pp. 239, 294-295.
[71] R. Kurzweil, *The Singularity Is Near: When Humans Transcend Biology*, Viking, London 2005, pp. 35-38.

Shchoukine, a cognitive neuropsychologist to establish *Rossiia 2045* (Russia 2045)[72], a socio-political movement aimed at promoting radical life extension and urging the Russian government to embrace the construction of artificial bodies as a unifying "universal idea". His primary objective is to transfer the human brain and mind into a series of progressively advanced robotic bodies, initially merging man and machine but ultimately *transcending* physical embodiment altogether. The year 2045 is the date by which the movement's main endeavor, the *Avatar Project*, is to be completed[73].

In 2010, Shchoukine was approached by practitioners of *Advaita Vedanta*, a philosophical school of Hinduism in Nizhnii Novgorod, Russia. Intrigued by phenomena such as lucid dreaming and fascinated by brain activity, they sought Shchoukine's expertise in constructing biofeedback devices to enhance their meditative practices. During his involvement with this community, Shchoukine discovered that Itskov was a regular visitor to the ashram. This connection led them to collaborate on the development of a spiritual and technological protocol aimed at achieving a specific form of immortality, which significantly differs from other contemporary theories that strive towards transcending the body through technological means[74].

The initiative encompasses the achievement of its initial phase, designated as *Body A*, by the year 2020. This phase involves the development of a robotic body that is regulated through a brain-computer interface. Itskov asserts that advancements toward Body A have already been realized, referencing research on cerebral implants that empower individuals with disabilities to control robotic limbs or computer systems solely through cognitive processes. To further enhance their efforts, they engaged the expertise of Aleksandr Kaplan, a professor specializing in neurophysiology and neuro-interfaces at Moscow State University. Kaplan's research has been informed by the pioneering work of Soviet scientist Vladimir Demikhov, who conducted transformative experiments involving dog head transplantation[75]. Building upon these foundational concepts, Kaplan posited the feasibility of isolating and transplanting a head and brain into an independent corporeal entity, termed *Body B*. Subsequently, through discussions with Vitalii Dunin-Barkovskii – another significant contributor to the 2045 initiative and former director of the Neurocybernetics Institute in Rostov on Don – they conceptualized *Body C*, also known as Rebrain, an approach which entails the isolation of a distinct neural substrate[76]. The concluding phase of this project involves the development of a

[72] A. Bernstein, *The future of immortality. Remaking Life and death in contemporary Russia*, Princeton University Press, Princeton 2019.

[73] The main goals of the 2045 Initiative: the creation and realization of a new strategy for the development of humanity which meets global civilization challenges; the creation of optimal conditions promoting the spiritual enlightenment of humanity; and the realization of a new futuristic reality based on 5 principles: high spirituality, high culture, high ethics, high science and high technologies. http://2045.com/ideology.

[74] Ivi, pp. 51-55.

[75] Ivi, p. 80.

[76] Ivi, p. 54.

holographic entity, referred to as *Body D*, which Itskov likely encountered through teachings from his Advaita Vedanta spiritual mentor who introduced the notion of the "rainbow body"[77].

By the year 2045, Itskov envisions a future in which "substance-independent minds" are not merely uploaded onto computer chips but instead integrated into bodies composed of intangible materials. A holographic body, for instance, could possess the ability to traverse solid objects or operate at light speed, whereas a construct made of nano-robots would be capable of assuming multiple forms at will. In light of such changes, he posits that "Humanity, for the first time in its history, will make a fully managed evolutionary transition and eventually become a new species"[78].

While this vision may initially appear implausible, evoking parallels with the Organians from the "Errand of Mercy" episode in Star Trek[79], Itskov's intrigue with the holographic body stemmed from accounts of Tibetan masters attaining profound wisdom and compassion through dedicated practice, often accompanied by reports of rainbows manifesting upon their passing[80]. Notably, one of Itskov's most significant endorsements stems from the Dalai Lama, who advocates for these endeavors within a framework that emphasizes "ethical responsibility" and "reverence for life" which he asserts will ultimately benefit humanity[81]. However, Itskov not only seeks feasibility but also aims for accessibility across socioeconomic strata, aspiring for enduring impacts on society's trajectory.

At this stage, prior to delving into further elucidations concerning the notion of the "rainbow body" as reinterpreted through Itskov's theoretical framework, it is imperative to acknowledge that Project 2045 emerged from an intellectual milieu significantly influenced by bio-hacking principles and its associated speculative imagery, which substantially shaped the epistemological context surrounding its experimental endeavors[82]. Nevertheless, in order to effectively clarify the unique *typological* characteristics of the *transparency* concept as articulated by the 2045 Initiative,

---

[77] C.N. Norbu, *Rainbow Body: The Life and Realization of a Tibetan Yogin, T.U. Tendzin*, North Atlantic Books, Berkeley 2012.

[78] C. Pinchefsky, *Dmitry Itskov Wants To Live Forever. (He Wants You To Live Forever, Too)*, «Forbes», June 18, 2013.

[79] *Star Trek: The Original Series*, Season 1, Episode 26, 1966.

[80] A. Bernstein, *The future of immortality. Remaking Life and death in contemporary Russia*, cit., pp. 53-55.

[81] Remarkably, this project has garnered support from prominent figures in various fields. Scholars such as roboticist Hiroshi Ishiguro, Ray Kurzweil (Google's director of engineering), and Peter Diamandis (chairman of the X-Prize Foundation) have endorsed Itskov's objectives during their presentations at GF2045 lectures.

[82] This episteme deserves recognition for establishing what is being discussed and how it is being discussed, thus constructing a common ground between different cultural systems. Foucault introduces the concept of episteme in his work *The Order of Things* referring to it as a framework defining the range of possibilities within which characteristic knowledge of a specific era is constituted and operates: an inherent set of rules for formation (and any meta-reflections on these rules) that enable discourses to function by discussing various themes along specific lines of coherence". (M. Foucault (1966), *The Order of Things: An archaeology of the human sciences*, Routledge, London-New York 2002 (2005), pp. 23-25).

it is essential to present a succinct overview of the intellectual lineage that informed such conceptualization.

In this scholarly context, the theoretical framework established by Itskov will be utilized to examine the degree to which a particular hypothesis regarding open access science, the re-codification of biological data within the digital environment, and the perspective that regards technology as subordinate to spiritual development, has contributed to the emergence of an epistemological paradigm that integrates scientific investigation with spiritual and socio-political considerations. In fact, as Itskov articulates, this initiative transcends mere veneration for physical immortality; it represents a conceptual framework designed to endure across temporal boundaries, aspiring to embody an eschatological archetype of humanity's future.

As such, within the realm of evolving epistemological paradigms and conceptual reconstructions geared towards bio-hacking the body and mapping its inner functions, "old" academic discussions regarding speculative futures have surfaced. This resurgence is shaped by unique ideological perspectives, which evolve based on particular "keys of accessibility" that originate from a specific cultural setting[83]. Remarkably, as various scientific-oriented practices seek to recalibrate the past toward constructing an improved society, Umberto Eco's concept of "possible worlds" becomes relevant, in order to explore how biohacking as a mainstream framework has been translated into one of the "internal languages" of a given *culture*[84]:

> The possible worlds as epistemic constructs are real in that they are embedded, not just syntactically, in the real world that produces them. [...] These "possibles" are not parallel; they are proportionally nested within each other, and each one participates to some extent in the reality of its container[85].

According to Eco, the ability to imagine possible worlds relies on the network of *shared encyclopedic knowledge* within a given community: this network enables the generation of inferences and the differentiation between possible worlds and the "real world", based on a plausible projective construction that is intricately linked to the logical articulation of the real world. These logical-semiotic systems are designed to narrate and propose the coordinates of a future possibility within a plausible scenario, whether real or abstract, always drawing on epistemic "keys of accessibility", which encompass political, social, scientific, intellectual and other subsystems that operate through ongoing processes of translation, based on dynamics of compatibility and incompatibility with a specific cultural system.

In this outlook, the significant rise of post-soviet Sovietology in the 1990s played a pivotal role in stimulating a reassessment of the Russian historical and intellectual heritage in light of the political and identity challenges of the post-Cold

---

[83] U. Eco, *Lector in fabula*, Bompiani, Milano 1979, pp. 154-173.
[84] J. Lotman, *O semiosfere*, in «Trudy po Znakovym Sistemam», 17, 1975.
[85] U. Eco, *Sugli specchi e altri saggi. Il segno, la rappresentazione, l'illusione, l'immagine*, Bompiani, Milano 1985, pp.209-210.

War era[86], potentially justifying an entire *economy of anticipation*[87]. This intellectual heritage has experienced considerable resurgence, facilitated by the dissemination of previously unpublished materials related to Russian philosophy of religion and the inaugural publications of Nikolaj Fyodorov's groundbreaking theories by Svetlana Semyonova, alongside those of other Soviet thinkers[88]. As such, it is evident that initiatives aimed at maximizing individual effort in response to societal changes align with a historical continuum that stretches from Filokalia, Russia's late counterpart to "The Imitation of Christ", to Nikolaj Chernyshevsky's influential novel, "What Is to Be Done?"[89].

In this outlook, the advent of Project 2045, which pertains to the concept of a holographic body, is not an arbitrary occurrence; rather, it warrants analysis through the lens of the Bio-Cosmism movement and the *anthropo-technical* paradigm[90] predominantly articulated by Soviet Communism[91]. Indeed those epistemological categories were, to some extent, deconstructed after 1991, while simultaneously giving rise to what may be referred to as "metaphysical compensation chambers", environments where postmodern individuals may draw upon a renewed reservoir of cosmic belonging. Within these "chambers" new ideological supplements are emerging as a result of an exchange that involves the replacement of Judeo-Christian religious frameworks with what is termed "Western Buddhism". In the sense that its purported influence is characterized by its association with a cultural ambiguity, a

[86] I. Perusko, *L'uomo sovietico sbarcò davvero sulla luna? Le trasgressioni di Victor Perelevin*, in L. Piccolo (a cura di), *Violazioni. Letteratura, cultura e società in Russia dal crollo dell'URSS ai nostri giorni*, Roma Tre Press, Roma 2017, pp. 83-97.

[87] In this outlook, the dimension of possibility possesses an encyclopedic nature, and should be understood within a rhizomatic constellation of possible worlds, each endowed with a topological property and a certain degree of operability. Notably, Anya Bernstein outlines that "Producing affective states of alternating hope and fear, such an ethos of preparedness requires action now to secure certain futures and avoid others. Anticipation produces political narratives of preemptive wars and dictates how we manage our financial futures" (A. Bernstein, *The future of immortality. Remaking Life and death in contemporary Russia*, cit., p. 11).

[88] M. Laruelle, *Totalitarian Utopia, The Occult, And Technological Modernity in Russia: The Intellectual Experience of Cosmism* in B. Menzel, M. Hagemeister, B. Glatzer Rosenthal (eds.), *The New Age of Russia. Occult and Esoteric Dimensions*, Kubon & Sagner, Munich 2012, pp. 238-258; see also D. Monticelli, *Thinking the new after the fall of the Berlin Wall: Juri Lotman's dialogism of history*, in «Rethinking History», 24, n. 2, 2020, pp. 184-208.

[89] The protagonist, Rahmetov, epitomizes a modern ascetic figure who subjects himself to extreme physical challenges, including sleeping on a bed of nails and adhering to a stringent dietary regimen.

[90] In 1926, the third volume of the Great Soviet Encyclopedia included the entry "Anthropotechnics". This term is defined as "an applied branch of biology that aims to enhance both the physical and spiritual qualities of humans using methods similar to those employed in zoology for the improvement and breeding of new domestic animal breeds" (in B. Groys, M. Hagemeister (eds.), with collaboration from A. von der Heiden, *Die Neue Menschheit. Biopolitische Utopien in Russland zu Beginn des 20. Jahrhunderts*, Suhrkamp, Frankfurt am Main, 2005, p. 54).

[91] A. Sloterdijk (2009), *You Must Change Your Life: On Anthropotechnics*, Polity, Cambridge 2013, pp. 391-392.

phenomenon that aligns with Zizek's well-documented critique[92], and which can be linked, to some degree, to a certain spiritual priority (implied as a gnoseological category) attributed to the East[93].

The Anarcho-biocosmists represented the most politically active segment within the ideological framework of the anarcho-futurist movement. Emerging around 1920 in Moscow, with significant activities also noted in Petrograd, this group explicitly endeavored to realize to initiate a program for Lenin's resurrection, culminating in 1926 with Alexander Bogdanov's establishment of the Moscow Transfusion Institute (*Institut perelivanija krovi*); an ostensibly nominal institution aimed at advancing a revolutionary medical theory known as *physiological collectivism*. Within this framework, Marxism was regarded as an irrefutable doctrine capable of actualizing immortality through scientific advancements while social revolution was perceived as an initial step toward fundamentally reshaping *reality itself*. Svjatogor, the author of the Biocosmism Manifesto (1922)[94], subjected the classical doctrine of Russian anarchism to a fundamental revision, integrating it with the necessity of conceptualizing immortality as a political objective: an ultimate principle to be pursued in order to guarantee the new socialist individual an inalienable right to immortality. Svyatogor viewed immortality as both the goal and prerequisite for a future communist society, positing that true social solidarity could only be established within a society of immortals: "death separates people; private property cannot be eliminated until time is collectivized" (1922).

In this perspective, total biopower would inherently necessitate the collectivization not only of space but also of time, thereby eliminating conflicts between the individual and society. Notably, Dmitry Shlapentokh observed that:

> The Bolshevik Revolution [...] implied not only social changes (e.g. absolute social harmony), but metaphysical-transcendent changes as well. This feeling of eschatological excitement was in various measures an element of all great revolutionary transformations [...] Toppling the Russian monarchy would lead not only to political but also to *cosmic changes*. [...] The same eschatological expectations arose during the Bolshevik Revolution and one might say that in the mind of many Russian intellectuals the Bolshevik Revolution had two "layers"; so to speak; one was social, the other millenarian[95].

In such a light, the anthropo-technical aspect of the 1917 Revolution commenced with Russian intellectuals's engagement with this phenomenon, coinciding with a transformation in the Western conception of political revolution

[92] S. Zizek, *Self-Deceptions. On being tolerant and smug*, «Die Gazette», August 27, 2001.
[93] A. Sloterdijk, *You Must Change Your Life: On Anthropotechnics*, cit.
[94] A. Svjatogor (1922), *The Biocosmist Manifesto*, in B. Groys (ed.), *Russian Cosmism*, MIT Press, New York 2018.
[95] D. Shlapentokh, *Bolshevism as a Fedorovian regime*, in «Cahiers du monde russe : Russie, Empire russe, Union soviétique, États indépendants», 37, n. 4, 1996, pp. 429-465.

that fundamentally depoliticized it, thereby rendering it a radically "meta-ontological experiment"[96].

In this perspective, as Sloterdijk posits, one might even assert that politics was infiltrated through the lens of *orientalization*. In this context, the term "East" denotes a propensity for the predominance of the spiritual dimension, a factor that persisted in exerting influence following 1991. It is posited that the emergence of the *hologram body* project would not have been possible without this underlying spiritual conversion, which keeps mirroring an extensive process of external transformations[97].

These elucidations are crucial for contextualizing this phenomenon while simultaneously offering a valuable framework for comprehending how cultural expressions became subjugated under the symbolic authoritarianism inherent in Soviet political ideology, while intricate interactions intertwining communism with religious elements began to permeate into the mythological fabric of early twentieth-century Russian society.

Indeed, the exploration of the relationship between biohacking and Leninism/Bolshevism is fundamentally contentious and poses significant challenges when attempting to offer interpretative insights that can reconcile Itskov's trajectory with Soviet political ambitions in the aftermath of the pivotal events of 1991. However, as will be clear in the next paragraph, two fundamentally opposing conclusions regarding biological futures emerge, each presenting itself as a grand narrative with significant geopolitical ramifications. Both narratives share a symmetrical focus on the necessity of transcending physical limitations while simultaneously converging on the imperative for humanity to act collectively in a manner that aligns with societal advancement. At the core of both discourses lies a recognition and aspiration to surpass current constraints on life, envisioned as a technological construct, while also invoking a "return to nature," suggesting an endpoint which is notably close to that envisioned by the novel paradigm of inquiry established by bio-hacking.

Based on these premises, we argue that biohacking establishes an epistemological foundation where bodily functions and chemical-molecular processes are situated within a broader interconnected paradigm in which all material and psychic aspects are interrelated, suggesting a reevaluation of humanity's place within the cosmos, and prompting a reflection on the physical and chemical dynamics within the human body as part of cosmic processes. This approach introduces innovative perspectives that might shape contemporary investigations on the *auto-poietic* potentialities inherent in the human body and matter itself, encompassing behaviors, experiences, physiological and chemical processes that may challenge traditional

[96] A. Sloterdijk, *You Must Change Your Life: On Anthropotechnics*, cit.

[97] The term "conversion" denotes a spiritual recalibration of existence, while "revolution" involves reimagining reality from a foundational perspective. Within the revolutionary crucible, matter previously fixed in various qualities is transformed into potent potentiality, harnessed for innovative endeavors (A. Sloterdijk (2009), *You Must Change Your Life: On Anthropotechnics*, Polity, Cambridge 2013, pp. 390-392).

understandings of human functionality and humanity's position within the cosmic order.

These perspectives prompted contemplation on resurrection and spiritual advancement, central themes in Itskov's belief, previously explored by the eminent philosopher Nikolaj Fydorov, whom Itskov undoubtedly is familiar with[98]. In this outlook, while certainly not interested in erasing the peculiarities of different religious traditions or in turning Tibetan Buddhism and Fyodov's unique interpretation of Christianity into masks of the technological, Itskov makes a number of daring claims, suggesting that there are peak mystical experiences in those traditions that share intriguing similarities with Konstantin Tsiolkovskij theory of *pan-psychism*, even if it's certainly not possible to affirm their identical character[99].

Remarkably, he describes *Body D* as a "body of light" and points Tsiolkovsky, regarded as the father of Russian space science, as part of his intellectual lineage. Tsiolkovskij discussed replacing biological bodies with others composed of pure energy, referring to it as *radiant mankind* (*luchistoe chelovechestvo*), putting an equal or even superior emphasis on studying the spiritual characteristics of matter over pursuing mere physical enhancement through technological means.

At this juncture, it becomes imperative to delve further into Tsiolkovskij speculations, who provides an interesting perspective to discuss specific idiosyncratic reactions to the concept of *autopoiesis thought transparency* established by the biohacking framework which we previously elucidated. As evidenced in the subsequent passage, the scientific ambitions of pan-psychism are currently undergoing a reevaluation by Itskov; this reassessment aligns with the *translation* of biohacking theories away from conventional frameworks towards new epistemological paradigms that prioritize matter as a pathway to directly engage with immortality itself. Crucially, Tsiolkovsky posited that the pinnacle of human intellect would not be achieved through biological methods but rather by tapping into electromagnetic fields. He introduced the concept of *cosmic matter* evolving within the brains of higher organisms into an irreversible state of radiant energy, imbued with a unique cosmic consciousness that permeates space. Once achieving such a state, humanity would bask in *eternal existence* and *salvation*.

This notion, while not novel, aligns with Einstein's mass-energy equivalence formula. However, Einstein's formula pertains to matter as it currently exists and is inherently reversible, as its asymmetry does not stem from the formula itself[100]. Tsiolkovskij hypothesizes a type of matter whose transformation into energy or radiation is one-sided and irreversible; this irreversible transformation will

---

98 A. Bernstein, *The future of immortality. Remaking Life and death in contemporary Russia*, cit., p. 58.

99 K. Tsiolkovskij (1925), *Monizm svellenoj*, in S.G. Semënova, A.G. Gačeva (eds.), *Russkij kosmizm*, Nauka, Moskva, 1993.

100 The reference to Einstein pertains to his equation $E=mc^2$, signifying that matter can be converted into energy (in the form of radiation) and vice versa. A prime example of this concept is nuclear reactions: during fission, a heavy atom splits into two lighter ones. The final mass (sum of both lighter atoms) is less than that of the original atom due to the missing mass being transformed into energy. Nevertheless, this equation does not dictate a specific directionality.

characterize the terminal phase of the cosmos, at which point a directional arrow might be added to Einstein's equation. This subtle addition would convey profound insights to future super-humans, as these advanced beings will not require matter but *energy* and their cosmic purpose will have been fundamentally resolved:

> Matter is one existing thing, regardless of its movement or displacement in space. Deeper knowledge of the structure of matter is not yet available to us. But someday there will come a turning point when mankind will approach this "esoteric" knowledge. Then it will come close to the question: why? But for this to happen, billions of years of the space age must pass [...] Rockets, the second beginning of thermodynamics, are the business of our day, but at night we live a different life if we ask ourselves that *damned question*. (i.e. Why?) And there is another important point: the question about the *randomness* or non-duality of matter was raised by the ancient sages. They taught that there is a spiritual world where "there are neither tears nor sighs, but endless life". The idea of the "randomness" of matter came to my mind after I learned that the average mass density of matter in the galaxy [...] occupies a vanishingly small volume in comparison with the volume of "empty" space[101]. Thinking further [...] the smallness of matter speaks of its randomness or temporality, because everything random or temporary has a small or vanishingly small value. [...] What is the implication of this? [...] a random quantity can disappear someday: either its lifetime will end, or, speaking the language of physics, it will be transformed into radiant energy. Generally speaking, small quantities and values are absorbed without residue by large ones, and this happens the sooner the greater the difference between large and small values, and here we have a colossal difference equal to 10^{33}. [...] It's a kind of monism. A monism. But don't think of it as entropy! God forbid, entropy will not exist in that world either, as it does not exist in this one for open systems[102].
>
> I am a pure materialist. I acknowledge nothing but matter. I see only mechanics at work in physics, chemistry, and biology. The entire cosmos is merely an endless, complex machine. Its complexity is so great that it borders on the arbitrary, the unexpected, and the accidental [...] Various parts differ only in the degree of their sensitivity, which varies continuously from zero to an indefinitely large magnitude in supreme beings, that is, in beings more perfect than people. [...] Everything is continuous; everything is one. The degree of sensitivity depends on the combination of matter. [...] In terms of mathematics, the entire universe is alive, but the power of its sensitivity is manifested in all its brilliance only among higher animals. All atoms of matter feel in keeping with the environment. Finding itself in highly organized beings, atoms live their lives and feel their pleasure and pain. If they find themselves in the inorganic world, they sleep, as it were, immersed in a deep state of unconsciousness, in nothingness [...][103].

[101] The Universe is predominantly characterized by vast empty spaces. While a planet may be highly dense, the void between planets is essentially prevalent. This principle extends to stars within a galaxy and galaxies within the universe. Ultimately, when averaged out, the density of the Universe (i.e., the number of atoms occupying a certain volume) is exceedingly low.

[102] The text is an interview between Alexander Chizhevsky and Konstantin Tsiolkovsky (1932) and is given according to the first publication in the journal «Chemistry and Life» (n. 1, 1977). https://www.tsiolkovsky.org/en/the-cosmic-philosophy/the-theory-of-cosmic-eras.

[103] K. Tsiolkovskij, *Panpsychism, or Everything Feels*, in B. Groys, *Russian Cosmism*, MIT Press, New York, 2018, pp.133-136.

> [...] Since all material, under favorable conditions, can always go into an organic state, theoretically we can say that inorganic matter is potentially alive.

The extracts regarding pan-psychism of matter suggest an implicit emergence of a new concept of transparency, wherein the notion entails the capacity to circumvent intermediaries, such as the body and its material substance, and directly engage with the energy of the universe. This direct engagement is facilitated by an irreversible interplay between energy and matter, elucidated by Einstein's equation E=mc2. It is therefore presupposed that "corporeal data" can hold significance autonomously from the intricate networks and relationships in which they are currently embedded, as well as apart from the scientific domain through which they are *constructed*. Intentionally or unintentionally, the proper semiotic dimension of the body is therefore obscured. In concrete terms, this implies that information concerning the body and its internal mechanisms, formulated through scientific methodologies and explained by the principles of physics, acquires intrinsic autonomy due to a sequence of interconnected procedures and dialogues that permit its placement within the same epistemological framework.

In this regard, these criteria of accessibility and transparency, which aim to bypass intermediaries and present us directly with body data and subsequently with physical mechanisms imply that in a distant future matter will become neutral itself, lacking its semiotic function of mediation. We could then suppose that this type of interaction has an influence on how to interpret the content of the data itself, thus framing a horizon of expectation regarding the kind of possible-world being portrayed.

In other words, this framework for interpretation may imply that if we believe in the reality of the world depicted in a data, nature as "thinking matter" should also obey the mode of interaction that led to it (physics).

In this context, the 2045 movement undertakes coordinated endeavors for societal reformation on a political level by aiming to democratize biology and establish a unified platform encompassing means of production, scientific research, and data collection. Moreover, it distinctly demonstrates a tendency towards revitalizing traditional scientific and emerging religious paradigms in order to formulate a novel form of (supposedly) neutral ideology, but which actually functions as a perfect ideological supplement.

4. Conclusions

As elucidated by our reflections, the emergence of the epistemological paradigm of bio-hacking demands a revision of *interpretive parameters* adopted in scientific discourse analysis up to this point. It suggests the adoption of approaches that, given the complex nature of the subject matter, invariably encroach upon theoretical domains from diverse backgrounds: fields of study that are partially related but sometimes distantly connected will find themselves interconnected transversally. However, as we

have consistently emphasized, it is precisely during certain historical periods that specific modes of inquiry surrounding the concept of transparency become infused with ideological, political, and cultural values capable of expressing themselves even through specific cultural representations.
In such a light, it has been observed that the "construction" and "fabrication" of the human being is of great importance in the context of human cultures. Each culture assumes a specific anthropological, rather than symbolic, model for itself. A process that affects both the social and the biological spheres. Some aspects of the process, eventually *anthropo-poietic*, and the chosen models of "making humanity" also involve some dramatic aspects, such as modifications and interventions on the body, which may be disabling (e.g. amputations) or additional (e.g. implants). These take the idea of being human to a different level and demonstrates that the process of becoming a human being is not a neutral and uncontracted one, even in terms of one's relationship with one's social context.

The diverse and idiosyncratic ways in which each culture conceptualizes humanity, if on the one hand they demonstrate its provisional and fictional nature, on the other hand they can lead to the assertion of its success, concealing its limitations and fragility.

The second assumption of the concept of *anthropo-poiesis*, for example, can lead to the idea of overcoming precariousness and to the assumption of a form of "domination" and control over time and space. This enables the transcendence of the idea of death. The same concept can be informed by contributions from both religious and technological sources. One potential outcome of this is the possibility for humanity to sublimate its corporeality in favor of a projection towards a horizon of immortality.

A paradigmatic illustration of this phenomenon is the emergence of novel epistemological frameworks that facilitate an increasingly "creative practice" in engaging with scientific inquiry, distinguishing themselves both independently and eccentrically from traditional institutional forms. The narratives, beliefs, and practices linked to the scientific exploration paradigm established by bio-hacking engender a manifestation of disquieting speculative imagery. This imagery arises from the re-codification of biological data within the digital environment, which at times may be perceived as "spiritualized."

Nevertheless, it is essential to clarify that despite the diversity of narratives involved, what remains constant in these speculations regarding the future of human nature is the fundamental structure through which varied experiences are categorized. This structural framework subsequently informs cultural trajectories and ideological impulses characteristic of those experiences.
In this perspective, the examination of the 2045 Initiative, along with a deliberately constrained yet adequately representative range of discussions, has concurrently unveiled two fundamental points: the extensive variety of meanings and potential discourses through which the concept of transparency is articulated and the significance of the typological viewpoint that encapsulates them. Furthermore, it

emphasizes the potential for categorizing this array within an emerging epistemological paradigm of speculative inquiry, as the intended interpretation of a particular corporeal representation of immortality serves as an effective instrument for evaluating the anthropological identity of a culture in relation to its historical context. Consequently, it is imperative to investigate transparency as an elusive semantic spectrum whose shifting significances, manifested in texts and discourses, reveal deeper insights into the culture articulating them at any given historical moment. This naturally establishes a foundation for contemplation that warrants further exploration through widely disseminated texts. It presents a compelling point of departure for examining the evolution of highly intricate and nuanced scientific classifications alongside cultural sensibilities.

The Moral Value of Transparency in the Use of Performance Enhancing Drugs. The Case of Bodybuilding[a]

Matteo Cresti[*]

Abstract
L'articolo ha l'obiettivo di sostenere il valore morale positivo della trasparenza riguardo all'assunzione di Performance Enhancing Drugs (PED) nel bodybuilding. Per prima cosa darò una definizione di trasparenza adeguata all'ambito sportivo. In secondo luogo descriverò l'uso di PED nel bodybuilding, in particolare di steroidi anabolizzanti, mostrando come negli ultimi anni si possa registrare un fenomeno di rivelazione dell'uso di PED. Proporrò poi il mio argomento in difesa della trasparenza sull'assunzione di PED basato su considerazioni consequenzialiste. La tesi è che i bodybuilder che rivelano di fare uso di PED stiano compiendo un'azione moralmente positiva, in quanto consentono a chi si ispira a loro come modelli di ricalibrare le proprie aspettative e di fare scelte più informate. Infine risponderò all'obiezione che questa pratica possa incentivare l'uso di PED.

Parole chiave: bodybuilding, doping, steroidi, social network, riduzione del danno.

Abstract
The paper aims to support the positive moral value of transparency regarding the intake of Performance Enhancing Drugs (PEDs) in bodybuilding. First, I will adequately define transparency for the sports sector. Secondly, I will describe the use of PEDs in bodybuilding, in particular of anabolic steroids, showing how, in recent years, there has been a phenomenon of disclosure of the use of PEDs. I will then propose my argument in defense of transparency on PEDs intake based on consequentialist considerations. The thesis is that bodybuilders who reveal that they use PEDs are doing a morally positive action, as they allow those who look up to them as role models to recalibrate their expectations and make more informed choices. Finally, I will respond to the objection that this practice could encourage the

[a] Saggio ricevuto in data 31/05/2024 e pubblicato in data 22/01/2025.
[*] Ricercatore, Università di Torino, email: matteo.cresti@unito.it.

use of PEDs.

Keywords: bodybuilding, doping, steroids, social network, harm reduction.

1. Introduction

In this article, I will analyze some ethical aspects of the use of Performance Enhancing Drugs (PEDs), that is, those drugs that are commonly called doping and which are therefore found in the list of prohibited substances drawn up by the World Anti-Doping Agency (WADA). In particular, I will focus on the moral aspects of transparency in using PEDs and the positive effects such transparency can have. The argument that I propose here is that being transparent about the use of PEDs has a positive moral and pedagogical value. I will not talk about sports in general, but I will focus in particular on bodybuilding. I will do this for two reasons. The first is that in this sport, the use of PEDs is pervasive and evident. Secondly, in recent years, there has been a powerful movement of disclosure of information regarding the use of PEDs.

First, I will define "transparency" in relation to sports. I will distinguish between two forms of transparency, namely, the disclosure of relevant information. The first sense is complete transparency, in which all information about the training program is provided, and the second is partial transparency, in which only general information is provided. I will explain that transparency, in the first sense, cannot be contemplated within sports.

Secondly, I will try to provide a sufficiently detailed picture of the use of PEDs in bodybuilding. I will show how at least three culturally distinct moments can be identified regarding transparency in the use of PEDs. When PEDs entered the sport, it was common to talk about them and promote them. Later, bodybuilders denied PEDs use to protect the sport's reputation. Today, however, we are witnessing a moment in which people go back to admitting to using PEDs and also promote them through social networks.

Thirdly, I will present my argument in favor of transparency in the use of PEDs. In essence, I ground my argument on the consideration that professional bodybuilders or fitness influencers are role models for aspiring bodybuilders or amateurs. Ignoring that a particular physique has been achieved through the use of PEDs can generate frustration and discomfort. On the other hand, transparency can help to recalibrate one's expectations or make autonomous choices.

Finally, I will respond to some objections regarding the use of PEDs being encouraged in this way. This can be the first step towards transitioning to a different model, from the current total ban to harm reduction.

2. Transparency in Sports

Transparency is not a recognized value in sports. At least at an official level. WADA, in its Code, makes a list of sports values that does not include it:

> The spirit of sport is the celebration of the human spirit, body and mind. It is the essence of Olympism and is reflected in the values we find in and through sport, including:
> - Health
> - Ethics, fair play and honesty
> - Athletes' rights as set forth in the Code
> - Excellence in performance
> - Character and Education
> - Fun and joy
> - Teamwork
> - Dedication and commitment
> - Respect for rules and laws
> - Respect for self and other Participants
> - Courage
> - Community and solidarity
>
> The spirit of sport is expressed in how we play true. [1]

It has also been argued that transparency cannot be a value within sports because athletes and their clubs have every interest in keeping secret all their training methods, the technologies they use, and the results of their research.[2] This applies to both legal and illegal methods and activities. Suppose a specific training program or a permitted technology can produce an advantage for the athletes and lead them to win. In that case, athletes, coaches, and sports societies would want these technologies and methods to remain secret or protected by a patent. On the other hand, if the methods and substances are illegal (or could soon become so because they comply with the criteria established by WADA to be included in the list of prohibited substances), there will be a greater interest in keeping everything confidential: not only so that the athletes retain this advantage, but also so that they are not disqualified or punished. Therefore, the first thing to do is to understand the meaning of the value of "transparency" and whether it is compatible with sports practice. Some researchers belonging to the International Network for Doping Research (INDR) instead claim that transparency is a positive moral value.[3] They show a critical attitude towards the

---

[1] WADA, *World Anti-Doping Code*, 2021, https://www.wada-ama.org/sites/default/files/resources/files/2021_wada_code.pdf, p. 13.

[2] G. S. Bullock et al. *The Trade Secret Taboo: Open Science Methods are Required to Improve Prediction Models in Sports Medicine and Performance*, in «Sports Medicine», LIII, 2023, pp. 1841–1849. S. Holm, *Doping under medical control – conceptually possible but impossible in the world of professional sports?*, in «Sport, Ethics and Philosophy» . I, n. 2., 2007, pp. 135–145.

[3] J. Mazanov, J. Connor. *Rethinking the management of drugs in sport*, in «International Journal of Sport Policy and Politics», II, n. 1, 2010, pp. 49-63; A. Petróczi, et al. *'Clean athlete status' cannot be certified: Calling for caution, evidence and transparency in 'alternative' anti-doping systems*, in «International Journal of Drug Policy», XCIII, 2021, 103030.

current anti-doping policies of WADA, although they "welcome programmes that encourage athletes to openly declare their commitment to clean sport"[4]. However, the central point regarding transparency in these authors does not concern athletes but rather anti-doping programs: they "call for transparent and rigorous scientific scrutiny via peer-review for alternative anti-doping systems"[5]. In this perspective, transparency in the construction of anti-doping programs is obtained through ethically approved studies and protocols, results published in scientific peer-review journals, and identified authorships that take responsibility for the statements they make.

As regards the ethics of sport, the term "visibility" has been used, which refers to various phenomena: on the one hand, the term "visibility" indicates how much a category manages to get noticed; in this case, we are talking about the visibility of disabled athletes, female athletes or other categories.[6] On the other hand, visibility has meant "the level of information that individuals have access to in regard to the kind of drugs or pharmaceuticals they are being administered, or the regimes or surgeries they undergo; and the level of transparency, and thereby accountability, that characterizes the professional sport context"[7]. Therefore, I use the term "transparency" precisely in this last sense: as the disclosure of information and results, which also leads to accountability, not only in legal terms but in moral terms to the consequences of one's training paths and the sports results obtained.

Transparency in sports can be of two types: complete and partial. Complete transparency means the complete disclosure of all training plans and technologies used. As said before, it is impossible, under penalty of losing the competitive advantage, to disclose all the information. Doing so would be problematic not only from a sporting point of view but also for the athletes' privacy, who would be forced to disclose a good part of their private life[8]. However, greater transparency and publicity in research results and training protocols could advance sports medicine.

The second sense of transparency is partial transparency. In this case, for example, one declares to be training in a certain way without specifying the methods in which the training takes place, or to be taking drugs but without indicating which ones, or to use certain technologies without specifying the technical details. For

---

[4] A. Petróczi, et al. *'Clean athlete status' cannot be certified: Calling for caution, evidence and transparency in 'alternative' anti-doping systems*, cit., p. 7.

[5] *Ibidem.*

[6] See K. P. DePauw, *The (In)Visibility of DisAbility: Cultural Contexts and "Sporting Bodies",* in «Quest», XLIX, 1997, pp. 416-430; P. Serra et al. *The (in)visibility of gender knowledge in the Physical Activity and Sport Science degree in Spain*, in «Sport, Education and Society», XXIII, n. 4, 2016, pp. 324-338; H, Gammelsæter, *Media visibility and place reputation: does sport make a difference?*, in «Journal of Place Management and Development», X, n. 3, 2017, pp. 288-298.

[7] S. Camporesi, M. J. McNamee, *Performance enhancement, elite athletes and anti doping governance: comparing human guinea pigs in pharmaceutical research and professional sports*, in «Philosophy, Ethics and Humanities in Medicine», IX, n. 4, 2014, p. 1-9: 2.

[8] L. Cox, A. Bloodworth, M. J. McNamee, *Olympic Doping, Transparency, and the Therapeutic Exempion Process*, in «Diagoras. International Academic Journal on Olympic Studies», I, n. 1, 2017, pp. 55-74.

example, let us take the case of a bodybuilder who publishes his workout on a social network. Suppose the bodybuilder discloses all the types of exercise, number of sets and repetitions, weight load, the type and grams of supplements taken, the calories and macronutrients consumed, and the dosage of any other PEDs. In that case, he will be exercising a form of complete transparency. If, on the other hand, he limits himself to the type of exercises, supplements taken, and the declaration of being a "natural" or "enhanced" athlete, then he will be implementing partial transparency. Partial transparency also responds to the objections raised against complete transparency. First, it does not create a competitive disadvantage: hidden details are essential to obtain similar results. Secondly, the violation of privacy is severely limited. For example, in the case of taking a drug for therapeutic purposes, one does not necessarily have to say what type of drug one is taking and the reason why one is taking it. Still, it will be enough to say that one is being treated for a pathology. This also violates the athlete's privacy, but much less. At the same time, this type of transparency is not very helpful for the progress of scientific research.

Therefore, I will adopt this second sense of transparency: making visible the strategies adopted to improve one's performance without disclosing the technical details of their implementation. The paper aims to show how adopting this policy on transparency is a positive moral practice. I will not show it in general but about a specific aspect: the use of PEDs in bodybuilding.

3. The Use of PEDs in Bodybuilding

The use of PEDs is widespread. This statement, which seems familiar among those involved in sports or practicing it at a competitive level, is challenging to support with statistical data. Studies that try to give numbers are few. A study published in 2017, but whose data collection dates back to 2011, showed that almost half of the athletes competing at the World Athletics Championship in Daegu (South Korea) had used illegal substances, and over half of the athletes competing at the 12th Quadrennial Pan-Arab Games (PAG) in Doha (Qatar) had used them.[9] This difficulty in quantitatively measuring the phenomenon is due not only to the reluctance of athletes, coaches, and sports clubs but also to the WADA rules themselves, which, for example, have made it extremely difficult to publish the results of these studies.[10]

Bodybuilding, in many ways, is a sport where doping is much more frequent and accepted. Even just from the point of view of common sense, the muscles brought by certain athletes on prestigious stages, such as that of Mr. Olympia, cannot

9 R. Ulrich et al. *Doping in Two Elite Athletics Competitions Assessed by Randomized-Response Surveys* in «Sports Medicine», XLVIII, 2018, pp. 211-219.

10 R. Ulrich, *Letter to Right Honorable Jesse Norman, Member of Parliament*, 2016. http://www.parliament.uk/documents/commons-committees/culture-media-and-sport/Correspondence/Letter-from-University-of-Tubingen-regarding-blood-doping-11-January-2016.pdf (last accessed 31 may 2024. See also: R. Pielke, *Assessing Doping Prevalence is Possible. So What Are We Waiting For?* in «Sports Medicine», XLVIII, 2018, pp. 207-209.

be imagined without the use of PEDs. However, many anthropological and sociological testimonies have documented the use of PEDs in bodybuilding.[11]

When we talk about PEDs in bodybuilding, we immediately think of the use of steroids, such as testosterone, nandrolone, and trenbolone. However, these are just some of the substances used. For example, drugs such as insulin are also used to improve carbohydrate management, or diuretics to increase muscle definition in the run-up to competitions. Not all PEDs have a good reputation. Even within the bodybuilding community, some substances are seen as worse than others. For example, the use of synthol is perceived very differently from that of steroids. Synthol is a drug that mainly contains oils, which is injected subcutaneously to increase muscle size. A minimal use can rebalance some imperfections, but a massive use makes the muscles "bloated" or "watery", only increasing their size in a disharmonious and unnatural way, therefore not favoring the athlete's performance in any bodybuilding competition. It is interesting to note that within the bodybuilding community itself, there is a negative perception of this PED. It is perceived both as something that makes it *fake* (the muscles have not grown because of training, but because they have been "inflated" with oil) and something that makes it aesthetically unnatural (the muscle is not "hard", "swollen" or "built").[12] Even among PED users, not all substances are created equal, and not all have the same reputation. In this article, I will focus specifically on steroid use, as its use is the most widespread and evident in bodybuilding and the group of substances that is undergoing the fastest shift in cultural perception.

A line of change in attitude has been drawn regarding the use of PEDs in bodybuilding. As in many other historical phenomena, different phases can be identified, characterized by different cultural attitudes. Bodybuilding is no exception, even if it has a relatively recent history. Regarding the use of PEDs, we can identify three distinct phases.[13] Bodybuilding was born in the late 19th century and from the

11 For example, A. V. Christiansen, *Gym Culture, Identity and Performance-Enhancing Drugs Tracing a Typology of Steroid Use*, Routledge, London, 2020; J. Andreasson, T. Johansson, *Bodybuilding and fitness doping in transition. Historical transformations and contemporary challenges*, in «Social Sciences», VIII, n. 80, 2019, p. 1-14; D. Liokaftos, *A Genealogy of male bodybuilding*, Routledge, London, 2018; L. F. Monaghan, *Accounting for Illicit Steroid Use: Bodybuilders' Justifications*, in A Locks, N. Richardson (eds,.) *Critical readings in bodybuilding*, Routledge, London, 2012, pp. 77-90; F. Monaghan, *Bodybuilding, drugs and risk*, Routledge, London, 2001.

12 M. Rutcofsky, *7 Alleged "Synthol freaks" who went too far. These "bodybuilders" decided to take the worst shortcut imaginable*, in «Muscles and Fitness», https://www.muscleandfitness.com/features/newsstand/5-synthol-freaks-who-went-way-too-far/ (last access 31 May 2024); N. Albers, *Synthol: freak effects and abuse*, in «FitSociety» https://www.fitsociety.io/bodybuilding/synthol-freak-effects-and-abuse/ (last access 31 May 2024); M. Šarčev, *How Synthol Almost Killed Milos Sarcev* in «Generation Iron Fitness & Bodybuilding Network», 3 April 2020, https://www.youtube.com/watch?v=4ZUn1r5c3YI (last access 31 May 2024).

13 See: J. Andreasson, T. Johansson, *Doping - Historical and Contemporary Perspectives*, In J. Andreasson, T. Johansson (eds.) *Fitness Doping*, Palgrave Macmillan, London, 2020, p. 21-46; J. Andreasson, T.

beginning, the first muscular men who performed sold programs and "miracle recipes" for men who wanted to be like them. However, it was only in the 1950s that we could detect the appearance of the use of PEDs in the sense we give them today. It is said that at the World Weightlifting Championship in Vienna in 1954, Dr. John B. Ziegler learned of the Soviet experiments with anabolic substances, and upon returning to the USA, he developed Dianabol, marketed in 1958 by Ciba.[14] From this moment on, the use of PEDs began to spread in bodybuilding to the point of becoming common. In the same years, an "anti-doping" culture began to spread in other sports. However, in bodybuilding, such an attitude seemed absent, perhaps also in relation to its being both a subculture and the little scientific data on the side effects of these substances. While in the 70s, the use of steroids and other substances was natural and was talked about calmly, so much so that there were even guides for their use, things began to change in the 80s, marking a new phase. The use of PEDs seemed to spread in sports, and anti-doping policies became increasingly stringent. This also affected bodybuilding, where increasingly advanced muscles became an accusation of being "fake". In this phase, PEDs continued to circulate; however, they became a taboo, something whose use was reserved for the initiated only and which could only be spoken about in secret. It is interesting to report an ethnographic anecdote by Allan Klein that offers the measure of this. Klein spent many years conducting ethnographic research inside significant gyms in the US, interviewing various bodybuilders, and producing some of the first academic works on bodybuilding. He once said:

> A few days later I came in as usual. A small cluster of bodybuilders were huddled over the latest issue of one of the premier publications in the sport, a ritual repeated in the gym each month on the day it arrives. The bodybuilder in question was flanked by his friends, poring over the magazine and commenting on each picture. When they reached the advice column he writes, he read aloud a question sent him by a teenager in Pontiac, Michigan. The question concerned what sort of steroids were best to take. As he read the question, he imitated the high-pitched voice of his fan. Laughter all around. Then he went on to read his advice to the young man, which went something like this: "Don't destroy yourself. If you want a physique like mine, don't take shortcuts." Convulsing laughter. "I didn't win my titles by taking drugs. Chemicals are not substitutes of hard work". He would have continued, except that he was wiping tears from his eyes. His friends were on the floor.[15]

The anecdote manages to show the canonical attitude towards doping: it is used, and everyone knows it, but it should not be talked about. This attitude has managed, on the one hand, to protect bodybuilding from accusations of being an intrinsically perverse activity compromised by doping. On the other, it has caused a

---

Johansson, *Bodybuilding and Fitness Doping in Transition. Historical Transformations and Contemporary Challenges*, cit.; D. Liokaftos, *A genealogy of male bodybuilding*, cit.

[14] M. Kremenik et al. *A historical timeline of Doping in the Olympics (Part 1 1896-1968)*, in «Kawasaki journal of medical welfare», XII, n. 1, 2006, p. 19-28.

[15] A. Klein, *Little big men. Bodybuilding Subculture and Gender Construction*, State University of New York Press, New York, 1993, p. 28-29.

closure, accentuating its subculture character. Today we are going through a new phase in which the use of PEDs is not only declared but also claimed. This phase can be dated between the end of the 2010s and the beginning of the 2020s of the 21st century. This change does not only concern the use of PEDs but bodybuilding more generally. It is not my intention to explain here the reasons that caused the change; I will only limit myself to presenting some interesting features that are discontinuous with the previous period. First of all, in the post-pandemic years, there has been an increase in the number of minors enrolling in gyms worldwide. People are starting to go to the gym earlier and earlier, even in parts of the world where this was not common.[16] Furthermore, until a few years ago, there was a certain mistrust towards the diffusion of bodybuilding content through social networks.[17] They have become particularly common at the moment, also thanks to the younger members of the community.[18] This growth in online content has been accompanied by increased ease in talking about PEDs, being transparent about their use, and how to take them.[19]

Even at the level of traditional media, this cultural shift can be noted in which there is more openness in stating that one uses PEDs. Take two television programs, for example. *Il Testimone* was an Italian television program that aired from 2007 to 2021, first on MTV channels, then TV8, and finally, Sky, hosted by the presenter, actor, and director Pif. In the second season, which aired in 2008, the host interviewed and documented the days of the bodybuilder Daniele Seccarecci, who was probably the most important Italian bodybuilder in the early 2010s. In the interview, Seccarecci categorically denied using doping substances, although a few years later, he was

---

16 M. Naglazas, *Gen gym: Why the young are leading the fitness revolution*, in «Western Australia Today», 29 May 2023, https://www.watoday.com.au/national/western-australia/gen-gym-why-the-young-are-leading-the-fitness-revolution-20230524-p5db18.html (last access 31 May 2024); M. Dogra, *Growing Gym culture among youngsters,* in «Daily Excelsior», 17 September 2023, https://www.dailyexcelsior.com/growing-gym-culture-among-youngsters/ (last access 31 May 2024); Anonymous, *Gym craze among young adults is rising: Here are few dos and don'ts*, in «Times of India», 28 May 2023, https://timesofindia.indiatimes.com/life-style/health-fitness/fitness/gym-craze-among-young-adults-is-rising-here-are-few-dos-and-donts/photostory/100548176.cms (last access 31 May 2024); M. Ierace, *Se la palestra si fa precoce* [*If the Gym comes early*] in «Radiotelevisione Svizzera Italiana» 3 February 2023, https://www.rsi.ch/info/ticino-grigioni-e-insubria/Se-la-palestra-si-fa-precoce--1809980.html (last access 31 May 2024).

17 M. L. Wellman, *What it means to be a bodybuilder: social media influencer labor and the construction of identity in the bodybuilding subculture*, in «The Communication Review», XXIII, n. 4, 2020, p. 273-289.

18 V. A. Goodyear, *Young People, Social Media and Health. A Pedagogical Perspective on Influencers* in S. Lawrence (ed.) *Digital Wellness, Health and Fitness Influencers*, Routledge, London, 2022, p. 161-174.

19 M. Underwood, *Taking 'the God of all Steroids' and 'Making a Pact With the Devil': Online Bodybuilding Communities and the Negotiation of Trenbolone Risk* in A. Henning, J. Andreasson, J. (Ed.) *Doping in Sport and Fitness*, Emerald Publishing Limited, Leeds, p. 111-136; L. Hilkens et al. *Social Media, Body Image and Resistance Training: Creating the Perfect 'Me' with Dietary Supplements, Anabolic Steroids and SARM's*, in «Sports Medicine», VII, n. 81, 2021; L. T. J. Cox, L. Paoli, *Social media influencers, YouTube & performance and image enhancing drugs: A narrative-typology*, in «Performance Enhancement & Health», XI, n. 4, 100266.

investigated for trafficking in doping substances.[20] On the contrary, recently, the two most important Italian bodybuilders in the open category with participation in Mr. Olympia, Andrea Muzi and Andrea Presti, have clearly and unequivocally admitted to using PEDs during one of the most watched Italian programs: *Le Iene*.[21]

The cultural change that is taking place is not only seen in the increase in content that talks about PEDs on social networks but also in the terminology. Alongside the term "doping", we increasingly find the terms "enhanced" or colloquially "juiced" as opposed to "natural" or "natty". This terminology finds a counterpart in a broader movement of re-evaluation of PEDs, which finds its peak in the planning of the *Enhanced Games*, a sporting event, in open contrast to the Olympics and the anti-doping rules imposed by WADA, which has the objective of having athletes who are openly enhanced compete and to break down the prejudice against PEDs.[22] As in the past, there are athletes who, while using PEDs, try to hide it, both in public statements and especially in anti-doping controls. For this reason, the word "natural" in bodybuilding comes to indicate someone who may have used PEDs but who, at the time of the competition or control, is not using them. On the contrary, the term "drug-free" in recent years has been used to indicate those who have never used illegal PEDs.[23] However, the contrast between PEDs users and non-users remains quite evident in terms of media content.

4. The Positive Effects of Transparency on PED Use

My thesis is that transparency in the narrow sense of PEDs use in bodybuilding is something to be encouraged and has a positive value. My argument is consequentialist: that is, being honest about PEDs use has more positive than negative effects. I am not arguing that transparency is a virtue of sport or should be part of the values listed in the WADA Code, but that it is a practice that should be encouraged.

The first premise of this argument is that more and more young people, and more people in general, are entering the world of bodybuilding.[24] There are, therefore, more and more people who want to get a muscular physique.

---

[20] The episode of *Il testimone* is no longer available in its entirety, however excerpts containing the point in question are available online on YouTube at the following link: https://www.youtube.com/watch?v=1blZfk9SviM (last access 31 May 2024).

[21] The first episode on PEDs in bodybuilding is available at this link: https://www.iene.mediaset.it/video/bodybuilding-naturali-dopati_1303415.shtml (last access 31 May 2024).

[22] https://enhanced.org/ (last access 31 May 2024).

[23] D. Liokaftos, *Defining and defending drug-free bodybuilding: A current perspective from organizations and their key figures*, in «International Journal of Drug Policy», LX, p. 47-55.

[24] N. Bennett, *A New Era of Bodybuilding at CC*, in «The Catalyst», 1 February 2024, https://thecatalystnews.com/2024/02/01/a-new-era-of-bodybuilding-at-cc/ (last access 31 May 2024); D. Collier, J. Anderson, *Bodybuilding, Weightlifting Gains Popularity Among Students and Staff*, in «The Rider Online» 8 November 2023, https://therideronline.com/top-

The second premise is the increase in the use of social networks also to transmit fitness content. Professional bodybuilders have become influencers, and people take them as models. Therefore, those who try to develop their physique have as a model that they are inspired by and try in some way (even imperfect or mediated) to reach that of the professional bodybuilder. This is not a new phenomenon because information passed through physical culture magazines before the advent of social networks. However, the pervasiveness of social networks has changed the phenomenon's scope. As often happens, a change in quantity causes a change in quality.

Given these two premises, it is reasonable to conclude that a *frustration* effect could occur. I could try to look like someone, wish to have a physique like them, but without succeeding. In fact, if my model denies being "enhanced" when, in reality, he is, my desire to equal him, to reach him or at least resemble him, or to achieve certain types of muscularity would be doomed to failure. I can follow the training protocols used for a diet similar to his, but I still need an essential piece of information: he uses PEDs, and I do not. Lying about using PEDs is essentially a "scam": it means encouraging the belief that certain levels of muscularity can be achieved in a "natural" or "drug-free" way when in reality, this is not the case.

For this reason, it is positive that bodybuilders, especially professionals who compete at high levels, are transparent about the fact that they use PEDs because they make it clear to the people who look up to them that a physique of that type cannot be achieved without the use of currently illegal substances. The argument has a pedagogical value in that it is a matter of not deluding people about the relationship between means and results. I have argued that the justification for the positive value of transparency is consequentialist. I do not want to argue that transparency is good; there may be other values with which it should be balanced or cases in which it does not produce positive effects. In this specific case, it produces a positive effect because, on the one hand, it allows aspiring bodybuilders, beginners, and especially young people who are new to this sport to reshape their expectations or to know that to reach specific goals, you must use PEDs; on the other hand, it avoids the frustration of not achieving results. This aspect is not secondary.

Many studies show how the image transmitted by fitness influencers on social networks such as Instagram or TikTok affects the psychology of followers. In some cases, continuous exposure to images of extremely muscular physiques, such as those of bodybuilders, can produce an extreme degree of dissatisfaction and an almost pathological search to transform one's body. To describe the condition of those who perceive significant discomfort for their body and subject themselves to exhausting training, a rigid diet, and severely limiting social interactions, the DSM V has coined

story/2023/11/bodybuilding-weightlifting-gains-popularity-among-students-and-staff/(last access 31 May 2023).

the label "muscle dysmorphia".[25] It is also related to exposure to online content that fuels the desire for a radically transformed body.[26]

There are no studies that show that knowing that your model uses PEDs reduces the psychological suffering experienced in trying to match his results. However, it is reasonable to imagine that having the belief that a muscular Mr. Olympia physique cannot be achieved without the use of PEDs can lead to lowering the bar of goals. Suppose I know that to achieve a particular goal, I have to do a whole set of things, such as training a certain number of times a week, eating a certain way, having a particular lifestyle, and taking certain supplements, but I ignore the fact that I also have to take PEDs. In that case, I may be overcome by despair when I cannot achieve the desired results. This is because essential information is hidden. However, If I know, I can act accordingly and decide if I have reasonable goals consistent with the available means. Furthermore, being transparent about the use of PEDs avoids the disappointment effect when you discover that your idol has used them. In other sports, the discovery of an athlete's use of PEDs triggers a whole series of legal and sporting consequences, such as public censure and the loss of titles and recognition; think of the case of Lance Armstrong in cycling.[27] What were thought to be the results of talent, dedication, and hard work become, in common sense, the results of someone who cheated and played dirty: a fake. Knowing from the beginning that the athlete uses PEDs avoids exposing the image of the athlete to all this.

It will be argued now that this argument is exposed to at least one very obvious objection, that being transparent about the use of PEDs encourages the use of such substances, and this goes against the idea that doping is wrong on the one hand and against the fact that it is legally prohibited on the other. Although I know there may be other objections, I focus only on this.

---

[25] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*, American Psychiatric Association, Washington DC, 2013; H. G. Pope et al. *Muscle dysmorphia. An underrecognized form of body dysmorphic disorder*, in «Psychosomatics» CCCLXXXVI, 1997, p. 548-57; I. H. Steele, H. G. Pope Jr, G. Kanayama, *Competitive bodybuilding: fitness, pathology, or both?* in «Harvard review of psychiatry», XXVII, n. 4, 2019, p. 233-240.

[26] E. Chatzopoulou, R. Filieri, S. A. Dogruyol, S. A. *Instagram and body image: Motivation to conform to the "Instabod" and consequences on young male,* in «Journal of Consumer Affairs», LIV, n. 4, 2020, p. 1270-1297; K. Schoenenberg, A. Martin, *Bedeutung von Instagram und Fitspiration-Bildern für die muskeldysmorphe Symptomatik,* in «Psychotherapeut», LXV, 2020, p. 93-100; L. Paulson, *#gotmuscles? Instagram and Body Image in College Men*, in «The Journal of Social Media in Society», IX, n. 1, 2020, p. 63-84; J. Cuadrado et al. *"Muscle Pics", a new body-checking behavior in muscle dysmorphia?* in «L'encéphale», XLIX, n. 3, p. 241-247.

[27] P. Dimeo, *Why Lance Armstrong? Historical Context and Key Turning Points in the 'Cleaning Up' of Professional Cycling*, in «The International Journal of the History of Sport», XXXI, n.8, 2014, p. 951-968; M. Spalletta, L. Ugolini, *Sports journalism between doping allegations and doping evidence. The coverage of Lance Armstrong in Italian newspapers*, in «Catalan Journal of Communication & Cultural Studies», VI, n.2, 2014, p. 221-238.

5. Legalizing Doping?

The question of decriminalization or the moral legitimacy of PEDs use cannot be resolved here. However, I will provide some answers to the objection that transparency in PEDs use incentivizes their use. I do not want to deny this phenomenon. This is an intuitive idea with some foundation. Since people have started talking more freely about PEDs on social networks, there has also been an increase in content that guides their use.[28] My argument is that this is not a bad thing.

Encouraging PEDs use, according to this objection, is bad for two reasons: first, because the use of PEDs that are on the WADA Prohibited List is currently illegal; second, because the use of PEDs is morally wrong. One can argue this objection for both reasons or just one. I will attempt to answer each separately to offer a well-rounded defense of my argument.

The first part of the objection, that transparency about PEDs incentivizes their use and therefore incentivizes the use of illegal substances, is missing the mark. My argument here is moral, not legal. Indeed, a certain number of drugs are currently illegal, but they may not be. Transparency about PEDs could be seen as a transformative movement to change these policies, an act of civil disobedience or conscientious objection. When abortion was illegal, some practiced it in secret. In many cases, these clandestine abortions became an occasion for protest: women who had had abortions and doctors who had performed abortions would report themselves in protest as an act of rebellion to trigger change. If we make the comparison with the case of bodybuilding and PEDs, it would be like saying that the transparency of women who declared they had had abortions was an incentive to perform abortions. That is precisely how it was, and that was the point. It was a practice that arose on the one hand from a desire for transformation and on the other from moral disagreement on the legal prohibition of abortion. The same can be said for transparency in the use of PEDs, in which, on the one hand, one would like to be able to change the way they are seen and, on the other, to abolish the legal prohibition on using them. The supporter of this objection could continue to say that my answer only works if one accepts the goodness or moral legitimacy of the use of PEDs. This brings us to the second part of the objection, that PEDs should continue to be prohibited and that their use is morally wrong. It is beyond the scope of this paper to

[28] K. van de Ven, Katinka, K. J. D. Mulrooney, *In a bid for the perfect profile pic, young men are increasingly turning to steroids*, in «The Conversation», XXIII, 2016, https://theconversation.com/in-a-bid-for-the-perfect-profile-pic-young-men-are-increasingly-turning-to-steroids-60874 (last access 31 May 2024); L. Cox, N. Gibbs, L. A. Turnock, *Emerging anabolic androgenic steroid markets; the prominence of social media*, in «Drugs: Education, Prevention and Policy», XXXI, n. 2, p. 257-270; L. Paoli, L. Cox, *Across the spectrum of legality: The market activities of influencers specialized in steroids and other performance and image enhancing drugs,* in «International Journal of Drug Policy», CXXIII, 2024, 104246. For example see the guide inspired by harm reduction principles edited by The Love Tank and Queer Health: B. Weil, *Demystifying Steroids. Your guide to safer anabolic steroid use for building muscle with fewer risks*, 2024, https://static1.squarespace.com/static/60be2f8a0cc8001044609e26/t/655ccb6d12b05c46c9c2f69a/1700580206131/Demystifying+Steroids.pdf (last access 31 May 2024).

argue for the moral goodness of doping and the use of PEDs. Here, I will limit myself to making some suggestions regarding the use of doping.[29]

First, there is no consensus on the immorality of doping, either in philosophical or sports contexts. The majority and the foremost philosophers of sport argue for its illegitimacy.[30] However, there is room for dissent: there are those who argue that there are reasons to support its legitimacy[31], and others nevertheless admit that the reasons supporting the prohibition are weak[32]. Even in the world of sports, the Enhanced Games campaign is correct at the level of proposing a different moral paradigm regarding sports and the use of PEDs.

Secondly, it is impossible to eradicate cheating from sports; it is inherent to it.[33] So, it is very likely that as long as PEDs exist, they will continue to be used. The current system, which criminalizes them, only allows their use to take place in the shadows. Making PEDs use visible could lead to a paradigm shift from total bans to harm reduction. Eric Moore and Jo Morrison have recently defended medically supervised doping.[34] In their argument, the starting point is that anti-doping programs do not work. This assumption is difficult to dispute, given that although WADA continues to be particularly severe in anti-doping policies, doping continues not to be eradicated from sport. If doping is unavoidable in sports, the authors argue, then let us make it legal under strict medical supervision. In this way, we also avoid some of the problems that PEDs generally cause, namely health problems, which are often accentuated precisely by the fact that they must be taken secretly. The two authors argue that in this way, a relationship of trust would be created between the athlete and

---

[29] A crucial point that I decide to leave aside is whether the use of PEDs is contrary to the spirit of sport. The expression "spirit of sport" although used in the WADA code is deeply ambiguous and has raised reflections on the part of the major philosophers of sport. On this subject see: S. Loland, M. McNamee, *Fair play and the ethos of sports: an eclectic philosophical framework*, in «Journal of the Philosophy of Sport», XXVII, n. 1., 2000, p. 63-80. S. Loland, M. J. McNamee, *Anti-doping, performance enhancement and 'the spirit of sport': A philosophical and ethical critique,* in N. Ahmadi, A. Ljungqvist, G. Svedsäter (eds.) *Doping and public health*, Routledge, London, 2016, p. 111-123; S. Loland, *Performance-enhancing drugs, sport, and the ideal of natural athletic performance*, in «The American Journal of Bioethics», XVIII, n. 6, 2018, p. 8-15; S. Loland, M. J. McNamee, *The 'spirit of sport', WADAs code review, and the search for an overlapping consensus*, in «International Journal of Sport Policy and Politics», XI, n. 2., 2019, p. 325-339.

[30] For example, S. Camporesi, *Partire (s)vantggiati? Corpi Bionici e atleti geneticamente modificati nello sport*, Fandango, Roma, 2023; M. J. McNamee, J. Parry (eds.) *Ethics and sport*, Routledge, London, 1998.

[31] B. Foddy, J. Savulescu, *Ethics of performance enhancement in sport* in W. J. Morgan (ed.) *Ethics in sport*, Human Kinetics, Campaign (IL), 2018 p. 307-320; J. Savulescu, B. Foddy, M. Clayton, *Why we should allow performance enhancing drugs in sport*, in «British Journal of sports medicine», XXXVIII, n. 5, 2004, p. 666-670.

[32] R. L. Simon, *Fair Play. The ethics of sport*, Westview Press, Boulder (CO), 2010.

[33] V. Møller, P. Dimeo, *Anti-doping–the end of sport*, in «International journal of sport policy and politics», VI, n. 2, 2004, p. 259-272.

[34] E. Moore, J. Morrison, *In defense of medically supervised doping*, in «Journal of the Philosophy of Sport», XLIX, n. 2, 2022, p. 159-176. A similar argument is proposed by J. S. Russell, A. Browne, *Performance-enhancing drugs as a collective action problem*, in «Journal of the Philosophy of Sport», XLV, n. 2, 2018, p. 109-127.

the sports physician, which would replace WADA and certify that the athlete is in the right health conditions to be able to compete safely. Moore and Morrison offer a list of nine principles in order to regulate the use of PEDs:

5. There should be a prohibited substance list specific to each sport;
6. There should be education on PEDs use;
7. There should be research on PEDs;
8. There should be doctors who specialize in PEDs;
9. There should be pharmaceutical tracking on PEDs safety;
10. Conscientious objection should be allowed;
11. Penalties should be provided for those who are negligent or corrupt;
12. Confidentiality should be safeguarded;
13. The rule prohibiting PEDs use within sports codes should be equal to others.

Having a culture of PEDs, both from a medical and social point of view, would perhaps guarantee the possibility of their use in a safer way. Since safety interests us, we can only admit substances that have been tested and whose traceability in the production system is guaranteed. According to Moore and Morrison, transparency on the use of PEDs would not affect competition between athletes since, in the end, the substances are more similar than one might think. One could argue that in many sports, this proposal is unfeasible.[35] However, this model works for bodybuilding. The substances used are not very many; they are mainly anabolic steroids, testosterone and its derivatives or analogs, anti-estrogens, and a few other substances, such as diuretics or drugs for the management of sugars and carbohydrates. There is scientific literature on their use for almost all of them; their side effects are known, and active medical monitoring during their use would help limit them.

Thirdly, it promotes individual decision-making autonomy. Suppose people are genuinely aware of the risks and benefits of PEDs and have the possibility of taking them in a controlled and as safe a way as possible. In that case, we will have made the exercise of true informed consent possible, and therefore, individuals will have made a truly autonomous choice. In reality, the complete ban is characterized by a paternalistic nature.[36] Furthermore, athletes who find themselves caught in the net, that is, who are forced to take PEDs to remain competitive, do not have the possibility of being monitored during their intake. Transparency on the use of PEDs could benefit both athletes who want to use them and those who do not wish to do so. Acceptance of the use of PEDs could lead to the establishment of a dual regime: that is, competition for those who use them and others for those who do not, in order to allow everyone to make an autonomous and informed choice.

[35] See S. Holm, *Doping under medical control - conceptually possible but impossible in the world of professional sport?,* cit.

[36] W. M. Brown, *Paternalism, drugs, and the nature of sports*, in W. J. Morgan (ed.), *Ethics in Sport* Human Kinetics, Campaign (IL), 2018, p. 253-261.

6. Conclusions

I believe that the recent phenomenon of greater transparency on the use of PEDs in bodybuilding, primarily through social media, is a fact to be evaluated in a morally positive way. This is because it reveals fundamental information. This information allows people who try to achieve that result or are inspired by it to make conscious choices, whether to take PEDs themselves or restructure their goals. Transparency in using PEDs should be encouraged, as well as the dissemination of the methods of intake, controls, and risks. Indeed, it would be desirable for the transfer of information not only to occur informally through influencers and athletes but also through doctors in order to move from a model of total prohibition of PEDs to one of risk reduction.

In summary, being transparent about the use of PEDs is a positive practice, at least within bodybuilding, as it produces positive effects. For athletes, especially beginners, who do not want to use PEDs, it allows them to avoid the frustration effect, allowing them to reshape their goals in light of the awareness that a specific type of result is possible only through the use of PEDs. For those who are willing to use them, it allows greater awareness of the methods and risks involved. Furthermore, this may give rise to a transformative process that leads to a harm reduction model, where the use of PEDs is done under medical supervision. In any case, greater knowledge and awareness of the risks and benefits of PEDs can only lead to more informed choices, ensuring a more conscious exercise of one's decision-making autonomy.

The phenomenon we are witnessing of greater transparency in bodybuilding on the use of PEDs seems to be the beginning of a transformation process that seems to have many positive sides and which should, therefore, be encouraged.[37]

[37] I would like to thank for the profitable discussions on philosophy and psychology dr. Lina Maria Lissia, and the two anonymous reviewers for the useful comments and suggestions. Also I would like to thank Simone Melotti and my students with whom I discussed this and other topics related to bodybuilding. The responsibility for what I wrote remains mine alone.

Now you see me, now you don't. The predicament of Gyges in Plato's *Republic*[a]

Richard Davies[*]

Abstract

In questo saggio esaminiamo il caso di un oggetto trasparente, inteso nella sua accezione fisica di base, cioè tale che la luce lo attraversa in modo da renderlo invisibile. Il caso centrale che consideriamo è quello di Gige, raccontato all'inizio del Libro II della *Repubblica* di Platone. Trattiamo questa narrazione come se rendesse evidente un caso estremo di impunità e le sue conseguenze, e cerchiamo di tenere conto di alcuni aspetti del *topos* degli agenti invisibili che ha visto una rinascita nell'ultimo secolo e mezzo. Dopo un breve sguardo a come gli esperimenti di pensiero figurano nell'argomentazione filosofica, notiamo alcune varianti nelle storie associate al nome di Gige. I due punti principali che ci proponiamo di evidenziare sono, in primo luogo, che l'opportunità di non essere visibile a piacimento che l'anello di Gige conferisce è in contrasto con la sua capacità di essere un agente efficace, perché sarà cieco, e, in secondo luogo, che gli svantaggi di tale opportunità possono, nel complesso, superare i vantaggi, perché perde il rispetto per se stesso e per coloro che lo circondano.

Parole chiave: Invisibilità, impunità, Platone, Gige, esperimento di pensiero.

In this essay, we look at a case of a transparent object taken in its basic, physical sense of being such that light passes through it so as to make it invisible. The central case we consider is that of Gyges, as recounted at the outset of Book II of Plato's *Republic*. We treat this narrative as making vivid an extreme case of impunity and its consequences, and we try to take account of some aspects of the *topos* of invisible agents that has seen a revival in the last century and a half. After a brief look at how thought experiments figure in philosophical argumentation, we note some of the variants in the stories associated with the name of Gyges. The two main points we aim to bring out are, first, that the opportunity not to be visible at will that Gyges'

[a] Saggio ricevuto in data 09/04/2024 e pubblicato in data 22/01/2025.
[*] Professore associato, Università degli Studi di Bergamo, email: richard.davies@unibg.it.

ring confers is at odds with his being an effective agent because he will be blind, and, second, that the disadvantages of such an opportunity may, overall, outweigh the advantages because he loses respect for himself and those around him.

Keywords: Invisibility, Impunity, Plato, Gyges, Thought experiment.

1. Thought experiments

In recent decades, considerable attention has been paid to what are often called thought experiments in philosophy and related disciplines, so much so that some of the tags adopted for them, from the Ship of Theseus to Twin Earth or the Ailing Violinist and its Trolley relatives, have generated identifiable literatures. Likewise, the uses that have been made of such imaginary cases have become an object of sustained reflection[1]. So it is hardly surprising that Tim Williamson dedicates the whole of the fourth chapter of his *Doing Philosophy*[2] to the use of thought experiments to bring into focus – and even to decide – some knotty questions in the subject. Williamson fairly explicitly endorses some thought experiments as at least close to decisive, even when they involve physical impossibilities, such as Einstein's supposition about what it would be like to ride a light ray. On the other hand, he is highly critical of others, especially David Chalmers' zombie story, because they seem to depend on some conceptual or logical muddle. In between, he allows that some thought experiments may be stimulating and innocuous even if the scenario they envisage is, in one way or another, impossible. In this last class, Williamson cites the case of Gyges, though he does not specify what sort of impossibility is in the offing in such a case. Yet, Martin Hollis entitles the seventh chapter of his *Invitation to Philosophy*, dedicated to what it is to know right from wrong, "The Ring of Gyges", and goes so far as to say that "we nowadays find nothing odd in the story, as it stands"[3].

At the risk of oversimplifying, the basic use of a thought experiment is to show that, if a certain scenario is at least possible, then a certain thesis is untenable. The simplifications here are, of course, in what is meant by a scenario, by possibility and by a thesis. The scenario we shall be considering comes in a variety of versions, of which those found in Plato and Cicero are philosophically the most interesting, though they do not coincide exactly[4]. The possibility that they depend on is the compatibility of invisibility with sightedness, which too can be declined in various

[1] With a certain emphasis on scientific uses, a leading text would be T. Szabò-Gendler's *Thought Experiment: On the Powers and Limits of Imaginary Cases*, Routledge, London 1999; see, also, more recently, the thematic collection on "New Perspectives on Philosophical Thought Experiments", in *Topoi*, 38, 2019.

[2] T. Williamson, *Doing Philosophy*, Oxford University Press, Oxford 2018.

[3] M. Hollis, *Invitation to Philosophy*, Blackwell, Oxford 1985, p. 128.

[4] E.g. M. Shell, "The Ring of Gyges", *Mississippi* Review, 17, 1989, pp. 21-84; A. Laird, "Ringing the Changes on Gyges: Philosophy and the Formation of Fiction in Plato's *Republic*", *The Journal of Hellenic Studies*, 121, 2001, pp. 12-29; R. Woolf, "Cicero and Gyges" *The Classical Quarterly*, 63, 2013, pp. 801-12.

degrees. The thesis they put to the test comes out at the very beginning of book II of Plato's *Republic*, where Glaucon wants Socrates to show that, in every case, it is better to be just than unjust (357b1-2), and to show the untenability of the common opinion that injustice pays when it is not punished (359a2-c5). So what Glaucon wants to investigate is what it would be reasonable to expect of the behaviour of a person who knows s/he can commit injustice with impunity. One way to acquire impunity is not to be seen when committing injustice. Gyges exemplifies a person who commits injustice taking precautions not to be seen.

2. A classical topos

Stories about Gyges abound in the ancient literature to such an extent that we may have to do not so much with the name of an individual as with something more like a dynastic title[5], which would go some way to explain the slight contortion in Plato's text, where Gyges is referred to as the ancestor of the Lydian (359d1), which latter reference may in turn denote Croesus[6]. Yet, when reference to the ring returns in *Republic* X (612b5), it is attributed directly to Gyges. For present purposes, we may downplay some of these complications, which have been extensively studied[7], to concentrate on the special use to which Plato puts his version of the tale. So far as I have been able to discern, the scholars who have excavated the diverse figures of Gyges have not been much exercised by the sort of worry that we wish to raise here, and have tended to take Plato as offering a mere variant on a theme. It may even be that classical scholars are so inured to stories with magical elements as not to be fazed by what we find in the *Republic*.

To simplify, then, we may stick to the two main versions of the tale[8].

One, whose *locus classicus* is in the first book of Herodotus' *Histories* (chapters viii and following), has Gyges ordered by Candaules, king of Lydia, to view his wife – whose name is not given[9] – while she undresses at night in order to prove to Gyges that she is the most beautiful of women. Despite his protests at the order, while Gyges watches from behind the bedroom door, the queen catches sight of him and, the following day, she offers Gyges a choice between killing Candaules for so

[5] K. Flower Smith, "The Tale of Gyges and the King of Lydia", *American Journal of Philology*, 223, 1902, pp. 261-82 and 361-87 counts as many as five different versions of his seizure of power (p. 267, n. 1).

[6] S.R. Slings, "Critical Notes on Plato's *Politeia* II", Mnemonsyne, 42, 1989, pp. 380-97; also Vegetti's note to this passage: *La Repubblica* Italian translation with commentary edited by M. Vegetti (Vol. II, devoted to books II and III), Bibliopolis, Naples 1998, p. 30.

[7] A splendidly well-documented and compact overview is offered by Francesca Calabi in note [B] in Vegetti ed. cit., pp. 173-88.

[8] Others who have made this simplifying move include G. Danzig, "Rhetoric and the Ring: Herodotus and Plato on the Story of Gyges as a Politically Expedient Tale", *Greece and Rome*, 55, 2, 2008, pp. 169-92.

[9] Referring to Photion, Calabi, notes the name Nysia and recalls also the names Tudo, Clyzia and Abro (in Vegetti, ed. cit., n. 29, on p. 186).

dishonouring her or being killed himself. Reasonably enough, Gyges goes for the former option and becomes the husband of the queen and, so, king, which promotion is later endorsed by the oracle at Delphi. In this version, which was recurrent in the subsequent Greek and also Latin literature, we have little more than some blood-thirsty palace politicking, with a bit of sexual voyeurism thrown in for good measure.

The other version of the Gyges story has a rather more supernatural edge to it. This is the version that we find in Plato (*Resp.*, II, 359d-60c) and that hardly resurfaces either in Greek or in Latin before Cicero's *De officiis* (III, 38-9), a text that names Plato as its inspiration. We (or at least I) do not know whether Plato was drawing on some pre-existing bit of folklore. But, from the curiously circumstantial way in which the story opens, with a humble shepherd not merely stumbling on a magic ring, but finding it on the finger of an outsize corpse hidden inside a buried bronze horse revealed by a rain-induced landslide, we might imagine a tradition of story-telling among shepherds around the camp fire that accreted circumstantial elements in the transmission. Or, more simply, it may have been Plato himself who conjured a scenario in which the discovery of so curious an object is no more surprising than the circumstances of its discovery. In this version, it takes Gyges some time to learn how to use and how to exploit the powers of the ring. But, as in the Herodotus version, he seduces the queen or rapes her: Plato uses *moikhao*, while Cicero uses *stuprum*. Then, with her help, he kills Candaules (who, in these versions is anonymous) to become king of Lydia himself.

3. The ring's powers

The question then is: what powers are we meant to attribute to the ring in the Plato/Cicero version? At the very least, we may say that, when Gyges turns the bezel or collet of the ring towards the inside of his hand, no-one else can see him. This is as far as Cicero goes, saying that he could not be seen by anyone (*a nullo videbatur*: III, 38). But Plato hazards a bit more, saying that Gyges has become *adelos* (360a6), a word whose most primitive sense seems to be something like «secret», but that gets extended to mean «invisible», which corresponds to what appears in all the translations into modern languages that I have consulted.

In the Herodotus version of the story, when Gyges hides behind Candaules' bedroom door, he should not be visible from within the bedroom. But he is surely visible to someone who happens to be passing by outside the bedroom, and he is not sufficiently well hidden to prevent the queen from catching a glimpse of him and recognising him. That is to say, the relative spatial positions of Gyges and potential viewers make all the difference to his visibility. But, in the Plato-Cicero version, when the ring is suitably adjusted, there seems to be no position from which Gyges can be seen.

Almost irrespective of what theory we might adopt of how light propagates, it is not unreasonable to say that light passes through Gyges' body at least in this sense: if he is standing against the door of Candaules' bedroom, what we see is the door and

not Gyges. One way of expressing this is to say that his body is transparent.

4. Being transparent

One way that a body can be transparent is for it to be made of glass. I do not see the glass of the window in my study because I can see the tree outside; or: I can see the tree outside because my window is made of glass and, so, transparent. As readers of Descartes' *Meditations* will recall[10], the delusion that one is made of glass was something of an epidemic in early modern Europe, especially among persons who were much in the public eye, such as king Charles VI of France in the fifteenth century. If he had indeed been made of glass, poor Charles would not have been «in the public eye». Not only did he feel himself to be as fragile as glass, but also wished not to be seen. In this fantasy, though his robes and crown would have been visible, they would have seemed to be moving on their own without a wearer. Yet, presumably, someone who had the audacity to place their hand under the crown would have encountered something solid, namely the glass that had taken the place of Charles' flesh and bone[11].

The glass delusion epidemic has sometimes been attributed to the relative novelty of paned windows in the period in question and it seems to have gone extinct by the beginning of the nineteenth century. But, over the last century and a half, variants of it have returned with a vengeance in fiction, perhaps beginning in 1897 with H.G. Wells' *The Invisible Man*[12]. In this case, we have a chemist called Griffin who concocts a brew that alters the refractive index of his body so as to make it invisible. Wells' deployment of the quasi-technical term «refractive index» is meant to make Griffin's experiment seem scientifically plausible. Indeed, this is something like the direction that some recent technological developments have been trying to exploit in order to render objects invisible by cloaking them with dielectric materials so as to disperse the light that would otherwise be reflected off their surfaces. It can hardly be a coincidence that, at least in their popularisations of their researches, the scientists working in this direction occasionally make reference to the cloak that Professor Dumbledore bestows on Harry Potter[13]. Harry can see through the cloak, but anyone looking in his direction cannot see him, nor indeed the cloak itself. Yet it seems that a cloak made of a dielectric material will be opaque and thus not allow someone behind it to see out.

---

[10] C. Adam, and P. Tannery (eds.), *Œuvres de Descartes* (12 voll., 1897-1913) corrected and added to by J. Beaude and P. Costabel (et al.), Vrin, Paris, 1964-76, VI, p. 19.

[11] In her *Real People: Personal Identity without Thought Experments* (Oxford, Oxford University Press 1988), K. Wilkes seems to think that an invisible person would also be intangible (p. 11) and that this would be a limitation on his/her efficacy in, for instance, committing theft.

[12] H.G. Wells, *The Invisible Man,* Oxford, Oxford University Press 2017.

[13] See, e.g. the report on the work of David R. Smith at Duke University: "Invisibility Cloak Demonstrated!" in *Computing News* 2006 (https://home.nestor.minsk.by/computers/news/2006/10/2003.html).

To return to Gyges, the advantage the ring confers rests on two conditions. One is that others cannot see Gyges when he is up to no good. The other is that Gyges can see what he is doing. If the first condition is satisfied by his invisibility, we might wonder whether his invisibility is compatible with the second condition.

In the Herodotus version, Gyges' not being seen was meant to be secured by his hiding behind the door, but this failed. In the Plato version, the second condition seems more problematic. There are of course circumstances in which one can see without being seen. Setting aside the technology of video-cameras and the like (in which at least the camera is potentially detectable by sight), the one-way mirrors that the police install in interrogation rooms allow the interrogator's colleagues to see what is going on, without the suspect seeing them. The trick is quite easy and indeed dates to just a few years after Wells' story: a certain Emil Bloch was granted US patent 720877 in 1903 for the design of a one-way mirror. In the interrogation room there is a glass surface that reflects light; and in the adjacent observation room there is the other side of the same glass that is transparent. While light does indeed pass in both directions through the glass, the observation room is dimly lit relative to the brightly lit interrogation room. As a result, the suspect is able to see *something* – what appears to be a normal mirror and hence, for instance, his own face in it – and yet the police have a window on him without themselves being seen. Even when, as in the Clint Eastwood movie *Absolute Power* (1997), the spyhole is very small, there is still something to be seen that might alert suspicion.

But this will not quite do for what Gyges needs, which is for there to be *nothing* that others see while he sees them. Moreover, Gyges needs to be mobile and not restricted to just one viewing post.

5. What the invisible man sees

It may be at this point that we run into what Williamson might have been thinking of as an impossibility in having a clear and distinct conception of Gyges' predicament. We may put the point in the manner that Soviet science populariser Yakov Isidorovich Perelman (1882-1942) used to debunk Wells' invisible man. In the second volume of his *Physics for Entertainment*[14], Perelman points out that, if Griffin is invisible, then light passes through him. If light passes through him, then it is not stopped even by the retinas of his eyes. Unless his eyes stop light, Griffin must be blind. If he is blind, then he will not be able on his own to find his way about. So Griffin's invisibility does not give him an advantage in putting through his affairs. And the same would, presumably, apply to Plato's Gyges.

Plato himself might not have been wholly impressed by this move, but if we restrict ourselves to at least moderately plausible theories of how seeing works, it seems that a person who is invisible will be transparent and, so, will be blind. If so, one sort of impossibility that Williamson might have been thinking of in saying that

[14] Y. I Perelman, *Physics for Entertainment*, (13th ed., 1936); tr. Eng., Hyperion, New York 1975, pp. 242-9.

the Gyges story is impossible arises precisely from the incompatibility of invisibility with sightedness.

As indicated, the incompatibility derives from what we understand about how seeing works. Up to a point, this is a matter of how eyes work. Yet there are in nature very different conformations of eyes, some of which, for instance, obviate the blind spot that, in humans and other mammals, is due to the connection of the optic nerve at the back of the retina. It is perhaps not completely irrelevant to note that, on Earth, there are at least forty kinds of animals – butterflies, squid, juvenile eels, shrimps, slugs, even frogs, but especially, deep-sea fish such as *salpa maxima* – whose bodies are transparent. But, again, only up to a point. In all the cases I have been able to identify, such animals are able to see because at least the retinas of their eyes are opaque: generally black and also large relative to body size. But there are reasons for thinking that their visual capacities are rather limited by the lack of the *camera obscura* effect provided by the opaque sclera of, for instance, the normal human eye.

Given that there is some leeway or contingency about how seeing actually works in the animals we know something about, we might find wriggle room for the possibility of there being a creature that was both transparent and sighted. We might say that this is a logical or metaphysical possibility, without committing to its being a physical possibility: the description of such a creature may not lead to a flat-out contradiction or other incoherence, but it is at least at the outer edge of conceivability. After all, Plato and Wells at least took themselves to be conceiving of such cases. As has been much discussed in recent literature[15], the fact of conceivability may be a poor guide to what is physically possible. Indeed, we might think that we are conceiving of St Anthony of Padua's bilocation or of Pegasus' musculature that allows him to fly. But that shows neither that such phenomena are physically possible, nor even that we are thinking the things through sufficiently to say that we really have such conceptions (especially not of the «clear and distinct» variety to which we have already made reference).

One impression that one might come away with is that the Plato/Cicero version of the story of Gyges is rather an intuition pump, in something like Dennett's sense[16], than a fully-fledged thought experiment: it encourages us to think about impunity in a certain way, but need not depend on our granting the physical possibility of an invisible man who can act effectively because he is sighted. For myself, Gyges' blindness is an obvious consequence of his invisibility and my experience is that, when I present this consequence even to seasoned philosophers who are acquainted with Plato's text, its obviousness is pretty uniformly accepted without need of further explanation. I would even go so far as to say that some of my interlocutors have had

[15] E.g. T. Szabò-Gendler and J. Hawthorne (eds.), *Conceivability and Possibility*, Oxford University Press, Oxford 2002.

[16] D.C. Dennett, *Elbow Room: The Varieties of Free Will Worth Wanting*, Clarendon, Oxford 1984, esp. pp. 17-8 and pp. 32-4.

«ah-ha» responses[17]: even if they had thought about the case of Gyges, for instance for the purposes for which Glaucon introduces it, they hadn't thought of it as raising a question of its physical or effective possibility.

6. What the ring brings

We have been speaking of various grades of possibility and impossibility – the logical, the metaphysical, the physical – without recourse to the technicalities of the modal systems that can be used to model them. We have suggested that any circumstance that is impossible in any of these grades is not the case. But there is a large range of uses of modal terms such as «must» and «have to» or «can» and «may» as well as «obligatory» and «forbidden» in which we cannot infer from a necessity that something is the case or from an impossibility that it is not. Some of these uses can help us to get a grip on another way in which the story of Gyges presents us with an impossibility.

At *Republic*, 362b3, Glaucon introduces the idea that, if there could be one magic ring like the one Gyges finds, there might be two. He hypothesises that, if one were given to a just man and the other to an unjust man, they would both act in the same way, robbing and raping and murdering. Glaucon's aim in invoking this scenario is to indicate that no-one, not even the most just person, would be ready to forgo the advantages of being able get away with injustice, and would likewise be ready to accept the disadvantages. So it may be worthwhile considering how the advantages of possessing a magic ring stack up against the disadvantages.

The most obvious advantages that accrue to Gyges' discovery of his ring and of how to use it are made explicit in Plato's text. Having once got into the palace, killed the king and married the queen, he no longer has to work as a shepherd and undergo the hardships of the outdoor life. He has easy access to food, which will be prepared for him at his whim by the palace cooks. The queen, as well as any other woman (or man) he takes a fancy to, will be at his disposal to satisfy his sexual urges. Money and anything that it can buy can be obtained either in his role as king or by further employment of the ring. He can exercise power over others both within Lydia, giving orders whose execution he can verify under cover of invisibility, and beyond, subverting other sovereigns and imposing his own rule. And, being recognised as king, he will be at least outwardly honoured even by those who are envious of his position.

In the terms of Glaucon's tripartition of goods at *Republic*, II, 357b-8c, leisure, nutrition and sexual satisfaction seem to fall at least near the category of things desirable for themselves (though Glaucon mentions only joy and harmless pleasures: *hēdonai ablabeis*), while money, power and honour are closer to being desired mostly as mere means and so desirable primarily for their effects. According to the common opinion that Glaucon sets out at 359-60, justice comes about as a compromise

[17] I am particularly proud of having produced this effect on an incumbent of the Bertrand Russell Chair of Philosophy at the University of Cambridge.

between the strong and the weak, so that those who practice it do so against their will, and injustice pays when it goes undetected and unpunished. If, by strength, cunning or luck, someone is able to avoid detection – as Gyges is –, then it looks as if it is not irrational for him to commit injustice and reap the benefits of so doing.

But at what cost? If what Socrates wishes to argue is that justice is to be pursued come what may, then, even when presented with the case of Gyges, he is free to bring to our attention what Gyges loses, and to argue that these losses are greater than the gains accruing to the ring of invisibility. The losses and gains in question may be hard to quantify, there not being a single obvious scale on which to measure them against each other. But things that are hard are sometimes worth trying[18]. Even if Gyges' use of the ring is only to be expected, it may be an open question whether or not it is really the rational course and, so, whether or not he would be better off without the temptations that are put in his way.

At a first stab, there may be as many (or as few) as four more or less interrelated dimensions to Gyges' losses in making his nefarious uses of the ring. At least in Plato's version, we are told nothing about what sort of person Gyges was before his find. All we know is that he was a shepherd, and shepherds may be caring and conscientious or exploitative and devious, with all the other combinations of character traits we might want to pull in. But the temptation that the ring exerts once Gyges grasps its powers may mean that, in some pretty strong sense, he loses his identity, whatever it was. He becomes, rather, the slave of the ring, not unlike Tolkien's Sméagol/Gollum.

Related, but distinct, is Gyges' loss of what we might call his sense of integrity. Even if, before discovering the ring, he was wily and exploited to his own benefit situations in which he could get away with unjust behaviour, the ring's powers mean that he cannot even congratulate himself on not having been detected. Conversely, if he had been sincerely (or merely unreflectively) law-abiding, when he finds himself drawn to all sorts of treacherous behaviour, he should (in one of the senses gestured at above) feel ashamed of himself. He knows that he is acting unjustly and does not deserve the benefits that accrue to him. By way of analogy, an athlete who takes illicit performance enhancers cannot – or at least should not – feel that the medal she wins is fully merited. One might even go so far as to say that she does an injustice to herself[19].

Perhaps as a corollary of the loss of integrity, the recognition that his ill-gotten gains are mere effects of fortune should lead to a loss of self-esteem. Here, the «should» is relatively weak. It has rather less than the force of a prediction, because of the unsightly fact about human beings that the unjustly advantaged, such as middle-class white males in most Western societies, tend to think of their advantages as deserved. Yet, while a classist, racist, sexist society is what most middle-class white

[18] The remaining books of the *Republic* (especially book VIII and IX) might be read as constituting Plato's own approach to elaborating such a scale.

[19] Thus, for instance, K. Gongaki, "The Platonic Myth of Gyges and the Concept of Justice and Injustice in Modern.Day Sport and the Contemporary World", *Electryone*, 5.2, 2017, pp 1-11: 6.

males are accustomed – and hence oblivious – to, Gyges' radical change of status might give him pause for reflection on the aleatory nature of his sudden promotion: it was no doing of his and, hence, no merit accrues.

A fourth feature of Gyges' new situation that pretty clearly counts as a loss regards his relations with the people with whom he comes in contact. At the very least, he has to be wary of having the powers of the ring discovered. Indeed, not even the queen can be let in on his secret: she must not know how Gyges pulled the trick in the first place, and then she must be kept in the dark about his subsequent uses of the ring. Here, the «has to» and the «must»s are on pain of the queen's snatching it for herself. Gyges is in mortal danger if anyone should uncover how he gets away with what he does. As a result, he is condemned to be distrustful and unsocial, an outcast in his own palace, so to say. Perhaps it is no surprise that, as in the cases of Griffin and Sméagol/Gollum, having the power not to be seen easily leads to mental breakdown.

Even if the four drawbacks to the possession and use of the ring of invisibility just outlined may not obviously outweigh the gains emphasised in Plato's text, it may be worth suggesting that they all seem to be harms corresponding to goods that should fall into Glaucon's first category: a sense of identity, of integrity, of self-esteem and the ability to socialise at ease all seem to be desirable for themselves and not in view of anything else. The loss of these goods may be tantamount to what Matthew (16,26) and Mark (8:36) call «losing one's soul».

We may return, then, to the cadaver from which Gyges took the ring in the first place (*Resp.*, II, 359d-e). How did it get to be buried inside a bronze statue of a horse? One attractive hypothesis is that it was a rather elaborate case of suicide, aimed at taking the ring out of circulation. The big man may have grasped that the ring presents a temptation to self-destruction[20], and decided to take it with him to the grave for the good of others. Such a gesture would mean that he had come to think that the worldly goods that one might obtain by the use of the ring are outweighed by the harms that a person may do to herself in merely possessing it. Even if the ring were not a physical impossibility, it would threaten a moral impossibility.

7. What have we seen?

Not everyone will be impressed by our attempt to illustrate why Gyges' ring must, if it is to exist, be a supernatural object. After all, not everyone thinks that, to exist, an object must be in line with what nature, as we currently understand it, allows. But we have tried to show that the description of it is at least close to being internally incoherent. Likewise, not everyone will be impressed with our suggestion that, all told, it may not be advantageous to possess the ring. After all, not everyone thinks that the loss of the moral goods at which have gestured would outweigh the gain of the

[20] As D.K. O'Connor puts it, in association also with Hades' cap, "they may as well be gifts from hell": "Rewriting the Poets in Plato's Characters" in *The Cambridge Companion to Plato's* Republic, (ed.) G.R.F. Ferrari, Cambridge University Press, Cambridge 2007, pp. 55-89, at p. 68.

material goods that are more readily imagined as accruing to the ring's possessor. But we have tried to suggest that the moral goods do at least have some weight.
Plato's use of the notion of invisibility remains a powerful stimulus to thought about the notion of justice. While he uses it to provoke us into considering an extreme case of impunity arising from not being seen by others, the tables can be turned. For instance, Rawls' use of the Veil of Ignorance can provoke us into considering how we would deliberate if we could not see ourselves and our own position in society, and what sort of society would emerge from the application of those deliberations[21]. But that is another story.

[21] J. Rawls, *A Theory of Justice* (1971) rev. ed., Harvard University Press, Cambridge Mass. 1999, esp. §24. For a vigorous application of his principles to a society like ours, see D. Chandler *Free and Equal: What Would a Fair Society Look Like*?, Allen Lane, London 2023.

Democratic Representation and Decision-making at the Time of Digital Disintermediation: A Critique of the Populist Erosion of the Role of Parliaments[a]

Paolo Monti[*], Graziano Lingua[†], Philippe Poirier[‡]

Abstract

Diversi autori hanno analizzato l'ascesa dei movimenti populisti in tutto il mondo come un fenomeno che deve essere inquadrato nel contesto di una trasformazione generale della democrazia rappresentativa in una forma di democrazia del pubblico, in cui il valore dell'intermediazione è sempre più contestato a tutti i livelli della vita sociale. Nell'esaminare questo cambiamento in corso, illustriamo innanzitutto alcune implicazioni generali che i fenomeni sociali di disintermediazione hanno per la pratica della democrazia rappresentativa, assottigliando e rimodellando i confini tra la sfera pubblica formale e quella informale. In particolare, esaminiamo la crescente influenza della leadership carismatica nella politica dei partiti e la spinta alla democrazia diretta digitale come alternativa al ruolo delle assemblee elettive, per mostrare come un ideale normativo di rappresentanza politica come specchio in tempo reale dell'opinione pubblica sia alla base di entrambe queste strategie populiste. Valutiamo poi criticamente queste implicazioni pratiche e teoriche della disintermediazione. Da un punto di vista pratico, scopriamo che le leadership carismatiche e le strategie populiste di democrazia digitale diretta non soddisfano gli standard di immediatezza e trasparenza su cui si basano e non possono sostituire la funzione democratica pluralistica delle assemblee elettive. Da un punto di vista teorico, sosteniamo che la premessa concettuale su cui si basano è fondamentalmente errata: la rappresentanza politica è un processo che implica sempre un grado rilevante di interpretazione e intermediazione, e pertanto le affermazioni dei rappresentanti non possono essere interpretate come riflessi speculari dei rappresentati. Concludiamo suggerendo che i

[a] This article is the result of a joint research, started within the PARREL Project of the University of Luxembourg and continued within the PRIN 2022 Project (JR8Z8P - Social Transformations and the Crisis of Expertise), University of Turin Unity, funded by the European Union – Next Generation EU. The first and second section has been mainly drafted by Philippe Poirier and Graziano Lingua. The third and fourth sections has been mainly drafted by Paolo Monti. The conclusions reflect the writing of all authors. Saggio ricevuto in data 31/05/2024 e pubblicato in data 22/01/2025.

[*] Ricercatore, Università degli Studi di Milano-Bicocca, email: paolo.monti@unimib.it.

[†] Professore ordinario, Università di Torino, email: graziano.lingua@unito.it.

[‡] Professore ordinario, Université du Luxembourg, email: philippe.poirier@uni.lu.

parlamenti dovrebbero invece adottare pratiche innovative come le audizioni pubbliche e la democrazia diretta avviata dai cittadini, che ricentrano la funzione rappresentativa dell'assemblea sull'ascolto attivo dei rappresentanti e sulla partecipazione dei rappresentati.

Parole chiave: rappresentanza, democrazia del pubblico, populismo, disintermediazione, crisi degli esperti politici, sfera pubblica, audizione pubblica.

Several authors have analyzed the rise of populist movements around the world as a phenomenon that must be seen in the context of a general transformation of representative democracy into a form of audience democracy, in which the value of intermediation is increasingly contested at all levels of social life. In examining this ongoing shift, we first illustrate some general implications that social phenomena of disintermediation have for the practice of representative democracy by thinning and reshaping the boundaries between the formal and informal public spheres. Specifically, we examine the growing influence of charismatic leadership in party politics and the push for digital direct democracy as an alternative to the role of elected assemblies, to show how a normative ideal of political representation as a real-time mirroring of public opinion underpins both of these populist strategies. We then critically assess these practical and theoretical implications of disintermediation. From a practical perspective, we find that charismatic leaderships and direct digital democracy populist strategies do not meet the standards of immediacy and transparency on which they are based, and cannot replace the pluralistic democratic function of elected assemblies. From a theoretical perspective, we argue that the conceptual premise on which they rely is fundamentally flawed: political representation is a process that always involves a relevant degree of interpretation and intermediation, and therefore representative claims cannot be construed as mirror reflections of the represented. We conclude by suggesting that parliaments should instead adopt innovative practices such as public hearings and citizen-initiated direct democracy, which refocus the representative function of the assembly on the active listening of the representatives and the participation of the represented.

Keywords: representation, audience democracy, populism, disintermediation, crisis of political experts, public sphere, public hearing.

1. Introduction: the rise of populist movements and the practice of representative democracy

The rise of populist movements around the world and its connection with general trends of social disintermediation in the public sphere has been subject of widespread scrutiny in the last decade. As Nadia Urbinati has aptly pointed out, this is a phenomenon to be intended as an attempt to transform constitutional democracy from its stabilized post World War II form into a new, substantially mutated model

of representative democracy.[1] Populist movements, indeed, do not entirely reject the logic of representation, but rather disfigure it by discrediting the role of political mediations, by undermining the checks on the power of majorities, and by vilifying views and groups that do not fit into their understanding of who 'the people' are and what they want.[2]

This kind of transformation has been in the making for quite some time, prepared by a general transition from parliamentarian and party based democratic models into new forms of audience democracy[3] where the relationship between representatives and represented is focused on the personal image and initiative of individual political leaders and their constant connection with the public through multiple means of communications and opinion polls.

This direct audience relationship has gradually become predominant as the value of intermediation has also been contested at all levels of social life and the fundamental institutions of representative democracy are regarded with suspicion. Pierre Rosanvallon has spoken of this trend as a form of counter-democracy: a phenomenon which encompasses the spread of anti-political sentiments among the population, a mounting request for more control over representative institutions, a systemic mistrust for political elites and traditional forms of political decision-making, and a noticeable demand for direct democracy.[4]

In this scenario, the role of political parties and parliaments in the democratic practice is frequently marginalized, manipulated and sometimes even by-passed at the hands of populist leaders who seek their legitimation in a supposedly direct and disintermediated relationship with their public. This paper aims to illustrate how this crisis of parliamentary representation is rooted in a misleading concept of political representation that construes the representative relation as a 'mirror reflection' of the people. The populist representative claims, we suggest, are based on the assumption that it is possible to offer a political image of the people's beliefs, wishes and needs that is not the fruit of a lengthy process of interpretation, which includes elements of mediation, interaction and compromise, but is rather the result of a direct, unmediated and mechanically accurate mirroring of the represented 'as they are'. We first trace the preconditions of this conception in the general phenomena of social and political disintermediation and how they are thinning and twisting the separation between formal and informal public sphere. We then look closer at the relationship between representatives and represented to discuss how the notion of representation as a mirror reflection is affecting both the everyday practices of elected assemblies and the conceptual framework they are based on. Finally, we argue that construing political

[1] See N. Urbinati, *Democracy disfigured. Opinion, truth, and the people*, Harvard University Press, Cambridge MA 2014.
[2] See N. Urbinati, *Me the people. How populism transforms democracy*, Harvard University Press, Cambridge MA 2019.
[3] See B. Manin, *The principles of representative government*, Cambridge University Press, Cambridge 1997, pp. 218-235.
[4] See P. Rosanvallon, *La Contre-Démocratie. La politique à l'âge de la défiance*, Éditions du Seuil, Paris 2006.

representation as a mirror is a faulty and self-defeating path that is ultimately at odds with democratic pluralism. Political representation is an interpretive relationship, not a mirroring one, and the interpretive process involves representatives and represented together through various layers of mediation. In light of these considerations, we suggest evaluating innovative democratic practices like public hearings and citizen-initiated direct democracy practices as promising strategies to integrate the representative process of elected assemblies with the participation of citizens.

2. Political disintermediation in the public sphere

To understand the push towards disintermediation and the consequent transformations of political representation we need to consider the impact that media innovation and the digital revolution are having on the public sphere and, specifically, on the practices of political communication. These transformations, initiated by the Modern individualization and horizontalization of social relations and accelerated by contemporary technological advancements, have notably found public justification based on the appeal to the ethical-political value of transparency.

The call for transparency as a condition for political participation dates back at least to the Enlightenment and its formulation of the ideal of 'publicity', according to which the free circulation of information and the fight against the secrecy and opacity of power have an essential emancipatory political function. This ideal has been crucial to develop the concept of the bourgeois public sphere as a space where knowledge and reasons can be freely exchanged, a genuine 'sphere of criticism'[5] where the processes of discursive mediation are fundamental not only to the articulation of ideas but also to the direct agency in the political field.[6] In this context, the mediation of experts is important not only because of the role played by journalist and intellectuals, but also because of the function of elected representatives during a time that Bernard Manin has designated as the age of parliamentarianism, whose origins can be traced back to the XVIII century.

This situation starts shifting with the advent of mass media, which undermines traditional forms of political representation and participation. The increasing influence of the media gradually weakens the Enlightenment idea according to which there would be a direct proportionality between publicity and emancipation.[7] Radio and television bring about a more democratic access to political information, but also tend to transform the public into a consumer audience and to create a context that favors the impact of media manipulation over the discursive exchange of reasons. In

[5] See R. Koselleck, *Kritik und Krise. Eine Studie zur Pathogenese der bürgerlichen Welt*, Suhrkamp, Frankfurt am Main 1973.

[6] See J. Habermas, *The Structural Transformation of the Public Sphere: An Inquiry Into a Category of Bourgeois Society*, MIT Press, Cambridge MA 1989.

[7] See S. Baume, *Publicity and Transparency: The Itinerary of a Subtle Distinction,* in E. Alloa, D. Thomä (dir.)*, Transparency, Society and Subjectivity. Critical Perspectives,* Palgrave Macmillan, Londra 2022, pp. 203-224.

Manin's reconstruction of the transformations of representative government,[8] this is the time of party democracy, where the core representative relation shifts from the individual trust between representative and represented to the ideological identification between the masses and the political parties. During this time, the processes of intermediation and the role of experts are still central, even though they are gradually transformed. The rise of mass parties is supported by party activists and bureaucracies that are crucial not only to direct the electoral choices of the general public, but more widely to nurture and shape the political discussion on all matters of public concern. Journalists also retain a prominent role in choosing and framing the news that reach the wider audiences.

The third stage of Manin's account, which encompasses the late XX century, when the mass media reach the peak of their influence, sees the rise of the audience democracy model, in which the relation of ideological identification is progressively weakened and substituted by a new direct bond between political leaders on one side and passive audiences of media consumers on the other.[9] The arrival of digital media, however, pushed this ongoing transformation in new and radical directions that Manin, at the end of the 1990s, could not entirely appreciate. With the possibilities opened by the Web 2.0, the citizens become 'prosumers': they aren't now limited to the consumption of news, as they can also actively contribute to their production, thus further undermining some of the most consolidated forms of political intermediation and communication. In this new framework, where everyone can be a source of information for the general public and everybody can join a political debate from home, the role of expert intermediators is not only widely delegitimized, but it is even perceived as an obstacle to the free circulation of ideas and the direct expression of the people's political will. As Byung-Chul Han's noted, mediation and representation are now merely "viewed as a lack of transparency and inefficiency – as temporal and informational congestion".[10] This push towards disintermediation profoundly affects the political sphere and directly hits the principle of representation in a preexisting context of severe crisis of the mass parties that originated in the 1980s and has, since then, gravely wounded the legitimacy of the political elites.

Social and technological transformations are certainly offering to the citizens a widespread horizontal access to an open sphere of public representations: every individual can formulate representative claims addressed to a wide audience with a chance of visibility that does not depend on traditional forms of intermediation (i.e.: old media, labour unions, political parties, churches, etc.). In many countries this new condition is positively fostering bottom-up processes of anti-authoritarian resistance and democratization, but its enduring impact on democratic institutions is still uncertain. In this fashion, the impact of technology on public discourse has effectively opened a new layer of horizontal transparency and direct exchange not just

[8] See B. Manin, *The principles of Representative Government*, cit.

[9] On the crisis of party politics and the rise of audience democracy, see also P. Mancini, *Il post partito. La fine delle grandi narrazioni*, Il Mulino, Bologna 2015, p. 48.

[10] See B.-Ch. Han, *In the Swarm: Digital Prospects*, MIT Press, Cambridge MA 2017, p. 15.

among citizens who belong to the same political community, but also among individuals all over the world. Social and civil rights movements have taken advantage of this new opportunity to defy existing power structures and enduring injustices on the global stage that the new media offer.

However, there is something fundamentally flawed in the assumption that, as expected by the emancipatory ideal of publicity, the ongoing extension of horizontal transparency will also inevitably result in a substantial increase of vertical transparency, by rendering obsolete old structures of political intermediation and granting the citizens equal access to information and decision-making on public issues. Previous forms of vertical intermediation have indeed been weakened, but they have been soon replaced by new ones, which reshaped the internal workings of the public sphere and had important repercussion on the institutionalized forms of political representation.[11]

The exponential growth of the digital public sphere has been not only fostered but also internally structured by the diffusion of social networks. These platforms operate as technologies of attention and constitute effectively "new media": new forms of vertical intermediation that are less centralized and evident compared to old media, but not less influential. The overabundant and pervasive flow of images and beliefs that fills the internet has, in fact, rendered almost irrelevant the impact of the individual claims that are constantly made on public issues, as they tend to be lost in the iconic ocean of the new media. What counts is now not the ability to fabricate representations, which is abundantly available to almost anyone, bur rather the capacity to orient the public's attention towards certain selected representations as the relevant representative claims. Internet social networks are, in this sense, great technologies of attention, built to train the attention of a vast public and then resell it for commercial or political purposes. Data about users are collected and sold, users are profiled and nourished specific forms of content and advertisement, echo chambers are formed where only like-minded individuals interact with one another.[12] The promise of 'transparent immediacy'[13] that comes with these digital platforms obscures the fact that they are not neutral, but 'programmed' environments that

[11] On the distinction between horizontal and vertical transparency see G. Lingua, *Transparence numérique et frontières de la désintermédiation politique*, in J. Bodini, M. Carbone, G. Lingua, G. Serrano (a cura di), *L'avenir des écrans*, Éditions Mimésis, Parigi 2020, pp. 193-205 and M. Carbone, G. Lingua, *Toward an Anthropology of Screen. Showing and Hiding, Exposing and Protecting*, Palgrave Mcmillam, London 2023, pp. 113-118.

[12] Among many others, see Y. Benkler, R. Faris and H. Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*, Oxford University Press, Oxford 2018; J.A.G.M. van Dijk, K.L. Hacker, *Internet and Democracy in the Network Society*, Routledge, London and New York 2018; J.P. Wihbey, *The Social Fact: News and Knowledge in a Networked World*, MIT Press, Cambridge MA 2019; S. Zuboff, *The Age of Surveillance Capitalism. The Fight for a Human Future at the New Frontier of Power*, Profile Books, London 2019; A. Jungherr, G. Rivero and D. Gayo-Avello, *Retooling Politics: How Digital Media Are Shaping Democracy*, Cambridge University Press, Cambridge 2020.

[13] See J.D. Bolter, R. Grusin, *Remediation. Understanding New Media*, MIT Press, Cambridge MA 1999, pp. 21-31.

respond to a series of choices that determine their structure and influence their usage. These choices are dictated by economic interests, commercial strategies, national policies, and technological options that end up governing from the inside the online landscape. The public is often scarcely aware of this internal opacity of the new media, while the logic behind their everyday functioning is in fact governed by narrow elites and subject to different sorts of manipulation and regulation.[14]

These new diffused forms of organization of the public sphere heavily impact the political practice and determine the emergence of new phenomena: institutional sources of information that are increasingly marginalized, political leaders that constantly and directly address their base, movements who advocate forms of real time digital consultation as the new frontier of democracy. Some of these phenomena are particularly worrying, such as in the case of authoritarian regimes that effectively transform the online social networks in systems of control over their population or use them to influence the democratic process in other countries. In this multifaceted transformation of political practice, social media emerge as crucial intermediaries for building and highlighting relevant representative claims in the new digitally oriented public sphere.

As a result of this combined process of horizontal disintermediation and vertical re-intermediation, the distinction between formal and informal public sphere has been not only effectively weakened, but specifically contested, often at the hands of populist movements. The normative value of this distinction relies on the assumption that a thick layer of intermediation is fundamental for proper democratic decision-making to happen. In this perspective, the formal public sphere needs to be separated from the informal one for rational deliberation to happen according to principles of fairness and reciprocity but needs also to be connected with it by a nurturing relation, for political representation to feed the decision-making process with the actual beliefs and wants of the citizenry. The increasing horizontal disintermediation, however, has led to the contestation of the separation between the two: the gap between formal and informal public discourse has become narrow and the political polarization is colonizing the space of informal public discourse. In this sense, the attention of the public opinion is increasingly directed to a political debate that develops on social media and outside of its institutional sites. At the same time, even political discourse articulated in institutional sites is often modeled after informal discussion, thus rendering effectively obsolete traditional concerns about the appeal to carefully defined boundaries of public reason or the exclusion of confessional religious language. New forms of re-intermediation favor also new models of political leadership: the populist leader borrows codes and rhetoric from the informal public sphere and relies on the delegitimization of traditional intermediators like institutional figures and intellectual elites.[15]

[14] See M. Flyverbom, *The digital prism*, Cambridge University Press, Cambridge UK 2019. pp. 1-24.
[15] M. Barberis, G. Giacomini, *La neo-intermediazione populista. Popolo, istituzioni, media*, in «Teoria politica», n. 10, 2020, pp. 317-340.

In the end, getting radically away from all forms of intermediation in politics is a highly problematic utopian project that aims at bypassing established forms of power imbalance but in practice ends up exchanging old forms of intermediation and power with new ones.

3. The impact of political disintermediation on the representative function of parliaments

This transformation of the political landscape under the pressure of disintermediating ideologies and practices is having important consequences on the inner workings of representative democracy and specifically on the role of parliaments as its central institutions. From the crisis of political parties and other representative bodies, comes a destabilization and liberalization of representative claims that as been labelled as 'hyper-representation': a political phenomenon that sees the flourishing of social actors that occupy the public sphere, claim an immediate relationship with important parts of society and adopt strategies of plebiscitary leadership and direct democracy.[16] We can appreciate the impact of this phenomenon on institutionalized political representation by examining the rise of a twofold practical and conceptual shift: (i) emerging practices of political representation that increasingly marginalize the representative function of parliamentary assemblies in favor of forms of direct relation with the public; (ii) a new normative understanding of political representation as a mirror reflection that is the premise upon which these new practices are established. Let us consider in turn these two interconnected aspects of the ongoing transformation.

(i) The populist quest for a disintermediated direct relation between "the people" and the sites of political decision-making takes forms that partially differ from movement to movement, also depending on to the components that shape the ideology of each specific group. Among these strategies, we consider most notably two: (a) The charismatic leader strategy, focused on the increased role of a prominent national figure that directly address 'the people' at the expenses of locally elected representatives that interact with their specific constituency;[17] (b) The technopopulist strategy, focused on digital direct democracy and consultation presented as an

[16] See A. Mastropaolo, *I partiti, la rappresentanza e la loro pretesa crisi*, in «EticaEconomia», n. 15, 2015; A. Mastropaolo, *Rappresentanza, partiti, governance*, in R. Sau (ed.) *La Politica. Categorie in questione*, Franco Angeli, Milano 2016, pp. 209–219.

[17] It is interesting to note that social media platforms play a decisive role in re-structuring the public conversation away from the local relationship between representative and constituency: "Social media platforms make it easier for the like-minded to socialize from their home environments and over great distances because digital technology facilitates geographically spread niche networks based on interest rather than location. So where mass media consumption to a larger extent is bound to geographically defined communities, social media platforms are bound to communities of peers and like-minded others" in U. Klinger, J. Svensson, *The emergence of network media logic in political communication: A theoretical approach*, in «New Media & Society», n. 17, 2015, pp. 1249-1250.

alternative form of political decision-making that can, in the long run, severely limit or even entirely by-pass the need for elected assemblies.

In the first case, (a) the leader is presented as an outsider, opposed to the established political elites and acting as the embodiment of the demands of "the people", to which they claim to have clear and unmediated access.[18] Usually men, these leaders supposedly act and speak like the people they represent, contest the authority of experts, and maintain that they are bringing the voice of the people inside political institutions that are otherwise close, unclear, opaque. The representations offered by the populist leaders are importantly self-representations as honest and simple persons that identify themselves with a homogenous people in order to contrast the unduly influence of the elites. The direct, apparently un-sophisticated and un-mediated manners of this self-representation are crucial to differentiate the populist leader from his adversaries: much of this strategy relies on the "populists' exposure of one particular aspect of mainstream politicians' behaviour that elites would wish to keep invisible: the constructed nature of their visible performance".[19] The leader is not just in a political relationship with the people as their representative: he embodies the people he represents and his legitimacy draws from his ability to constantly reshape his own image to reflect the represented, to show himself in tune with the sentiments of the population to mobilize their support.[20] The alleged authenticity of the leader's claim is supported by the corresponding reactions of the audience[21] and is presented in stark contrast with the manufactured and stale communication of the adversary.[22] Within this strategy, the role of representatives in

[18] See H. Kriesi, *The Populist Challenge*, in «West European Politics», 37, (2014), pp. 361-378; B. Krämer, *Populist online practices: the function of the Internet in right-wing populism*, in «Information, Communication & Society», n. 20, 2017, pp. 1293-1309.

[19] L. Sorensen, *Populist communication in the new media environment: a cross-regional comparative perspective*, in «Palgrave Communications», n. 4, 2018.

[20] R.R. Barr, *Populism as a political strategy*, in C. de la Torre (ed.), *Routledge Handbook of Global Populism*, Routledge, New York, 2019, pp. 44-56. Significant evidence supports the efficacy of this strategy and shows a correlation between the constant presence of the leader's messaging on old and new media and his approval levels, see G.J. Love and L.C. Windsor, *Populism and Popular Support: Vertical Accountability, Exogenous Events, and Leader Discourse in Venezuela*, in «Political Research Quarterly», n. 71, 2018, pp. 532-545; G. Bobba, *Social media populism: features and 'likeability' of Lega Nord communication on Facebook*, in «European Political Science», n. 18, 2019, pp. 11-23.

[21] "Arguably, social media contribute to dramatising populist communication because they are platforms suited to producing emotional, controversial, even violent contents typical of much populist activism, and to stimulating a 'remix' activity, a creative collage of video clips, sound bites, clickbaits, graffiti, parodies, memes, and many other contents, including insults and fake-news, that can prove crucial in boosting the popularity of the leader, of his/her creed, of his/her movement" in G. Mazzoleni, R. Bracciale, *Socially mediated populism: the communicative strategies of political leaders on Facebook*, in «Palgrave Communications», n. 4, 2018.

[22] In this sense, "[p]opulism is related to a destabilisation of the norms of mainstream politics, not least when it comes to language use. To violate the norms and conventions in the language of politics is a way to perform being anti-establishment" M. Ekström, A. Morton, *The Performances of Right-Wing Populism: Populist Discourse, Embodied Styles and Forms of News Reporting*, in M. Ekström, J. Firmstone

parliaments is only marginal: the assembly serves as an audience for the leader's claims, which are legitimate because of the alleged direct endorsement of 'the people' and need only to be procedurally translated into the legislative process.[23] At its core, the charismatic leader strategy introduces a mutation of the representative relation between representatives and represented, thus sidelining the role of those who still operate within the 'previous' logic of party politics and elected assemblies. As insightfully noted by Camil Ungureanu and Alexandra Popartan, in this sense populism can be examined as a specific political narrative that borrows from mythical and religious repertoires to present the leader as a messianic figure that defies ordinary political logic:

> Although the leader can be elected according to democratic procedures, the relationship with the electorate pertains not to the logic of representation through deliberation and general rules but to that of emanation. According to this logic, the leader as 'natural' emanation of the people has a privileged and immediate access to their interests and needs; the leader is the incarnation of the voice of the people. As such, the leader is not bound by general rules, but is a 'trickster' who transcends them. He places himself above democratic procedures and the basic moral norms of the interaction in the public sphere. As a corollary, the political party becomes a tool or an accessory at the service of the leader who has direct access to the masses through Twitter, Facebook or TV.[24]

In the second strategy, (b) the implementation of digital democracy tools aims at using information technologies to produce a real-time showcase of the genuine will of "the people": according to this view, the institutions of representative democracies are increasingly obsolete, and they are to be substituted by pervasive practices of direct democracy and citizens' consultation. In this perspective, the open and always accessible technological platform is the ultimate promise of getting away from the need of political intermediation, as every citizen will soon be able to directly express their own will on all matters. The role of representatives in parliaments in this process is only secondary and temporary: insofar as digital democracy is realized, the elected parliamentarians are at best conduits for the will expressed online by the people to be

(eds), *The Mediated Politics of Europe: A Comparative Study of Discourse*, Palgrave Macmillan, Cham 2017, p. 293.

[23] For an interesting analysis of how the rise to prominence of populist movements in Italy determined a further marginalization of the parliamentary institution, see: C. Fasone, *Is There a Populist Turn in the Italian Parliament? Continuity and Discontinuity in the Non-legislative Procedures*, in G. Delledonne, G. Martinico, M. Monti and F. Pacini (eds), *Italian Populism and Constitutional Law. Strategies, Conflicts and Dilemmas*, Palgrave Macmillan, Cham, 2020, pp. 41-74. Among several elements that concurred to that end, it is worth noting: the attack to the principle of free mandate, the erosion of parliamentary procedures and immunities, the use of social media to sabotage ongoing political negotiations or to direct the attention of the public away from parliamentary discussions, the use of committees of enquiry and parliamentary questions outside of their institutional boundaries as part of the populist communication strategy.

[24] C. Ungureanu, A. Popartan, *Populism as narrative, myth making, and the 'logic' of political emotions*, in «Journal of the British Academy», n. 8, 2020, p. 43.

translated into their vote in the legislative assembly. In this sense, the technopopulist strategy develops a meta-discourse where the specific contents of the digital consultations are flexible and generic, to mobilize disillusioned citizens coming from different ideological backgrounds: the main theme is not the specific issue at stake, but the general promise of giving them direct control over the political decisions.[25] Like the charismatic leader seeks his own legitimization through the opposition to the cold and distant elites, similarly the legitimization of the technopopulist strategy is achieved through a parallel delegitimization of the traditional forms of political representation.[26] In this sense, the technopopulist forms of citizens' involvement fundamentally differ from the deliberative ones, which aim at the participation of citizens into inclusive processes where the focus is on collective discussion and the eventual mediation among different claims. Here, instead, "[t]he myth of online direct democracy is an outcome of direct democracy […] It is considered an opportunity (mainly arising from democratic participation platforms) to develop 'real' direct democracy (online) at a low cost and without party interference".[27] For the most part this strategy is incarnated by practices of quite limited scope, as in the case of local referendums on specific questions or as internal consultations among the members of the populist movement to support or decline a proposal whose terms have been previously framed and formulated by the leadership. The relevance of these practices within the strategy is mostly symbolic, as a utopian anticipation of a future when these practices could entirely substitute the logic of representation, and rhetorical, to confirm the cohesive identification of the base with the leadership.

(ii) These emergent strategies are consistent with a conceptual understanding of political representation as a mirror of the public and tend to decisively sideline the role of elected assemblies, whose functioning and procedures fail to approximate the standards of immediacy and transparency dictated by the ideal of a perfect mirror reflection. The representative claims raised within these populist strategies are formulated so as to translate this understanding into a political practice. By borrowing some basic elements from Michael Saward's analysis of the representative claim,[28] we can characterize the populist claims as conflating the claim-maker with both the

[25] L. Manucci, M. Amsler, *Where the wind blows: Five Star Movement's populism, direct democracy and ideological flexibility*, in «Italian Political Science Review», n. 48, 2018, pp. 109-132.
[26] If digital disintermediation and the transparency it brings about are the main political message, the natural antagonist are all those traditional forms of intermediation (including parliamentarians and all elected representatives) who still belong to a previous political order that, because of its opaqueness, is now superfluous if not outright damaging. See G. Bobba, G. Legnante, *Italy. A Breeding Ground for Populist Political Communication*, in T. Aalberg, F. Esser, C. Reinemann, J. Strömbäck, and C.H. de Vreese, *Populist Political Communication in Europe*, Routledge, New York, 2016.
[27] E. De Blasio, M. Sorice, *Populisms among technology, e-democracy and the depoliticisation process*, in «Revista Internacional de Sociología», n. 76, 2018, p. 10.
[28] M. Saward, *The Representative Claim*, Oxford University Press, Oxford 2010.

audience and the object of the representation.[29] While the specific subject that is put forth, as we have noted earlier, may be quite elusive when it comes to its contents (from taxation and immigration, to welfare policies and international relations), what lies at the core of the claim is that the entire audience is construed as a homogenous whole ('the people') that is perfectly reflected by the claim-maker (the charismatic leader or the digital democracy platform) and this disintermediated identification is the actual object of the claim itself. The claim is really about the mirroring correspondence between 'the people' and the leader or the digital platform, which is qualitatively different from the representative relation attributed to other claim-makers, like traditional political parties and even elected representatives. The subject that is put forth is just an occasion to highlight this correspondence.[30]

More specifically, in the case of (a) the populist leader, he looks and talks like his audience, 'the people', and in turn the core content of the claims he makes is, again, 'the people' as opposed to 'the elites'. The specific subject of each claim may change, but the impact of the claim does not depend on it. The internal logic of the charismatic leader strategy is that of an oxymoronic 'direct representation':[31] the leader represents the people by embodying the people with his demeanor and rhetoric and by making the people the central object of what he claims, through a strategy of continuous communication with his audience to prove their mutual identification. In this sense, the legitimacy of the leader is not rooted in democratic procedures, but in his ability to embody a mirror reflection of the people, in contrast with 'the establishment'.[32]

In the case of (b) digital direct democracy practices, the identification of claim-maker, audience and content of the claim is made through the real time technological mirror. 'The people' can finally take the decision in its own hands instead of waiting for someone else to represent it in the decision-making process, and this is also the main content of the claim that is made: the fact that disintermediation is achieved through a platform that merely reflects the will of the people, so that it can then be procedurally applied to the policy- and law-making process.

In the light of this mirror-like understanding of representation, the notion that elected assemblies serve as intermediate political expressions of a certain society, and thus formulate representative claims at a highest degree of legitimacy, appears outdated. Even if one accepts that the electoral system provides a procedurally made

[29] Saward's complete formula for the 'general form of the representative claim' goes as follows: "[a] maker of representations ("M") puts forth a subject ("S") which stand for an object ("O") that is related to a referent ("R") and is offered to an audience ("A")" Saward, *The Representative Claim*, cit., p. 36.

[30] For an interesting analysis that highlights this functioning of populist representative claims on the subject of taxation in the United States and Canada, see D. Laycock, *Tax revolts, direct democracy and representation: populist politics in the US and Canada*, in «Journal of Political Ideologies», n. 24, 2019, pp. 158-181

[31] N. Urbinati, *Political Theory of Populism*, in «Annual Review of Political Science», n. 22, 2019, p. 120.

[32] One of the most influential recent accounts of representation in terms of hegemony and embodiment is offered in E. Laclau, *On Populist Reason*, Verso, London and New York 2005.

'photograph' of the constituency at a certain point in time, its validity will still be derivative if compared to the promise of a real time mirror, especially within a digital public sphere that is constantly filled, in the minute-by-minute experience of the public, by an everchanging flow of images and bits of information. Therefore, the ideal of disintermediated representation is the mirror: if the represented can see their own image constantly reflected in overabundant visibility of the leader or in the promise of real time digital voting, the gap of intermediation is closed.[33]

The idea of representation as a mirror reflection is consistent with the logic of the Schmittian principle of identity, of which it essentially offers an updated application within the contemporary political space. In his *Constitutional Theory*, Carl Schmitt states:

> […] the people can achieve and hold the condition of political unity in two different ways. It can already be factually and directly capable of political action by virtue of a strong and conscious similarity, as a result of firm natural boundaries, or due to some other reason. In this case, a political unity is a genuinely present entity in its unmediated self-identity. This principle of the self-identity of the then present people as political unity rests on the fact that there is no state without people and that a people, therefore, must always actually be existing as an entity present at hand. The opposing principle proceeds from the idea that the political unity of the people as such can never be present in actual identity and, consequently, must always be represented by men personally.[34]

The disintermediated digital public sphere now seems to offer a chance to drastically reduce the distance between the "unmediated self-identity" of the people and its embodied presence in the forms of political representation. In a political setting where old and new media ensure a space of ubiquitous public visibility, the populist promise is that self-identity can be made present in the form of a mirror image of 'the people' that is reflected back to the audience in real time; the gap between the principles of identity and representation has never been so narrow. However, as Nadia Urbinati has appropriately pointed out, this contemporary Schmittian revival is as much at odds with a genuinely pluralist representative democracy as its original version was:

> Clearly, since Schmitt thought of representation as a synthesis of identity and the presence of the sovereign, party pluralism and parliamentary competition were anathema to him. […] In similar manner, populism uses representation to constitute the political order above the society and through the expulsion of pluralism. As per Schmitt, who thus gave populism an important argument, representation is political insofar as it repels the liberal calls of advocacy, control, monitoring, and a constant dialogue between society and politics, and narrows the distance between the elected leader and the electors so as to incorporate society within the state.[35]

---

[33] See, M. Carbone, G. Lingua, *Toward an Anthropology of Screen*, cit., pp. 120-121.
[34] C. Schmitt, *Constitutional Theory*, translated and edited by J. Seitzer, Duke University Press, Durham and London 2008, p. 239.
[35] Urbinati, *Democracy disfigured*, cit., p. 137.

Before we move to a more comprehensive critical assessment of this populist understanding of representation as a mirror image, it is useful to note that the practices we examined, and their underlying conceptual premises, stem out of a justified awareness of the increasing difficulties of institutionalized representation is a rapidly changing landscape. An apt example is the long running debate on deliberative democracy and its prospects as a model to reform stale party-based systems of representation. In this sense, in political theory, discontent with the state of representative democracy and the excessive distance between representatives and represented has found abundant theoretical articulation in the past few decades, with a prominent focus on the notion of deliberative democracy and the conditions of the participation of citizens to political decision-making.[36] According to deliberative democracy theorists, the intrinsic limitations of representative democracy have been further aggravated by the rise of technocratic approaches to public management, the growing gap between elites and the general public, and the loss of sovereignty due to the increasing influence that international institutions and global markets have on national societies. To contrast the citizens' apathy and the delegitimization of democratic rule spurred by these factors, the answer would be to recover the dimension of direct participation to the political decision-making by rendering available deliberative practices and open forums. In this perspective, both the formal and the informal public spheres are essentially argumentative and deliberative spaces that need to be re-connected by establishing appropriate discursive practices.

The recent push towards disintermediation, however, seems to by-pass this strategy rather than support it. Citizens' participation and collective deliberation ideally aim to address the same gap between representatives and represented that populists focus on, but as practices of bottom-up political engagement they still entail, in important ways, complex and sometimes taxing forms of intermediation. The push towards disintermediation has thus shifted the focus from the open process of argumentation and deliberation to a more immediate ideal of political representation: the real time mirroring reflection of the public, as it supposedly is, with its own needs and wants. In this perspective, the prospect of a fruitful connection between the deliberative interaction among citizens and the deliberative function of parliaments is marginalized. The representative function remains alone at the center of the political stage, and it seems to be better served outside of elected assemblies and their procedures rather than inside.

[36] Among others, see: J.S. Dryzek, *Deliberative Democracy and Beyond: Liberals, Critics, Contestations*, Oxford University Press, Oxford 2000; M. Saward (ed.), *Democratic Innovation: Deliberation, Representation and Association*, Routledge, London and New York 2003); A. Gutmann, D.F. Thompson, *Why Deliberative Democracy?*, Princeton University Press, Princeton 2004; S. Besson and J.L. Martí (eds), *Deliberative Democracy and Its Discontents*, Ashgate, Burlington VT 2006.

4. *Critical assessment of the impact of political disintermediation on the representative function of parliaments*

Both the deliberative and populist critiques we considered, although fundamentally different, highlight important difficulties that institutionalized forms of political intermediation have been recently facing in a landscape of increasing social disintermediation. However, both the (i) emerging populist practices of political representation and the (ii) normative understanding of political representation that underpins them can be subject, in turn, to a profound scrutiny that underlines their fundamental weaknesses. Let us address them in turn.

(i) Populist representative strategies rely on a radical pretense of political disintermediation that they cannot in fact realize. As we have seen, populist representative claims conflate the claim-maker with both the audience and the object of the representation, but this identification of the three elements is fictious and ultimately incompatible with the pluralism of claims that is typical of democratic systems.

In the case of (a) charismatic leaders, far from simply mirroring their average constituent, they constantly shape and resell themselves as a representation of the beliefs and wants of society, thus actively fabricating the collective identification with such a representation. This 'shape-shifting representation' is, to some extent, a structural component of representative democracy, especially in settings that have the prevalent traits of an audience democracy.[37] But the artificial nature of this identification of the leader with 'the people' is fundamentally incompatible with the alleged disintermediated and direct nature of their relationship, upon which the validity of the leader's claim is grounded. As Lone Sorensen has observed by thoroughly examining the representative strategies of UKIP in the United Kingdom and the EFF in South Africa:

> Despite denying that they adapt their practices to the media, these movements engage in disruption of political norms that catches the media's attention and lends them control of both their own and the elite's visibility. The two populist cases thus address the challenges of the paradigm of visibility through entrepreneurial forms of meta-performance, designing their own performances to expose the crafted and crafty nature of elite visibility management.[38]

The widespread use of social media, often portrayed as the epitome of direct interaction and communication between leaders and people, is in fact fabricated to provide an illusion of immediacy in both content and practice. In terms of content, the thin ideology of populist message is kept simple, ambiguous and malleable, to

[37] M. Saward, *Shape-Shifting Representation*, in «American Political Science Review», n. 108, 2014, pp. 723-736.
[38] L. Sorensen, *Populist communication in the new media environment*, cit.

render easy for a large audience to identify with.[39] The very public performance of the leader is not even his own, for the most part: social media interactions are documented to be heavily performed by parliamentary assistants and communication teams rather than by elected representatives and leaders themselves.[40] The disintermediated identification of the leader with his audience and object of the representation is, thus, for the most part, a fiction built upon the actual mechanics of how the populist claim is formulated and diffused.

In the case of (b) digital direct democracy platforms, they are presented within a promise of absolute transparency, immediacy and self-rule, but are in fact often governed by less accountable forms of intermediation than traditional electoral processes. Therefore, barriers of access to the consultation and tight control over the timing, narrative and information provided to the participants play a decisive role in determining the outcome of these voting and consultation practices. Multilevel analyses of citizens' orientations in direct democracy votes show that political elites still play a decisive role in providing the citizens with signals and information that are crucial to the formation of their ability to make a competent choice.[41] Moreover, significant evidence indicates that in direct votes on specific issues, voters' factual beliefs on policy issues can become systematically distorted to align with their pre-existing cultural and political orientations[42] and that voters tend to align their arguments with their preferred party's position.[43] In this sense, instances of digital direct democracy, especially when restricted to consultations among movement and party members, do not provide a disintermediated image of 'the will of the people', as if it emerged in a vacuum to be reflected by the technological platform. On the contrary, they rather register an orientation that has been inevitably and profoundly affected by a long history of interactions with the intermediation of political elites, media infrastructures, cultural leaders, and religious authorities, which all played a significant part in creating the conditions of the citizens' choice. Proponents of digital direct democracy suggest that some of these shortcomings are only due to the limited nature and diffusion of these early practices, however it is also questionable that the project of rendering such practices pervasive would actually result in an improvement of the citizens' participation: a purely direct democratic regime, in fact, "requires that

[39] See N. Ernst, S. Engesser, F. Büchel, S. Blassnig and F. Esser, *Extreme parties and populism: an analysis of Facebook and Twitter across six countries*, in «Information, Communication & Society», n. 20, 2017, p. 1359.

[40] See C.S. Ben-Porat, S. Lehman-Wilzig, *Electoral system influence on social network usage patterns of parliamentary assistants as their legislators' stand-in: The United States, Germany, and Israel*, in «New Media & Society», n. 5, 2020, pp. 1022-1044.

[41] See C. Colombo, *Justifications and Citizen Competence in Direct Democracy: A Multilevel Analysis*, in «British Journal of Political Science», n. 48, 2018, pp. 787-806.

[42] See J. Gastil, J. Reedy, and C. Wells, *Knowledge Distortion in Direct Democracy: A Longitudinal Study of Biased Empirical Beliefs on Statewide Ballot Measures*, in «International Journal of Public Opinion Research», n. 30, 2017, pp. 540–560.

[43] See C. Colombo, H. Kriesi, *Party, policy – or both? Partisan-biased processing of policy arguments in direct democracy*, «Journal of Elections, Public Opinion and Parties», n. 27, 2017, pp. 235-253.

the public agenda be broken down into discrete issues that are voted on separately. This further undermines reasonable democratic deliberation"[44] as it prevents each time the specific issue at stake to be considered, debated and decided upon within a general framework that also includes other relevant issues that are systemically connected with it.

By looking closer at the (ii) conceptual underpinnings of these populist representative strategies, we argue that construing political representation as a real time mirror of the public is a misleading premise that is also at the basis of the practical faults we just illustrated.

Because of their common conceptual foundations, both the (a) charismatic leadership and (b) technopopulist strategies have in fact strong anti-pluralistic implications. The underlying logic of the populist representative claim is conflating the claim-maker with their audience and the object of the claim, but this effectively excludes the legitimacy of other representative claims: if the maker of the populist claim is identical with the audience and their correspondence is the actual object of the representative claim, anyone who makes a different claim is necessarily an impostor. If the leader is a direct representation of 'the people', it means he speaks like 'the people', on behalf of 'the people' and in doing so he re-instates the people in its legitimate position of sovereignty. Should this be true, different claims are by default to be considered illegitimate, as anything the people is and wants has been already mirrored. Similarly, if the digital platform allows 'the people' to speak as they want, directly on behalf of themselves, thus effectively re-instating themselves in their position of sovereignty, any different claim, even made by legally elected representatives, becomes secondary if not outright meaningless.

This mirroring framework "plays into the populist ideology that there is a single collective will that can be represented in its entirety, and is therefore fundamentally at odds with the view of representative democracy as pluralistic".[45] The irreducible plurality of people's beliefs, wishes, and needs cannot be reflected in any single mirror image, but has to be articulated through a multitude of competing representative claims that rather operate like portraits that differ substantially according to the author and with which the represented are engaged in an active process of recognition and critique, identification and rejection.

In the public sphere, mechanisms of vertical re-intermediation manage the flow of attention and thus render certain portraits more likely to be successful, certain representations more prone to become culturally and political hegemonic. Populist leaders and movements capitalize on this aspect of contemporary public sphere,

---

[44] A.J. McGann, *The Logic of Democracy: Reconciling Equality, Deliberation, and Minority Protection*, The University of Michigan Press, Ann Arbor 2006, p. 128.

[45] R. Van Crombrugge, *Are referendums necessarily populist? Countering the populist Interpretation of referendums through institutional design*, in «Representation», n. 1, 2020, p. 110. This critique has also important consequences on the understanding and design of referendums, which, like populist claims, may be erroneously construed as a perfect mirror image of the will of the people at a certain point in time.

sometimes even substantially by-passing and antagonizing institutionalized procedures and forms of political representation.[46]

Elected assemblies, on the other hand, fully enact their representative function by embracing the pluralist and portrait-like nature of political representation through their internally diverse composition and the free mandate of their members. Parliamentary representative practices are to be assessed based on how they establish and constantly enact this kind of interpretive relationship, rather than to how closely they 'mirror' an alleged image of their constituency. Political representation, in this sense, is a process that always includes some relevant degree of interplay between representatives and represented. This mutual interpretive relation remains healthy insofar as all parties involved accept that the gap between the representation and what is represented cannot be entirely dissolved into the immediacy of a perfect reflection, but it is rather the space of difference and change.

To sum up, the ideal of representation as a mirror is a promise of radical disintermediation that covers new forms of hidden intermediation and unjustifiably delegitimizes the democratic pluralism of representative claims.[47] If the representation is a reflection in a mirror identical with the object itself, then there is no room for other representations: what counts is the mirror – the leader or the digital platform – which at different times may show different images – various claims, elusive and sometimes even in conflict with each other – whose validity is exclusively granted by the mirror itself. The representative relation as a mirror reflection, however, is a mere

[46] The relationship between populist communication and the attention economy of social media is twofold: on one side, the social networks are convenient to the populist movements, as they present an apparently direct medium that is in tune with their political message of disintermediation, but in turn the populist communication is convenient to the inner workings of the social networks: "In terms of online opportunity structures, the concept of attention economy implies that attention is a scarce resource over which information providers have to compete. On the Internet, this competition is particularly fierce due to the abundance of content. Therefore, the Internet favors content that 'maximizes attention'. The populist style of simplification, emotionalization, and negativity increases our attention by addressing fundamental perceptual patterns and news values. Therefore, populism is particularly well-suited to be communicated online" in S. Engesser, N. Fawzi & A.O. Larsson, *Populist online communication: introduction to the special issue*, «Information, Communication & Society», n. 20, 2017, pp. 1285-6.

[47] It is worth mentioning that not all calls for disintermediation or critiques to the role of established elites are necessarily anti-pluralistic or un-democratic: "On the one hand, populist support for direct democracy is found to reflect confidence in the virtuous character of ordinary people (in contrast to politicians), which is usually associated with more participatory democracy. On the other hand, citizens with populist attitudes are portrayed as preferring to play an 'essentially passive' role and favouring a 'responsive government, i.e. a government that implements policies that are in line with their wishes', rather than more participatory forms of democracy. We argue that this ambiguous description of populist preferences in studies of populism is due to a conflation of populist and stealth-democratic attitudes. While citizens sharing either political-ideological orientation reject elite rule and would prefer (more) direct democracy, their motivation differs fundamentally" in S. Mohrenberg, R.A Huber, T. Freyburg, *Love at first sight? Populist attitudes and support for direct democracy*, in «Party Politics», n. 3, 2021, p. 529.

fiction. Political representation always comes with some normative limits based on which we can assess and compare the validity of representative claims or distinguish which acts of political person are legitimately representative and which are not. Representation as a mirror reflection is an attempt to get away from these limits, based on the assumption that the mirror – the charismatic leader or the digital platform – can only reflect, because of its intrinsic properties, a perfect image of what has to be represented. Without limits, though, the representative claims become non-pluralist and unaccountable, as there are no external standards based on which the claims can be assessed and compared with others. As Howard Schweber has argued:

> Political representation names a relationship among actors who have the capacity to engage in relationship of authorization in accordance with the norms of a representative map. The activities of a representative may include advocacy, deliberation, mechanisms of accountability, or mediation, and they may take place in the context of formal or informal institutional settings. The limits of political representation, however, exclude activities or relationships that go beyond the limits of political representation. Representation in its political conception is inherently normative, implicating standards for both legitimation and legitimacy as the basis for contestation, critique, or analysis. A substantive political conception of representation is a necessity for either normative or empirical analysis of representative claims and practices.[48]

In this sense, in political representation there are no real time mirror reflections, but only multiple representative interpretations. The representative claim-maker is not identical with the object or the audience of the claim: these elements are distinct and are variously connected by different possible interpretations that lead to different claims. Because of this irreducibility of the representative claim to a single element, the democratic pluralism of claims is always possible and different representative claims are legitimate. The representative process enacted by elected assemblies is such a form of active interpretation of society, both in the cognitive and performative sense: this interpretive nature of parliaments should not be hidden or sidelined, but rather rendered more visible, open and accessible.

5. Conclusion: beyond mirrors, back to the bond between representatives and represented

Social disintermediation is indeed seriously questioning the traditional understanding and practice of institutionalized representative democracy. These social transformations are deeply intertwined with technological innovations and are unlikely to be reverted in the foreseeable future. In this sense, it is imperative for institutional sites of democratic representation to take on a path of reflection and reform that takes the ongoing tectonic shifts into account. However, in the light of the two lines of critical assessment we presented, practical and conceptual, we

[48] H. Schweber, *The Limits of Political Representation*, in «American Political Science Review», n. 110, 2016, pp. 394-395.

maintain that the populist strategies of radical disintermediation of the constitutional democratic representative system are fundamentally flawed answers to a justified call for renovation.

Instead of following a misleading infatuation with a self-defeating ideal of total disintermediation, we suggest to rather look at how innovative forms of political intermediation, through means of popular accountability and direct participation, can help the responsiveness of our existing model of political representation through elected assemblies.[49] Experiences like public hearings[50] and citizen-initiated mechanisms of direct democracy[51] seek not to sideline or eliminate the role of elected representatives, but rather to reshape the way elected assemblies operate with the integration of new procedurally regulated avenues through which the public can directly participate in articulating the relevant questions, formulating their own representative claims, contributing to the deliberative exchange among the representatives, and eventually directly intervene in the policy-making process on the most important issues. Public hearings open-up the deliberative role of elected assemblies to include the voice citizens inside their own workings, thus translating into fair institutional practices the ongoing thinning of the boundaries between formal and informal public sphere.[52] On the other hand, citizen-initiated mechanisms of direct democracy substantiate the possibility for the citizens to directly formulate their own representative claims and bring them into the democratic system through appropriate practices that allow all voices to be heard instead of taking plebiscitarian shortcuts.

These kinds of emerging democratic practices embrace the inevitably interpretive and intermediated nature of representative claims but try to re-focus the interpretive process around the ongoing active listening of the representative and the participation of the represented rather than on the discrete mechanism of periodic delegation or the fictious disintermediation of the populist mirror.

---

[49] See G. Katsambekis, *The Populist Surge in Post-Democratic Times: Theoretical and Political Challenges*, in «The Political Quarterly», n. 88, 2017, p. 208.

[50] See R. Eising, F. Spohr, *The More, the Merrier? Interest Groups and Legislative Change in the Public Hearings of the German Parliamentary Committees*, in «German Politics», n. 26, 2016, pp. 314-333; C. Moreira de Castro, *Public hearings as a tool to improve participation in regulatory policies: case study of the National Agency of Electric Energy*, in «Revista de Administração Pública», n. 47, 2013, pp. 1069-1087.

[51] See D. Altman, *Citizenship and Contemporary Direct Democracy*, Cambridge University Press, Cambridge 2019.

[52] K.P. Hunt, N. Paliewiscz and D. Endres, 'The Radical Potential of Public Hearings: A Rhetorical Assessment of Resistance and Indecorous Voice in Public Participation Processes', in J. Goodwin (ed.), *Confronting the Challenges of Public Participation: Issues in Environmental, Planning and Health Decision-Making*, (Charleston SC, 2016), pp. 65-79.

Trasparenza e democrazia monitorante. La trasparenza integrale come occasione di partecipazione dei cittadini[a]

Nicola Pedretti[*]

Abstract

La garanzia di una reale trasparenza della pubblica amministrazione rappresenta indubbiamente un'occasione di partecipazione della cittadinanza alla gestione della *res publica*. In quest'ottica, va osservato come la normativa italiana abbia introdotto gradualmente strumenti di accesso civico che hanno reso maggiormente fruibili dati e informazioni ai cittadini. In tale ambito, verrà esaminato il caso studio del rapporto Rimandati che applica il community-based monitoring al delicato settore dei beni confiscati.

Parole chiave: trasparenza integrale, partecipazione politica, beni confiscati

Abstract

The guarantee of real transparency of the public administration undoubtedly represents an opportunity for citizen participation in the management of the *res publica*. From this perspective, it should be noted that Italian legislation has gradually introduced civic access tools that have made data and information more accessible to citizens. In this context, the case study of the Rimandati report will be examined, which applies community-based monitoring to the delicate sector of assets confiscated.

.

Keywords: full transparency, political participation, confiscated assets

[a] Saggio ricevuto in data 31/05/2024 e pubblicato in data 22/01/2025.
[*] Giurista, email: nicola.pedretti@gmail.com.

1. La casa di vetro?

"Dove un superiore pubblico interesse non imponga un momentaneo segreto, la casa dell'amministrazione dovrebbe essere di vetro"[1]. Con queste parole Filippo Turati creava una delle più efficaci metafore con cui è stato descritto il modo di agire ottimale della pubblica amministrazione: la trasparenza nei confronti della cittadinanza doveva assurgere a regola e, pertanto, occorreva che l'apparato burocratico diventasse un involucro totalmente trasparente con un contenuto visibile in qualsiasi momento da parte della comunità interessata.

Quando parliamo di trasparenza occorre essere precisi nell'individuare una corretta definizione per descriverla e differenziarla dal concetto di pubblicità. Mentre quest'ultimo si identifica con la pubblicazione degli atti, affinché si possa fare riferimento alla "casa di vetro", è imprescindibile che essi siano anche denotati dall'elemento della comprensibilità e risultino del tutto fruibili e intelligibili da quanti ne prendano visione[2], in modo che venga loro garantita la possibilità di esercitare un controllo diffuso sull'esercizio dei pubblici poteri e sul buon andamento della pubblica amministrazione[3]. Possiamo quindi concludere evidenziando come il principio della trasparenza amministrativa risulti centrale sotto un duplice profilo: se da un lato è funzionale a garantire imparzialità e buona amministrazione dall'altro risulta anche indispensabile per assicurare eguaglianza e rispetto dei diritti dei cittadini di fronte allo stato[4].

L'importanza della trasparenza amministrativa risulta ormai assodata al punto da essere annoverata tra i diritti umani. Vale infatti la pena evidenziare come le Nazioni Unite abbiano ritenuto opportuno prima inserirla nella Dichiarazione Universale dei Diritti Umani del 1948, all'art. 19, seppur in forma estremamente vaga, e, nel 1998, nel rapporto annuale del Relatore Speciale per la Libertà di Opinione e di Espressione, affermando che *"il diritto di accesso alle informazioni in possesso del governo dev'essere la regola piuttosto che l'eccezione"*[5].

Una simile concezione, tuttavia, non ha trovato storicamente un'immediata e pacifica accoglienza da parte degli apparati burocratici, i quali spesso sono risultati restii a condividere i documenti e le informazioni in loro possesso. In epoca moderna, il primo provvedimento legislativo che portò a sancire il diritto dei cittadini di

[1] F. Turati, Atti del Parlamento italiano, Camera dei Deputati, sessione 1904-1908, 17 giugno 1908.
[2] N. Pedretti, *Comunità monitoranti e trasparenza amministrativa*, in «Scienza e pace», 2, 2020, p. 84.
[3] Cfr. M.C. Cavallaro, *Garanzie della trasparenza amministrativa e tutela dei privati*, in «Diritto Amministrativo», 1, 2015.
[4] E. Carloni, *Alla luce del sole trasparenza amministrativa e prevenzione della corruzione*, in «Diritto Amministrativo», 3, 2019.
[5] E. Belisario, G. Romeo, *Silenzi di stato*, Chiarelettere, Milano, 2016, p. 19. Ivi pp. 19-20 si riporta opportunamente come a livello internazionale tale principio sia ormai pacifico e presente anche in altri importanti documenti quali la Raccomandazione (81)19 sull'accesso alle informazioni detenute dalle pubbliche amministrazioni del Comitato dei Ministri e la Convenzione di Arhus, seppur limitatamente alla materia ambientale.

accedere ai documenti della pubblica amministrazione fu il *Tryckfrihestörordningen*[6]. Questo è stato approvato in Svezia nel 1766 e ancora oggi i suoi principi sono recepiti dalla costituzione di quel paese. Frutto del lavoro di Anders Chydenius, pastore e illuminista radicale, la norma si basava sull'idea di uno stato garante delle libertà dei cittadini e sanciva che tutti i documenti in possesso delle pubbliche amministrazioni dovevano essere immediatamente rilasciati a chiunque ne avesse fatto richiesta[7]. Indubbiamente una pietra miliare della garanzia del principio di trasparenza a livello globale è rappresentata dal FOIA statunitense del 1966. Provvedimento frutto di lunghe discussioni e battaglie politiche[8], approvato nonostante l'aperta ostilità del presidente Johnson, che mantiene ancora oggi alcuni capisaldi che rappresentano un modello globale per gli strumenti di accesso civico: fruibilità totale alle informazioni detenute dal governo a ogni livello, tutela giurisdizionale di tale diritto e onere della prova in capo all'amministrazione nei casi in cui si ritenga che un documento debba restare riservato[9].

Un'analoga visione è emersa recentemente anche a livello europeo, se prendiamo in considerazione il recente *Recovery Plan*, che ha rappresentato una vera svolta nelle politiche dell'Unione Europea. Possiamo notare come da esso emerga un'idea di amministrazione particolarmente attenta *"al rafforzamento dei processi decisionali e della trasparenza, della fiducia e dell'integrità nel settore pubblico, oltre che alla trasformazione digitale del settore pubblico, declinando una funzione pubblica attrattiva e dinamica, adeguata alle nuove sfide"*[10].

Appare chiaro, pertanto, come si stia affermando a livello globale, nonostante alcune resistenze delle amministrazioni, l'idea dell'indiscutibilità del diritto di accedere alla "casa di vetro" da parte di chiunque ne faccia richiesta, sia a tutela di inalienabili diritti umani sia in ragione della necessità di contribuire al buon andamento dell'amministrazione anche mediante l'esercizio di un controllo diffuso da parte della cittadinanza. A conclusione della riflessione introduttiva, pare opportuno evidenziare come non ci si possa limitare a prendere in considerazione la, pur fondamentale, evoluzione della normativa. Se i meccanismi di accesso reattivi e proattivi rappresentano un architrave ineludibile, occorre che siano affiancati da un approccio culturale che permetta di comprendere appieno la natura della trasparenza amministrativa e la sua connessione con la possibilità di favorire la partecipazione attiva dei cittadini. Su questo aspetto si può fare riferimento a quanto scritto da Carloni

> le riforme dei singoli meccanismi vanno legate a due aspetti essenziali. Il primo, è quello (auspicabile, ma in concreto raramente presente) di legare l'introduzione di

[6] Testo inglese reperibile in https://www.chydenius.net/tiedostot/worlds_first_foia.pdf
[7] E. Belisario, G. Romeo, *Silenzi di stato*, *cit.*, pp. 15-16.
[8] Già nel 1956 il deputato John E. Moss propose di introdurre l'istituto dell'accesso civico nell'ordinamento degli USA senza però riuscirci.
[9] Ivi, pp. 21-23.
[10] E. Carloni, *Il dovere pubblico alla trasparenza oltre i diritti di accesso e gli obblighi di pubblicazione*, in «Lo Stato», 18, 2022, p. 25.

> singoli strumenti e misure di trasparenza ad un'idea complessiva di pubblica amministrazione, ad una certa visione del rapporto tra istituzioni e cittadini (e, in ultima istanza, ad una certa idea di funzionamento in concreto della democrazia amministrativa). Il secondo, di considerare che il cambiamento di paradigma dipende dal suo riconoscimento non meno che dalla definizione delle regole che lo supportano: un riconoscimento che è consolidamento di una cultura della trasparenza, che è accettazione della sua portata generale, supporto all'esercizio dei relativi diritti, propensione all'apertura[11].

Una simile visione mostra molto chiaramente come la trasparenza non possa che essere connessa a un'idea di pubblica amministrazione aperta ad un rapporto attivo con la cittadinanza. Un'ottica di *open government* da intendersi come una cultura della governance basata su principi di trasparenza, integrità, *accountability* e partecipazione finalizzata allo sviluppo della democrazia e della crescita inclusiva[12].

2. La trasparenza in Italia.

Nel nostro ordinamento la trasparenza trova una profonda connessione con i principi costituzionali di imparzialità e buon andamento della pubblica amministrazione. Tale assunto è stato recentemente fatto proprio anche dalla Corte Costituzionale che ha avuto modo di affermare come *"i principi di pubblicità e trasparenza, riferiti non solo, quale corollario del principio democratico (art. 1 Cost.), a tutti gli aspetti rilevanti della vita pubblica e istituzionale, ma anche, ai sensi dell'art. 97 Cost., al buon funzionamento dell'amministrazione*"[13] e analogamente si è espresso il Consiglio di Stato sostenendo che *"la luce della trasparenza feconda il seme della conoscenza tra i cittadini e concorre, da un lato, al buon funzionamento della P.A. ma, dall'altro, anche al soddisfacimento dei diritti fondamentali, se è vero che organizzazione amministrativa e diritti fondamentali sono strettamente interrelati"*[14].

Tuttavia, nonostante questi importanti orientamenti giurisprudenziali, per lungo tempo il sistema si è basato sulla segretezza dell'attività amministrativa rispetto alla quale la trasparenza acquisiva unicamente un carattere di occasionalità, a fronte di una discrezionalità molto ampia degli apparati che avevano la possibilità di decidere sostanzialmente a loro discrezione in merito alla fruibilità di qualsiasi cosa fosse in loro possesso[15]. Tale condizione iniziò a mutare gradualmente a partire dal 1990 quando, con la legge 241, venne introdotto per la prima volta il diritto di accesso in capo ai cittadini seppur limitatamente ai casi nei quali sia presente un interesse concreto attuale e diretto connesso ad uno specifico documento[16]. Tale disposizione risulta finalizzata alla tutela di posizioni giuridiche qualificate, solo gradualmente si è

---

[11] *Ivi*, p. 28.
[12] Cfr. Rapporto OCSE, *Open Government: The Global Context and the Way Forward* 2016, p.3.
[13] Corte Costituzionale sentenza n. 20 del 2019.
[14] Consiglio di Stato, Adunanza Plenaria, sentenza n. 10 del 2020.
[15] S. Di Pietro, *La tutela della privacy tra esigenze di trasparenza e nuove regole di riservatezza*, in «Rivista di diritto amministrativo», 9-10, 2018, p.4.
[16] In realtà già la legge 142 del 1990 aveva iniziato a normare la tematica dell'accesso ai dati e alle informazioni in possesso delle pubbliche amministrazioni.

giunti a superare questa impostazione limitata alla tutela di specifici portatori di interessi in favore di un approccio volto all'accessibilità totale. Una prima svolta fondamentale è indubbiamente rappresentata dal dlgs. 33 del 2013 (Decreto Trasparenza)[17], questo già all'art. 1 chiarisce che la trasparenza va intesa come *"accessibilità totale dei dati e documenti detenuti dalle pubbliche amministrazioni, allo scopo di tutelare i diritti dei cittadini, promuovere la partecipazione degli interessati all'attività amministrativa e favorire forme diffuse di controllo sul perseguimento delle funzioni istituzionali e sull'utilizzo delle risorse pubbliche"*. Il citato provvedimento legislativo introduce il cosiddetto accesso civico semplice, in base al quale chiunque può richiedere che una pubblica amministrazione renda disponibili documenti e informazioni in suo possesso, per i quali sia prevista l'obbligatoria pubblicazione, che non sono stati divulgati tramite i canali istituzionali dell'ente. Pur rappresentando un primo fondamentale cambio di paradigma, tale istituto è unicamente un rimedio rispetto all'inosservanza di specifici obblighi di pubblicità e per poter rilevare la definitiva scelta di garantire un'accessibilità generalizzata occorrerà attendere il 2016.

Con il dlgs. 97/16 di quell'anno viene infatti introdotto l'accesso civico generalizzato comunemente noto come il *"FOIA italiano"*. Il nuovo istituto ha introdotto la possibilità di richiedere e ottenere qualsiasi documento o informazione in possesso di una pubblica amministrazione, senza che occorra specificare alcuna motivazione in riferimento all'istanza, purché non sussista una specifica eccezione assoluta o relativa esplicitamente prevista dalla legge. Va, tuttavia, evidenziato come la citata disciplina presenti, in particolare all'art. 5 bis co. 1-2 del dlgs 33/2013[18], importanti limitazioni consistenti nella forte discrezionalità dell'amministrazione che, in assenza di procedure per l'analisi del danno e la valutazione dell'impatto, può rigettare la richiesta per la presenza di un preponderante interesse pubblico o privato costituzionalmente tutelato[19].

In ragione di tali considerazioni, si ritiene opportuno fare una rapida illustrazione degli esiti dei primi anni di attuazione di tale normativa. Il "FOIA italiano" è stato infatti sottoposto a numerose azioni di monitoraggio di natura sia civica sia istituzionale. Ad esempio, possiamo citare il rapporto "Ignoranza di Stato", curato dalla Ong Diritto di Sapere nel 2017, dal quale emersero considerevoli criticità[20]. Appare particolarmente interessante anche il monitoraggio periodico dal Dipartimento della Funzione Pubblica sulle istanze di accesso civico generalizzato rivolte ai ministeri[21], effettuato l'ultima volta in relazione all'anno 2020. A fronte di

[17] Pare, tuttavia, opportuno evidenziare che in realtà era già stata introdotta una modalità di accesso civico che prescindeva dalla necessità della presenza di una motivazione con la legge n. 195 del 2005. Va però specificato che questa trovava operatività unicamente nell'ambito delle informazioni ambientali.

[18] Come modificato dal dlgs. 97/2016.

[19] M. Trapani, *Il diritto di accesso generalizzato e l'emergenza: rischi ed opportunità di uno stato tecnologico*, in «DPCE Online», 3, 2020, pp. 4349-4350.

[20] Si faccia riferimento a https://cild.eu/blog/2017/04/19/ignoranza-di-stato-il-difficile-accesso-alle-informazioni/

[21] Reperibile in https://foia.gov.it/osservatorio/monitoraggio

1176 richieste pervenute ne sono state evase l'82% nel termine dei 30 giorni stabilito dalla legge e il 62% di queste è stato accolto (il 53% totalmente e il 9% parzialmente)[22]. Possiamo quindi evidenziare come la maggior parte delle istanze formulate, almeno in riferimento a quelle presentate ai ministeri, venga accolta ma il risultato finale risulti distante dall'effettiva totale accessibilità.

Nonostante i limiti illustrati e altri[23], con l'introduzione dell'accesso civico generalizzato il nostro ordinamento ha compiuto un fondamentale passo verso la *full disclosure*, con il riconoscimento del diritto di ogni individuo di avere piena conoscenza dell'azione amministrativa come strumento per promuovere la partecipazione attiva e consapevole dei cittadini alla vita democratica.

3. Comunità monitoranti e partecipazione democratica.

Come già evidenziato, il principio della trasparenza non solo è legato a quello di buon andamento della pubblica amministrazione ed al contrasto dei fenomeni corruttivi ma rappresenta anche uno strumento di partecipazione attiva dei cittadini alla vita pubblica, in quanto risulta funzionale a creare un dialogo cooperativo e proattivo tra amministratori ed amministrati[24]. In questo senso, lo stesso dlgs. 33/2013, all'art. 1, afferma chiaramente le finalità di "favorire forme diffuse di controllo sul perseguimento delle funzioni istituzionali e sull'utilizzo delle risorse pubbliche" e "attuare il principio democratico".

Non si intende affermare che l'introduzione di strumenti di accesso civico sia sufficiente a risolvere ogni problematica di trasparenza nel rapporto tra le pubbliche amministrazioni e i cittadini. Esistono anche altre criticità da considerare, basti pensare alle problematiche relative al problema del *free rider*[25] e alla possibilità che l'azione di controllo sia influenzata dalle élite locali[26]. Tuttavia, fornire la possibilità di una migliore conoscenza delle informazioni in possesso delle pubbliche amministrazioni può favorire la promozione di una vita democratica matura basata su una partecipazione attiva costante e non limitata al momento elettorale[27]. Si ritiene di condividere la tesi secondo cui un monitoraggio effettuato dalle popolazioni, che beneficeranno dal successo dell'azione efficiente del settore pubblico, possa risultare più efficace rispetto ad uno che basato unicamente su controlli di carattere burocratico

---

[22] Le motivazioni più ricorrenti dei rigetti sono, in relazione all'art 5 bis del dlgs. 97/2016, sicurezza pubblica e ordine pubblico (23%) e, in riferimento a diverse motivazioni da quelle del menzionato articolo, dati non posseduto dall'amministrazione (44%).

[23] Ad esempio l'assenza di un'adeguata tutela giurisdizionale realmente efficace in caso di inattività dell'amministrazione interpellata.

[24] S. Di Pietro, *La tutela della privacy tra esigenze di trasparenza e nuove regole di riservatezza*, *cit.*, p. 8.

[25] B. A. Olken, *Monitoring Corruption: Evidence from a Field Experiment in Indonesia*, in «Journal of Political Economy», 2, 2007, p. 204.

[26] P. Bardhan, *Decentralization of Governance and Development*, in «Journal of Economic Perspectives», 16, 2002, p. 192.

[27] Cfr. M. Savino, *Il FOIA italiano. La fine della trasparenza di Bertoldo*, in «Giornale di diritto amministrativo», 5, 2016.

delle autorità centralizzate[28]. In tale ottica, la possibilità di disporre di informazioni tempestive e accurate rappresenta un fondamentale elemento per la partecipazione civica e gli stessi dati assurgono alla categoria di bene comune[29].

Pare opportuno esaminare il concetto di trasparenza integrale, consistente in un sistema volto a tutelare gli interessi collettivi e a prevenire fenomeni corruttivi mediante l'esercizio di un controllo diffuso dell'esercizio del potere delegato ai rappresentanti da parte dei cittadini con un approccio cooperativo[30]. Proprio in questa categoria possiamo ritrovare le ragioni dell'iniziale metafora vitrea che dovrebbe rappresentare la prassi per le pubbliche amministrazioni risultando

> *un apparato pubblico dalle mura di vetro, anzi di cristallo infrangibile dove tutto sia osservabile e valutabile da tutti, ma con porte blindate per assicurare protezione dai ladri, rendendo il vivere collettivo (e la cosa pubblica) inospitale per i corruttori, inaccessibile ai corrotti, indisponibile alle mafie. Non solo: è necessario un impegno ulteriore per creare un presidio nella pubblica amministrazione a opera di una società civile capace di vigilare con la cura e l'attenzione che servono[31].*

Emerge chiaramente come questa concezione di trasparenza necessiti di una partecipazione attiva della popolazione che realizzi prassi civiche di monitoraggio dal basso accanto alle strategie istituzionali di contrasto alla *maladmnistration*. Pertanto, i due pilastri che legano la trasparenza ai principi costituzionali, partecipazione democratica e contrasto della corruzione, appaiono assolutamente connessi nella realizzazione della "casa di vetro"[32]. In tale ottica, va considerato come la trasparenza integrale da un lato richieda e dall'altro stimoli l'esistenza di una democrazia che vada oltre quella classica meramente rappresentativa. Il protagonismo civico non può essere limitato al mero momento elettorale ma va estenda alla partecipazione a nuove istituzioni di controllo che vigilino costantemente sull'esercizio del potere delegato ai decisori politici. Questa nuova forma può essere definita democrazia monitorante, concetto di non semplice definizione che trova la sua caratterizzazione nel pubblico controllo di ogni aspetto della vita politica e sociale oltre gli istituti della democrazia rappresentativa tramite un apparato di organi non partitici, extraparlamentari e spesso non eletti che operano sia all'interno che all'esterno dei confini statali[33]. Keane al riguardo individua accuratamente i principi su cui essa si basa

---

[28] J. E. Stiglitz. *Participation and Development: Perspectives from the Comprehensive Development Paradigm*, in «Review of Development Economics», 6, 2002, pp. 165-167.

[29] U. Di Maggio, G. Notarstefano, G. Ragusa, *Ri-conoscere i beni confiscati. Un percorso tra partecipazione, condivisione e trasparenza*, in R. Ingrassia (a cura di), *Economia, organizzazioni criminali e corruzione*, Aracne editrice, Canterano, 2018, p. 158.

[30] L. Ferrante, A. Vannucci, *Anticorruzione pop*, Edizioni Gruppo Abele, Torino 2017, p. 130.

[31] *Ibidem*.

[32] Pare opportuno evidenziare come il legame tra la qualità della democrazia e la lotta alla corruzione è già stata evidenziata da eminente dottrina, cfr. A. Vannucci, *Atlante della corruzione*, Edizioni Gruppo Abele, Torino, 2012, pp. 210-216.

[33] J. Keane, *Potere e umiltà. Il futuro della monitory democracy* (2021), tr. it. di P. D'Ortona, Hopefulmonster, Torino, 2021, p. 122.

> Nell'età della democrazia monitorante è come se si applicassero i principi della democrazia rappresentativa – carattere pubblico, uguaglianza dei cittadini, elezione dei rappresentanti – alla democrazia rappresentativa stessa. L'effetto più notevole di questo processo è modificare i modelli di interazione – la geografia politica – delle istituzioni democratiche[34].

Cuore di questo modello di democrazia non possono che essere le organizzazioni civiche, anche informali, che si pongono il fine di esercitare l'azione di monitoraggio collettivo a tutela dell'integrità del sistema: le comunità monitoranti[35]. Si tratta di realtà che, sebbene non ancora particolarmente sviluppate in Italia, sono ben conosciute nel contesto internazionale[36] e hanno contribuito a promuovere l'impegno civico e il miglioramento strutturale dei servizi monitorati[37]. Nel descrivere la loro azione sono stati individuati tre fondamentali passaggi consistenti nell'esaminare i processi decisionali tramite gli strumenti della trasparenza (illuminare), nell'elaborare e diffondere le informazioni raccolte (vigilare) e nel coinvolgere della cittadinanza in alcune fasi dei processi decisionali (partecipare)[38].

4. *Esempi di monitoraggio civico tramite gli strumenti della trasparenza: i beni confiscati alle mafie.*

Come evidenziato, la trasparenza offre importanti occasioni protagonismo per la cittadinanza e allo stesso tempo necessita che questa si dedichi ad una partecipazione attiva per non rischiare di essere ridotta alla mera dimensione burocratica. Nel nostro paese, negli anni, si sono sviluppati importanti esempi di "comunità monitoranti" che hanno vigilato con gli strumenti a loro disposizione sull'integrità di settori strategicamente fondamentali[39]. Un ambito in cui si sono sviluppati importanti esempi di monitoraggio dal basso organizzati dalle organizzazioni della società civile è quello dei beni confiscati alle mafie.

La normativa di riferimento si basa su due pilastri principali ossia un sistema di misure patrimoniali volte a colpire i beni intestati agli associati alle consorterie mafiose molto più efficace di quelle previste nei confronti della criminalità comune e il riutilizzo sociale di quanto incamerato in questo modo dallo stato. La *ratio* sottesa a tale impianto consiste nel garantire un parziale ristoro dei territori danneggiati dalla

---

[34] *Ibidem*.

[35] L. Ferrante, A. Vannucci, *Anticorruzione pop*, cit., pp. 133-134.

[36] A titolo di esempio, si riporta l'esperienza indonesiana in B. A. Olken, *Monitoring Corruption*, cit.. Altri esempi relativi ad esperienze realizzate in Italia verranno riportati nel paragrafo successivo.

[37] M. Björkman Nyqvist, D. de Walque D., J. Svensson, *Information is Power. Experimental Evidence on the Long-Run Impact of Community Based Monitoring*, in «World Bank Policy Research Working Paper», 7015, 2014, p. 3.

[38] L. Ferrante, A. Vannucci, *Anticorruzione pop*, cit., pp. 135-189.

[39] Basti pensare alla campagna "Illuminiamo la salute" che promuoveva un monitoraggio per l'integrità della sanità, ben precedente all'inserimento dello strumento dell'accesso civico generalizzato http://www.illuminiamolasalute.it/.

presenza delle organizzazioni criminali mediante la restituzione dei patrimoni mafiosi sequestrati e confiscati ed ha rappresentato un elemento strategico fondamentale per colpire il consenso popolare costruito dalle mafie. Proprio in ragione di questo, trasparenza e conoscibilità da parte della popolazione in questo ambito rappresenta un elemento fondamentale e irrinunciabile.

In tale contesto, sono nate alcune campagne di monitoraggio civico che sono risultate particolarmente efficaci. Ci si riferisce, in particolare, l'analisi realizzata dal progetto Confiscati Bene[40], nato nel 2014 dalla collaborazione tra Libera e onData con il sostegno della Fondazione Tim, ossia un percorso partecipativo che mira a valorizzare il riutilizzo dei beni confiscati attraverso monitoraggio, raccolta e analisi dei dati che li riguardano. Nella versione 2.0, il suo portale online è stato poi implementato proprio per permettere alle comunità di contribuire attivamente alla creazione del patrimonio informativo[41].

A partire dal 2022 è nata un'ulteriore esperienza di monitoraggio in civico in tale settore: il report "Rimandati"[42], avente ad oggetto trasparenza dei beni confiscati nelle amministrazioni locali, che nel 2024 è giunto alla sua terza edizione[43]. Il progetto si fonda sul metodo del *community-based monitoring* che prevede una vigilanza diffusa da parte dei cittadini. Infatti, pur mantenendo una cabina di regia centrale, è stata creata una comunità monitorante, composta da 41 volontari e 6 tirocinanti dell'Università di Torino[44], diffusa in tutto il territorio italiano[45]. Questa ha messo in atto l'azione di monitoraggio su 1127 enti territoriali al cui patrimonio indisponibile sono stati assegnati dei beni confiscati e che sono stati ritenuti inadempienti rispetto ai relativi obblighi di pubblicazione[46]. Il lavoro della comunità monitorante si è articolato in tre fasi. Inizialmente, mediante la consultazione del portale OpenRe.g.i.o. sono stati individuati gli enti destinatari, successivamente un esame dei relativi siti istituzionali ha permesso di individuare i 1127 organismi già citati[47] indicati e di rivolgere ad essi istanze di accesso civico semplice. L'ultima fase è consistita in una seconda ricognizione dei siti istituzionali degli enti che hanno risposto all'istanza di accesso civico[48]. Per sintetizzare i risultati dell'azione, possiamo evidenziare come il 63% degli enti interpellati ha risposto all'istanza di accesso civico e tra questi 296 (il 44,6%) ha

[40] https://www.confiscatibene.it/

[41] U. Di Maggio, G. Notarstefano, G. Ragusa, *Ri-conoscere i beni confiscati. Un percorso tra partecipazione, condivisione e trasparenza*, cit., pp. 167-168.

[42] I promotori in questo caso sono Gruppo Abele il Dipartimento di Culture, politica e società dell'Università di Torino, e, limitatamente all'ultima edizione, ISTAT. Vale la pena però evidenziare che molte degli attivisti coinvolti nel progetto *Confiscati Bene* hanno collaborato anche a questo report portando un importante contributo in termini di *know-how*.

[43] https://www.libera.it/schede-2438-beni_confiscati_monitoraggio_libera_trasparenza

[44] L'avvio del percorso monitoraggio ha inizialmente coinvolto 107 volontari che hanno partecipato ad una specifica formazione.

[45] AA.VV., *Rimandati*, terza edizione, 2024, p. 32.

[46] Nello specifico 1100 comuni, 11 province e città metropolitane e 6 regioni. Ivi p. 34.

[47] L'inadempienza riguardava l'assenza di dati, la loro incompletezza o il loro mancato aggiornamento.

[48] Nello specifico hanno risposto 710 enti sui 1127 enti destinatari di istanza di accesso civico.

risposto in maniera pienamente soddisfacente determinando un notevole miglioramento nella situazione relativa alla trasparenza dei beni confiscati assegnati agli enti locali.

5. Conclusioni.

In un momento storico nel quale le giovani generazioni, ma non solo, paiono orientarsi verso individualismo e socialità ristretta[49], emergono invece forme peculiari di impegno. Queste si caratterizzano per una dimensione maggiormente individuale e meno mediata dai soggetti collettivi[50], per una partecipazione che si caratterizza per il mantenimento dell'individualità dei partecipanti e per la connessione con la vita quotidiana[51].

Come abbiamo visto gli strumenti della trasparenza e il metodo del *community-based monitoring* sembrano proprio in linea con tali caratteristiche in quanto permettono la valorizzazione dell'azione individuale nell'ambito di quella collettiva e consentono di intervenire su aspetti fortemente legati alle condizioni materiali della vita quotidiana grazie alla possibilità di operare sul livello istituzionale più prossimo ai cittadini.
La produzione del report Rimandati dimostra esattamente che mediante tali strumenti è possibile realizzare l'attivazione di un numero consistente di cittadini, molti dei quali appartenenti alle nuove generazioni, e creare un sensibile miglioramento nell'ambito di uno specifico ambito di azione che ha portato ad una conoscibilità nettamente maggiore di un settore strategico come quello dei beni confiscati alle mafie e trasferiti al patrimonio degli enti locali. Rispetto alla situazione di partenza, va evidenziato infatti come i casi di inadempienza agli obblighi di pubblicazione siano diminuiti notevolmente in quanto ben 296 enti hanno dato una risposta pienamente corretta[52] alle istanze di accesso civico[53]. Pertanto, possiamo concludere che la pratica della democrazia monitorante rappresenta un importante occasione di partecipazione dei cittadini alla vita pubblica tramite cui essi possono incidere profondamente sull'azione degli apparati istituzionali.

---

[49] Cfr. A. De Lillo, *I sistemi di valore*, in C. Buzzi, A. Cavalli, A. De Lillo (a cura di), *Giovani del nuovo secolo. Rapporto IARD sulla condizione giovanile in Italia*, il Mulino, Bologna, 2002.
50 A. Pirni, L. Raffini, *Giovani e politica. La reinvenzione del sociale*, Mondadori, Milano, 2022, p. 14.
[51] *Ivi*, p. 106.
[52] Si sottolinea che altri 76 enti hanno dato una risposta non del tutto corretta in quanto l'accessibilità ai dati inviati non era immediata per i richiedenti.
[53] AA.VV., *Rimandati*, cit., p. 70.

La trasparenza nei mercati finanziari: approccio classico e nuovi paradigmi[a]

Giulia Miotti[*]

Abstract
Il sistema finanziario è un sistema sociale molto complesso con una forte ramificazione all'interno di tutti gli altri sistemi sociali dei Paesi avanzati: ha un impatto fortissimo sull'economia reale e i tempi degli scambi finanziari hanno ormai determinato un'accelerazione anche nei tempi della vita sociale anche fuori dai mercati. Questo sistema è un sistema sociale anche perché basa il proprio funzionamento sulla ricerca e lo scambio di informazioni; eppure, sebbene l'informazione sia un concetto cardine all'interno della pratica e delle teorie dei mercati finanziari, questa sembra slegata dal concetto di trasparenza intesa come apertura, comunicazione e accountability. Questa mancata corrispondenza produce degli effetti rilevanti all'interno dei mercati e nella possibilità di una scelta informata ed equa degli agenti del mercato. Vedremo due possibili risposte a questa disfunzione interna ai mercati, una di ordine politico e l'altra di governance.

Parole chiave: mercati finanziari; trasparenza; disfunzione informativa.

Abstract
The financial system is a highly complex social system, deeply entangled with all other social systems in developed Countries. It exerts a disruptive impact on the real economy sector and the speed of financial exchanges seems to have determined a similar acceleration also in the speed and nature of social life itself. Another reason why the financial system can be considered a social system lies in the fact that the financial system ground its functioning in the research and exchange of information. In this context, information represents a pivotal concept around which financial practice and financial theories alike turn. Notwithstanding this, the notion of information seems detached from the notion of transparency meant as openness, communication and accountability. The lack of such correspondence engenders critical effects on financial markets, especially when it comes to the possibility of

[a] Saggio ricevuto il 31/05/2024 e pubblicato il 22/01/2025.
[*] Ricercatrice post-doc, Labont, email: giulia.miotti@unito.it.

making fair and informed choices by market agents. We shall describe and discuss two possible alternative scenarios for such market dysfunction; the first one is of a political kind, the second one of a governance-oriented one.

Keywords: fiancial markets; transparency; informational dysfunction.

1. Introduzione

Il concetto di "informazione" è di primaria importanza all'interno della teoria finanziaria e della pratica finanziaria. Il modo in cui le informazioni circolano all'interno dei mercati finanziari è centrale per comprendere il comportamento "epistemico" e le strategie di decision-making degli agenti e ha un'influenza diretta sulle possibilità di successo di una strategia di investimento. In quest'ottica, lo scambio finanziario è, in linea di principio, uno scambio di informazioni e la principale attività dei player finanziari è legata alla ricerca della migliore informazione, che in questo senso assume quasi precedenza sugli oggetti che vengono infine offerti e acquistati sul mercato[1]. Curiosamente, all'interno della finanza classica il concetto di informazione è slegato da quello di trasparenza intesa come apertura, comunicazione, accountability.

Come accennato, il ruolo svolto dall'informazione nei mercati è centrale tanto nella teoria quanto nella pratica finanziaria, e ogni teoria sul funzionamento dei mercati dedica largo spazio a questa indagine, occupandosi di descrivere in quale modo le informazioni presenti sul mercato determinano le possibilità epistemiche degli agenti; domandandosi, ad esempio, attraverso quali modelli e con quali capacità le informazioni vengono processate.

Proporremo una breve descrizione del ruolo ascritto all'informazione da parte della teoria denominata "ipotesi dei mercati efficienti" (EMH, *Efficient Market Hypothesis*): questa è una delle prime teorie ad essersi affermata tanto a livello accademico quanto a livello operativo ed è ancora una delle più largamente accettate nell'ambito della finanza classica.

All'interno di questa teoria il concetto di "informazione" assume un'importanza centrale. Sono proprio l'esistenza di informazione corretta e la sua piena circolazione all'interno del mercato che permettono che il mercato finanziario sia un contesto razionalmente comprensibile e che gli agenti finanziari possano non solo compiere scelte razionali, ma anche avere ragionevoli probabilità di prevedere in parte futuri andamenti dei mercati; o, almeno, di non essere sorpresi da quelli che vengono descritti come "eventi estremi". Non ci soffermeremo sull'analisi della validità scientifica della teoria dei mercati efficienti e dei suoi modelli, ma ci concentreremo sulla descrizione di come l'informazione agisce nella descrizione degli oggetti finanziari e cercheremo di mostrare perché l'attività legata alla ricerca di informazione in finanza e la quantità di informazione raccolta non sono proporzionali

[1] Cfr. F. Mirowski, E. Nik-Khah, *The Knowledge we have lost in Information*, Oxford University Press, Oxford 2017.

al livello di trasparenza relativa agli oggetti finanziari che restano invece opachi e alle strategie e scelte degli operatori finanziari che restano poco comprensibili. Questa separazione tra informazione e trasparenza ha due effetti particolarmente evidenti a livello economico e sociale: uno a livello "sistemico", l'altro a livello "locale". Forniremo per entrambi degli esempi.

Il problema dell'informazione che non sembra in grado di fornire alcuna trasparenza sugli oggetti che dovrebbe contribuire a formare, rappresenta un vulnus significativo all'interno della pratica finanziaria. Il sistema finanziario è infatti un sistema sociale, e un sistema sociale che non si dimostra in grado di veicolare informazione in maniera aperta e comprensibile non è un sistema equo e democratico. La trasparenza rappresenta infatti quella condizione che secondo l'Open Government partnership «*empowers citizens to exercise their rights and participate in decision-making processes*»[2] e che quindi si pone come unica strategia efficace per mantenere la salute democratica di qualsiasi istituzione sociale. È questa una questione che negli ultimi anni sta diventando sempre più pressante, anche all'interno della cultura finanziaria. Ci riferiamo in particolare a due approcci: il primo è la *Sustainable Finance Disclosure Responsibility* (SFDR), il regolamento europeo entrato in vigore a marzo 2021 che promuove un'informativa sulla sostenibilità nel settore dei servizi finanziari, il secondo è rappresentato da forme di finanza etica e responsabile che propongono forme di governance alternative a quelle tipiche della finanza classica. Esamineremo entrambi questi approcci, e mostreremo come la richiesta di trasparenza possa essere esaudita all'interno di forme di gestione finanziaria alternative, che considerano la trasparenza intesa come apertura, comunicazione e accountability uno dei cardini del proprio operare.

2. La teoria efficientista dei mercati e l'informazione

La teoria dei mercati efficienti è la prima teoria ufficiale sul funzionamento dei mercati finanziari, ad oggi ancora la più studiata e accettata[3]. Muove dalle assunzioni dell'economia neoclassica, riprende infatti concetti cardine quali l'equilibrio del mercato e la razionalità dell'agente economico[4]. Secondo questi due concetti, un mercato è tendenzialmente (ovvero nel lungo temine) un sistema in equilibrio perché l'insieme delle scelte operate dagli agenti economici lo rendono stabile: sono infatti delle scelte operate da individui pienamente razionali, in possesso di informazioni complete e rilevanti, e questo fa sì che tendenzialmente non avvengano squilibri fra i due grandi perni dei mercati finanziari, la domanda e l'offerta. La teoria efficientista è infatti una teoria che intende descrivere a tutto tondo la natura dei mercati finanziari e procede a una modellizzazione piuttosto articolata del mondo sociale: accanto a

[2] https://www.opengovpartnership.org/glossary/transparency/

[3] Cfr. E. Fama, *Efficient Capital Markets: A Review of Theory and Emipirical work,* in «The Journal of Finance», vol 25 (2), 1970, pp. 383-471.

[4] Cfr. E. Ippoliti, *Un filosofo a Wall Street. Speculazioni sulla finanza da Aristotele ai bitcoin*, Egea, Milano, 2020.

delle assunzioni antropologiche (su tutte, l'agente economico descritto come massimizzatore di profitto) propone ad esempio dei modelli matematicamente raffinati. Non ci occuperemo però di questi aspetti, ma ci concentreremo esclusivamente sulla descrizione del ruolo dell'informazione nella formazione dei prezzi e nella previsione dei futuri andamenti del mercato.

Seguendo l'analisi proposta da Keen[5], definiremo il concetto di "efficienza" del mercato attraverso i seguenti punti:

- le aspettative collettive degli investitori nel mercato azionario rappresentano predizioni accurate delle prospettive future delle società sul mercato;
- i prezzi delle azioni riflettono pienamente tutta l'informazione rilevante per la comprensione delle prospettive future delle società sul mercato;
- le oscillazioni nei prezzi delle azioni sono dovute a cambiamenti relativi all'informazione rilevante per le prospettive future, quando quell'informazione arriva in maniera randomica e non prevedibile;
- quindi, i prezzi azioni seguono una "*random walk*", cosicché i movimenti passati dei prezzi non forniscono alcuna informazione sui futuri movimenti.

Chiariamo meglio due di questi aspetti, fondamentali all'interno della nostra analisi. Partiamo dal secondo aspetto. Affermare che un mercato è efficiente quando i prezzi delle azioni riflettono pienamente tutta l'informazione rilevante disponibile equivale ad affermare che in un mercato efficiente il prezzo di un qualsiasi oggetto rispecchia il suo valore fondamentale. In questo contesto, infatti, gli agenti sul mercato hanno processato in maniera corretta tutte le informazioni disponibili su un dato oggetto finanziario e hanno così creato delle aspettative collettive razionali sui futuri flussi che ne determineranno il prezzo. In questo modo, vi sarebbe un rapporto diretto tra le informazioni utilizzate razionalmente (ovvero, da tutti allo stesso modo) e il risultato finale, il prezzo.

Continuiamo con il terzo aspetto, relativo alle oscillazioni dei prezzi e a come gli agenti sul mercato reagiscono. Come affermato dal principio, le oscillazioni dei prezzi sono il risultato di informazione arrivata in maniera casuale che influisce sull'equilibrio raggiunto dal valore delle azioni. Il fatto che queste informazioni siano sì randomiche e imprevedibili ma comunque a disposizione di tutti una volta entrate nel mercato, fa sì che gli agenti siano in grado di reagire razionalmente, valutando in maniera oggettiva le nuove informazioni e, grazie a questa valutazione, possano "riaggiustare" i prezzi a un nuovo punto di equilibrio.

Se questi principi reggessero, allora il mercato sarebbe perfettamente efficiente; molti autori sostengono però che le condizioni ipotizzate siano sostanzialmente irrealistiche e irrealizzabili[6]. La loro realizzabilità, infatti, presupporrebbe che tutti gli investitori sul mercato (dai più grandi ai piccoli

[5] Cfr. S. Keen, *Debunking Economics. The Naked Emperor Dethroned?* Zed Books, London, 2011.
[6] Cfr. G. Soros, *The Alchemy of Finance*, Wiley Publishing, New York, 2003.

risparmiatori) possano avere accesso allo stesso tipo di informazioni allo stesso momento e che quindi tutti gli agenti sul mercato abbiano a disposizione le stesse, necessarie possibilità, strumenti e tempo per raccogliere e processare le stesse informazioni. Come sostiene Keen[7] questo scenario è fondamentalmente irrealizzabile, basti pensare a due condizioni da questo richieste: il completo accordo tra gli agenti del mercato e disponibilità di liquidità pressoché illimitata per tutti (ovvero, chiunque può prendere in prestito e dare in prestito quanto desidera). Tornando all'informazione, il problema è simile: come afferma Ippoliti[8] «nei moderni mercati finanziari l'informazione è tale per cui la sua acquisizione può essere anche estremamente costosa e l'ipotesi (neoclassica) di informazione perfetta (ossia l'onniscienza degli attori finanziari) si rivela quanto meno descrittivamente inaccurata. I mercati si comportano molto spesso in modo imperfetto dal punto di vista informativo e nei mercati imperfetti c'è spazio per il potere. […] Quando la conoscenza di prodotti, processi e ambienti complessi non è uniforme ed è ineguale, la conoscenza è potere e alcuni attori finanziari, se sono egoisti e razionali, saranno in grado di sfruttare la situazione a loro vantaggio»[9].

3. Due esempi di malfunzionamento informativo

Come anticipato nell'introduzione, la non-coincidenza tra informazione ed effettiva trasparenza delle dinamiche all'interno dei mercati finanziari produce degli effetti di forte impatto. Questi effetti possono essere distinti come effetti a livello "sistemico" ed effetti a livello "locale"; i primi interesserebbero il sistemo finanziario nel suo stesso funzionamento e nella sua resilienza agli shock, mentre i secondi coinvolgerebbero un aspetto specifico dell'operare finanziario ed economico-sociale in senso più ampio.

Il primo esempio che vorremmo qui riproporre è relativo alla crisi finanziaria del 2008 e rappresenta un caso di asimmetria informativa, la condizione per cui l'informazione disponibile non viene condivisa in maniera paritaria tra gli attori coinvolti in uno stesso processo finanziario o economico. Cercheremo di ricostruire brevemente la genesi e lo sviluppo di questa crisi.

La crisi finanziaria del 2008 è una crisi che nasce a partire dall'esplosione della bolla del mercato immobiliare statunitense e che si è poi diffusa molto velocemente a livello globale, fino a diventare la peggior crisi finanziaria dopo la Grande Depressione degli anni '20 del 1900. Nel 2006, il mercato immobiliare statunitense viene interessato da una forte crescita, e durante quello che sarà poi descritto come il "boom" di una bolla speculativa viene erogata una grande quantità di mutui, concessi anche a soggetti generalmente non considerati bancabili. In questo caso si è parlato infatti di mutui sub-prime, ovvero dei prestiti ad alto rischio perché concessi a clienti con alto rischio di insolvenza (basso reddito, lavoro a tempo determinato, precedenti insolvenze).

[7] Cfr. S. Keen, *Debunking Economics*, cit.
[8] Cfr. E. Ippoliti, *Un filosofo a Wall Street*, cit.
[9] *Ibidem*, p. 267.

Seguendo un ciclo tipico delle economie capitaliste avanzate detto *boom-bust*, per cui ad un processo di forte espansione ne segue uno di veloce contrazione, alla fine del 2006 la bolla immobiliare si sgonfia. Molti dei possessori di un mutuo sub-prime diventano insolventi a causa del rialzo dei tassi di interesse, portando così il mercato finanziario sull'orlo del collasso.

Il passaggio da quella che poteva sembrare una crisi finanziaria ristretta a un settore specifico a una crisi dell'intero sistema finanziario globale (o almeno, dei paesi più ricchi) avviene in diverse tappe di raffinata ingegneria finanziaria, all'interno delle quali l'informazione rilevante relativa agli oggetti scambiati sul mercato non è certamente stata incorporata nel valore attribuito loro. Il primo passo ha visto la scelta di riunire i singoli mutui in più grandi pacchetti, creando dei *mortgage-back security* (MBS). Un MBS è un titolo che deriva i propri flussi di cassa dal portafoglio di prestiti ipotecari che lo costituiscono: la tenuta finanziaria di una simile struttura dipende direttamente dal flusso dei pagamenti dei mutui sottostanti. All'interno di questi MBS, i vari mutui che li compongono vengono divisi in più livelli, a seconda del rischio di insolvenza loro associato: in basso i mutui a rischio di insolvenza più alto (più rischiosi per gli investitori ma che offrono dei tassi di interesse elevati), in alto i mutui a rischio di insolvenza più basso (meno rischiosi, ma con tassi di interesse più bassi). È evidente, a questo punto, che le informazioni rilevanti relative alla struttura dei diversi *mortage-back securities* non sono state utilizzate in maniera tale da garantire comprensione e accountability né a favorire la formazione di aspettative razionali da parte degli agenti del mercato. Potremmo interrompere qui l'analisi della crisi partita dai sub-prime, ma è interessante comprendere come la disfunzione informativa abbia avuto un ruolo critico nello sviluppo di una delle più grandi crisi del mercato finanziario.

Il secondo passo consiste in un'ulteriore azione di ingegneria finanziaria grazie alla quale vengono costruiti dei nuovi strumenti finanziari a partire dagli originari mutui, ovvero dei *Collateralized Debt Obligations*, CDO. Un CDO è uno strumento formato da un insieme di titoli obbligazionari che hanno come garanzia un credito. Vengono definiti strumenti derivati perché derivano il proprio andamento, e legano quindi il proprio rimborso, a un insieme di strumenti collaterali sottostanti. In questo caso, gli strumenti collaterali sono rappresentati da obbligazioni ipotecarie, mutui. L'unica garanzia a tenuta dei CDO è quindi quella offerta dal valore globale dei collaterali. Per costruire un CDO viene raggruppato un insieme piuttosto numeroso di singoli collaterali (obbligazioni) di dimensione modesta, queste obbligazioni vengono poi suddivise in tranche che hanno diversi livelli di priorità e solidità[10] secondo un modello a cascata per cui in caso di fallimento delle obbligazioni sottostanti, le tranche a priorità alta verranno pagate per prime, le altre a seguire fino

[10] I diversi livelli di priorità sono rappresentati dalle seguenti categorie: *supersenior*, *senior*, *mezzanine* e *junior*. In un CDO *supersenior* o *senior*, vengono generalmente raggruppate tranche di titoli ipotecari con rating alto; in un *mezzanine*, invece, vengono raggruppate tranche con un rating più basso. Questa ripartizione fa sì che i *mezzanine* siano più rischiosi ma che offrano, al tempo stesso, tassi di interesse più alto; i *supersenior* o i senior più sicuri ma con tassi di interesse più bassi.

a disponibilità finanziaria. Un CDO viene generalmente diviso al suo interno in diverse sezioni. Poniamo, ad esempio, che venga diviso in sette sezioni: almeno tre di queste saranno verosimilmente valutate con un rating molto alto, mentre almeno una avrà un rating molto basso.

Sebbene la maggior parte dei CDO emessi tra il 2006 e il 2007 contenesse mutui subprime con un rating molto basso (che esprime quindi un rischio molto alto), sono stati valutati con un rating alto: i mutui subprime, infatti, sono stati tagliati in diverse tranche inserite in contesti che ne hanno fatto diminuire il rischio offrendo alti rendimenti. La loro origine, e la valutazione sulla loro potenziale tenuta, non sono in questo contesto ritenuti informazione rilevante per la corretta valutazione di questi prodotti. Il loro potenziale rischio, infatti, viene considerato accettabile perché bilanciato da un calcolo probabilistico: questo, a meno che non ci sia un fenomeno di correlazione molto forte tra il fallimento di più mutui, ovvero a meno che non avvenga una crisi generale del settore immobiliare. Non è nostro interesse ripercorrere le diverse fasi attraverso le quali si è formata ed è poi scoppiata la crisi finanziaria legata ai mutui subprime: ricostruire invece la struttura dei diversi strumenti finanziari creati a partire da mutui subprime e la manipolazione dell'informazione relativa alla natura dei mutui come strumenti collaterali è di importanza centrale al nostro discorso.

Ciò che in questo caso sembra particolarmente importante notare è come l'informazione rilevante relativa ai mutui subprime come building-blocks di strumenti finanziari più sofisticati non venisse rispecchiata nel valore attribuito ai CDO. Il valore cui facciamo qui riferimento ha duplice valenza: ci riferiamo qui sia al valore di mercato inteso come il prezzo attribuito ai diversi CDO, sia al rating loro accordato, che dovrebbe esprimere una forma di equilibrio tra il rischio che un dato prodotto finanziario comporta e il tasso d'interesse che può offrire. Oltre a una mancanza di trasparenza, questo aspetto mostra come la manipolazione dell'informazione all'interno della pratica finanziaria classica contravvenga ai criteri stessi che definiscono l'efficienza di un mercato finanziario: la corrispondenza tra valore/prezzo e informazione è stata infatti sostanzialmente tradita, facendo sì che all'interno del mercato circolassero e venissero scambiati dei prodotti sostanzialmente inconoscibili nella loro struttura e nel loro potenziale effetto sulla tenuta del mercato stesso.

Come abbiamo provato a spiegare con l'esempio dei mutui subprime e dei loro effetti su scala globale, il problema della manipolazione dell'informazione può avere effetti sistemici sul mercato finanziario. Il secondo tipo di effetto cui accennavamo è invece di tipo "locale", ovvero riguarda la possibilità per gli investitori di comprendere pienamente e con facilità quanti e soprattutto quali attori vengano coinvolti nelle attività finanziarie sostenute grazie ai loro investimenti o risparmi e dunque comprendere e stabilire il tipo di impatto a livello sociale ed economico che contribuiscono o meno a generare.

A livello "locale", dunque, il problema della gestione dell'informazione assume più marcatamente i contorni di un problema di *transparency* vero e proprio. Tornando alla definizione di trasparenza proposta all'inizio, un comportamento è trasparente se

permette agli attori coinvolti di esercitare i propri diritti e di partecipare attivamente in processi decisionali. Rileggendo quest'affermazione in un'ottica che possa includere il punto di vista degli investitori nel mercato finanziario, dovremo pensare alla possibilità di questi ultimi di esercitare i propri diritti come alla possibilità di avere garantito il loro diritto all'informazione, e alla possibilità di partecipare attivamente in processi decisionali come alla possibilità di sostenere finanziariamente lo sviluppo di attività con un impatto sociale, ambientale ed economico positivo per la collettività.

Nel contesto della finanza classica il problema della trasparenza assume una duplice connotazione, la prima riguarda la forma tipica di governance della finanza e la seconda riguarda quale ruolo debbano avere considerazioni di ordine etico all'interno delle scelte di finanziamento e investimento.

Come abbiamo visto nel caso della crisi dei mutui subprime, la costruzione di nuovi strumenti finanziari spesso avviene seguendo un meccanismo piuttosto complesso, all'interno del quale risalire allo strumento originario e alle sue caratteristiche più rilevanti può rivelarsi difficile. Un meccanismo simile avviene per gli investimenti, all'interno dei quali i risparmiatori o gli investitori possono conoscere l'obiettivo finale di investimento ma non riescono ad avere acceso al percorso dei flussi che loro stessi contribuiscono a finanziare, perdendo così la possibilità di comprendere quante e quali realtà sostengono nel corso del proprio investimento.

Questa mancanza di trasparenza assume i contorni di un problema di governance perché implica una forma di gestione delle informazioni di tipo gerarchico e piramidale, all'interno del quale solo pochi partecipanti hanno diritto a un accesso diretto e completo alle informazioni.

Come accennato prima, il problema della trasparenza solleva anche un ulteriore aspetto che dovrebbe avere un'importanza centrale all'interno della pratica e delle scelte finanziarie, quello del ruolo dell'etica nelle scelte e nell'operare finanziario. Le due funzioni basilari che il sistema finanziario assolve per garantire una crescita economica stabile[11] sono rappresentate dalla redistribuzione del reddito e dalla creazione di valore; nonostante queste funzioni siano di grande importanza e inseriscano l'operare finanziario all'interno di un contesto di azione vasto che include un pubblico molto ampio, non sembrano aver ancora accolto delle istanze di ordine etico e sviluppato strategie per rendere l'operare finanziario pienamente trasparente e sostenibile. La finanza classica, infatti, non sembra aver affrontato la questione se sia sempre giusto investire in qualsiasi settore che presenti delle opportunità di sviluppo o se alcuni ambiti, pur offrendo grandi opportunità di mercato nel breve e medio termine, dovrebbero essere ignorati in favore di altri in grado di garantire redistribuzione di risorse e creazione di valore nel lungo termine. Ad esempio, i settori del mercato che secondo un approccio etico e sostenibile non dovrebbero essere incoraggiati sono quelli che generano degli impatti sociali e ambientali particolarmente negativi. Gli operatori di finanza etica, dunque, si impegnano a non investire nel mercato delle armi, nella produzione e nel commercio di tabacco, o anche nel gioco

[11] Cfr. E. Ippoliti, *Un filosofo a Wall Street*, cit.

d'azzardo. Altri mercati che non vengono considerati sono quelli relativi allo sviluppo o utilizzo di energia nucleare, all'ingegneria genetica, al settore petrolifero ed estrattivo e alla produzione di pesticidi. Nonostante gli alti rendimenti garantiti, questi settori generano un impatto sociale e ambientale fortemente negativo sulle future generazioni. In un'ottica di finanza etica e sostenibile, questi impatti negativi non solo rendono questo genere di investimenti ingiusti, ma esercitando un impatto negativo critico sulle possibilità di uno sviluppo sociale e ambientale buono per le generazioni future[12], sono anche economicamente non convenienti nel lungo termine.

4. Nuove proposte di trasparenza per la pratica finanziaria

Vi sono due principali controproposte alla gestione dell'informazione e alla mancanza di trasparenza all'interno della finanza classica: la prima viene direttamente dalla Comunità Europea all'interno delle politiche volte ad aiutare gli investitori a riconoscere e scegliere strategie finanziarie sostenibili; la seconda dalla pratica ormai consolidata di practiotiners e studiosi di forme di finanza etica e sostenibile.

La prima controproposta è quella relativa alla *Sustainable Financial Disclosure Regulation* (SFDR), introdotta a partire dal 2018 ed entrata in vigore dal 2021 come strategia centrale al Piano d'Azione per la Finanza Sostenibile, insieme alla Tassonomia EU[13] e alla *Low Carbon Benchmarks Regulation*[14]. La SFDR intende migliorare la chiarezza e la comparabilità delle strategie di disclosure all'interno di prodotti e policy finanziarie proponendo una griglia comune a tutti gli agenti finanziari che permetta una lettura e comprensione omogenea per gli investitori, così che possano comprendere agevolmente la direzione delle linee d'investimento cui partecipano. La *Regulation* si intreccia quindi alle politiche proposte a livello europeo per la promozione di una produzione industriale e di uno sviluppo economico sostenibili; in questo senso, la SFDR rappresenta l'innovazione del settore finanziario necessaria e complementare all'innovazione del settore industriale ed economico auspicalmente promossa dalle politiche europee di sostenibilità. Questa strategia include quindi una chiara richiesta di maggiore trasparenza, che permetta agli investitori di integrare i rischi relativi alla sostenibilità socio-ambientale nelle loro decisioni di investimento e permetta loro di valutare i potenziali impatti avversi dei loro investimenti sull'ambiente e sulla società. Nello specifico, la SFDR ha il compito di elencare e descrivere gli strumenti finanziari utilizzati negli investimenti e durante le operazioni finanziarie, di certificare e garantire che i flussi finanziari prodotti dalle

---

[12] Per un approfondimento su forme di giustizia transgenerazionale, si vedano T. Andina, F. Corvino, *Transgenerational social structures and fictional actors: community-based Responsibility for future Generations*, in «The Monist», 106 n. 2, 2023, pp.150-164 o J. Roemer, *The Sustainabilitarian Approach: Utilitarianism, Discounting of future Welfare levels, and Sustainability,* in S. Gardner (a cura di), *The Oxford Handbook of Intergenerational Ethics*, Oxford University Press, Oxford, 2021.

[13] Cfr.: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32020R0852

[14] Cfr.: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R1011-20220101

diverse operazioni non finanziano quelli che vengono considerati mercati non etici o controversi.

Sebbene la SFDR rappresenti un primo, importante passo a livello istituzionale ed europeo, molti attori della finanza etica e sostenibile non la considerano ancora sufficiente.[15] Una prima ragione di insoddisfazione è legata all'elemento di volontarietà della disclosure: i singoli intermediari finanziari possono scegliere se fornire o meno maggiore trasparenza sui propri prodotti e strategie d'investimento. L'elemento di volontarietà è inoltre spia del fatto che il nuovo approccio alla trasparenza promosso dalla SFDR non rappresenta un vero e proprio nuovo paradigma all'interno della pratica finanziaria, quanto piuttosto un'aggiunta non strutturale a quell'apparato finanziario che manipola l'informazione senza renderla realmente accessibile o comprensibile.

In particolar modo, l'elemento di volontarietà sembra porre due questioni, una prima di ordine più teorico, l'altra più pratica. La prima questione relativa all'efficacia di una disclosure di questo genere, riguarda il fatto se una pratica volontaria sia sufficientemente forte da far sì che, partendo dall'evoluzione della pratica finanziaria nel senso da essa auspicato, questa possa avere poi degli effetti tali da portare a riconoscere concetti quali la sostenibilità a lungo termine e la preservazione del valore economico per le future generazioni come cardini teorici di un nuovo approccio economico-finanziario. La seconda questione è, come accennato, di ordine più pratico: una disclosure su base esclusivamente volontaria può riuscire a coinvolgere un numero "critico" di practitioners tale da generare un impatto significativo e, a posteriori, poter valutare l'effettiva efficacia di un approccio finanziario più trasparente ed etico?

Una proposta radicalmente alternativa è invece rappresentata dalla finanza etica e sostenibile, che propone un modello di governance nuovo, che si mostri in grado di includere gli investitori e la società come attori partecipi all'interno della pratica finanziaria, garantendo loro una piena inclusione come accesso a informazioni complete e facilmente comprensibili. Secondo il Manifesto della Finanza Etica proposto dall'Associazione Finanza Etica nel 1998, la trasparenza è una forma di garanzia di partecipazione perché interna alla pratica finanziaria stessa: «I depositanti hanno il diritto di conoscere i processi di funzionamento dell'istituzione finanziaria e le sue decisioni di impiego e di investimento. Sarà cura dell'intermediario eticamente orientato mettere a disposizione gli opportuni canali informativi per garantire la trasparenza sulla sua attività»[16]. Attraverso meccanismi di governance trasparenti, la partecipazione attiva degli investitori assume quindi i contorni di «meccanismi diretti di indicazione delle preferenze nella destinazione dei fondi, sia meccanismi

[15] Interessante a tal proposito l'intervista rilasciata da Anna Fasano, presidente di Banca Etica: https://fundspeople.com/it/entra-in-vigore-lsfdr-ma-la-finanza-etica-e-unaltra-faccenda/

[16] IL MANIFESTO DELLA FINANZA ETICA, a cura di Associazione Finanza Etica e Bancaetica – 1998, p. 1.

democratici di partecipazione alle decisioni. La finanza etica in questo modo si fa promotrice di democrazia economica»[17].

In quest'ultimo caso, la trasparenza nell'attività e nei prodotti finanziari si muove all'interno di un paradigma radicalmente alternativo perché a differenza dell'approccio finanziario efficientista classico, l'apertura e la circolazione di informazioni non si comporta come uno strumento (generalmente disatteso) per stabilire il valore o il prezzo di un prodotto, né si comporta, come nel caso della SFDR, come parte di una strategia per forzare i temi della sostenibilità all'interno dei mercati finanziari. All'interno di un approccio finanziario etico e sostenibile, la trasparenza è uno strumento potente e imprescindibile nella pratica finanziaria.
Una tale importanza può essere letta tanto da un punto di vista pratico-applicativo, quanto da un punto di vista teorico. Nel primo caso, come chiaramente stabilito nel Manifesto, la trasparenza è uno degli strumenti necessari a rendere la pratica finanziaria stessa più democratica ed equa e renderla un motore efficiente per uno sviluppo economico-sociale equo e sostenibile. L'equità e la sostenibilità, infatti, vengono garantite tanto dalla disponibilità di informazione, quanto dalla sua inclusività: quanto più l'informazione è trasparente, dunque, tanto più sarà comprensibile ai diversi attori coinvolti nella pratica finanziaria. In questa prospettiva, volendo continuare l'analisi del concetto di trasparenza da un punto di vista teorico[18], questo assume una centralità assoluta, tanto da poter essere utilizzato come concetto cardine, che discrimina la teoria economica e finanziaria etica e sociale da altre teorie, in particolar modo da quella classica. Il ruolo ricoperto dal concetto di trasparenza all'interno di quest'ultimo approccio, infatti, potrebbe essere equiparato e contrapposto al ruolo ricoperto dal concetto di informazione all'interno della teoria classica dei mercati; garantendo così una chiara distanza tanto nell'approccio teorico quanto negli effetti pratici.

[17] *Ibidem*, p. 2.
[18] Cfr. C. Felber, *Gemeinwohl-Ökonomie, das alternative Wirtschaftsmodell für Nachhaltigkeit*, Piper Verlag, München, 2018.

Delle trasparenze economiche[a]

Gian Vito Zani[*]

Abstract
Il presente articolo esplora il concetto di trasparenza nell'ambito economico, concentrandosi sulle prospettive offerte dalla Scuola Austriaca di economia, qui rappresentata da Ludwig von Mises e Friedrich von Hayek. L'analisi si articola in tre parti: la prima introduce il ruolo centrale della trasparenza nel discorso economico e nei suoi dibattiti sulle crisi. La seconda esamina la teoria di Mises, per cui la trasparenza del mercato è essenziale in quanto qualsiasi interferenza statale rappresenta una distorsione delle informazioni e delle dinamiche economiche. La terza si concentra sull'approccio di Hayek, per il quale, diversamente da Mises, il mercato opera in un contesto di opacità intrinseca, dove la conoscenza è dispersa e spesso tacita. Questi autori offrono interpretazioni opposte della trasparenza per giungere al medesimo obiettivo, cioè, rendere il mercato un'istituzione incontestabile. Il lavoro evidenzia come attraverso il binomio trasparenza/opacità, si siano sviluppati strumenti concettuali per legittimare il primato del mercato, negando al contempo validità alle possibili alternative.

Parole chiave: trasparenza, Scuola Austriaca di economia, von Mises, von Hayek, neoliberalismo

This article explores the concept of transparency in economics, focusing on the perspectives offered by the Austrian School of Economics, as represented by Ludwig von Mises and Friedrich von Hayek. The analysis is divided into three parts: the first introduces the central role of transparency in economic discourse and its debates on crises. The second examines Mises's theory, which posits that market transparency is essential, as any state interference distorts information and economic dynamics. The third focuses on Hayek's approach, which, in contrast to Mises, views the market as operating in a context of intrinsic opacity, where knowledge is dispersed and often tacit. These authors provide opposing interpretations of transparency to reach the same goal: making the market an uncontestable institution. The study highlights how the transparency/opacity dichotomy has been used to develop conceptual tools that

[a] Saggio ricevuto in data 24/05/2024 e pubblicato in data 22/01/2025.
[*] Borsista di ricerca, Università di Torino, email: gianvito.zani@unito.it.

legitimize the primacy of the market while simultaneously dismissing the validity of alternative models.

Keywords: Transparency, Austrian School of Economics, von Mises, von Hayek, Neoliberalism

Introduzione

Nel seguente articolo si vuole indagare il concetto di trasparenza in ambito economico e più specificatamente in riferimento alla Scuola austriaca di economia – in due suoi rappresentanti chiave del Novecento, Ludwig von Mises e Friedrich von Hayek – la cui influenza negli ultimi quarant'anni è sempre più forte nell'informare il dibattito riguardante le politiche economiche di diverse istituzioni[1]. Nell'analisi si tenterà di mostrare il duplice ruolo che la nozione di trasparenza ha all'interno del discorso economico in generale e di tale Scuola economica in particolare. Infatti, se da un lato, in Mises, è proprio in nome della trasparenza che viene negato qualsiasi ruolo dello Stato nel gioco economico, in quanto "falsa le carte", dall'altro, in Hayek, lo Stato non può intervenire perché il mercato e l'informazione che trasmette sono in sé frutto di un soggetto e di una società tutt'altro che trasparente. Ciò che accomuna le prospettive è l'uso, seppur inverso, della coppia trasparenza/opacità per inibire l'intervento pubblico che, come si mostrerà, poggia su due visioni differenti della società umana, da un lato cristallina dall'altro nebulosa.

Nella prima parte si analizzerà brevemente come il concetto di trasparenza sia centrale nel discorso economico: a riprova esso diventa un acceso punto di dibattito durante ogni spiegazione delle crisi economiche. Nelle due parti successive si analizzerà come tale concetto sia sviluppato da due rappresentanti del neoliberalismo[2]:

[1] A titolo di esempio Bruce Caldwell evidenzia come, durante il periodo precedente la prima campagna presidenziale di Barack Obama e quello successivo alle elezioni, il testo *La via della servitù* di Hayek ha subito un rialzo delle vendite, cfr. B. Caldwell, *The secret behind the hot sales of 'The Road to Serfdom' by free-market economist F. A. Hayek*, in «Washington Post», 17 febbraio 2010. Rispetto all'influenza della prospettiva della Scuola austriaca nelle politiche economiche cfr. R. G. Holcombe, *Public Choice and Austrian Economics*, in C. J. Coyne, P. Boettke (a cura di), *The Oxford Handbook of Austrian Economics*, Oxford Academic Online, 2015, pp. 491-507; A. Godart-van der Kroon, J. Salerno (a cura di), *The Austrian School of Economics in the 21st Century: Evolution and Impact*, Springer, Cham 2023.

[2] Per una ricostruzione delle diverse famiglie neoliberali e delle rispettive differenze, cfr. S. Audier, *Néo-libéralisme(s). Une archéologie intellectuelle*, Éditions Grasset & Fasquelle, Paris 2012; D. Plehw, Q. Slobodian, P. Mirowski (a cura di), *Nine Lives of Neoliberalism*, Verso, Brooklyn 2020, B. Walpen, *Die offenen Feinde und ihre Gesellschaft. Eine hegemonietheoretische Studie zur Mont Pèlerin Society*, VSA-Verlag, Hamburg 2004; P Mirowski, D. Plehwe (a cura di), *The Road from Mont Pelerin. The Making of the Neoliberal Thought Collective*, Harvard University Press, Cambridge (MA)-London 2009. Un cenno sulla scelta di utilizzare il termine neoliberalismo e non neoliberismo: secondo gli autori neoliberali non c'è distinzione tra liberalismo e liberismo. L'impossibilità di giustificare tale distinzione è ben esemplificata dal neoliberale Wilhelm Röpke che, come il suo amico Luigi Einaudi, si confrontò proprio con l'autore di tale distinzione, Benedetto Croce. Secondo Croce, il liberismo, cioè la prospettiva riguardante la libertà legata al piano economico, intesa come *laissez-faire*, si situerebbe in una posizione di secondarietà rispetto al liberalismo, che invece si affaccia su un orizzonte etico e

Mises e Hayek. Nella parte conclusiva, infine, si mostrerà come le differenti argomentazioni dei due autori rispetto al concetto di trasparenza trovano fondamento ultimo su visioni differenti di cosa sia l'ambiente sociale in cui l'essere umano si muove.

1. L'economia trasparente

Quanto la trasparenza sia un concetto rilevante nel discorso economico è subito evidente se si nota il ruolo chiave che essa ha in uno dei suoi teoremi cardine, cioè quello della concorrenza perfetta – *modus operandi* necessario per un leibniziano miglior mondo possibile. Tra i suoi vari requisiti è infatti necessario che tutti gli attori del mercato, consumatori e produttori, abbiano una informazione completa e simmetrica[3]. Completezza e simmetricità che garantiscono ai diversi attori la trasparenza del mercato e quindi la possibilità del raggiungimento di un punto di equilibrio. Che il ruolo della trasparenza sia necessario per l'esistenza del mercato stesso è evidenziato dall'influente articolo di George Akerlof *The Market for 'Lemons': Quality Uncertainty and the Market Mechanism*[4], per cui l'asimmetria informativa, cioè la non trasparenza, può comportare il possibile fenomeno del *missing market*.

Quanto il discorso economico sia legato all'idea di trasparenza risulta ancora più evidente se si fa riferimento ai diversi dibattiti avvenuti durante e dopo la crisi economica del 2007-2008, che ha avuto nel fallimento della *Lehman Brothers* la sua epifania. La celebre domanda che la regina Elisabetta II poneva ai membri della *London School of Economics* il 5 novembre 2008, cioè in piena crisi economica, sul perché gli economisti non fossero stati in grado di prevedere il *crack* economico si fonda sulla fede – o su una possibile critica – della trasparenza del mercato. Infatti, o tale trasparenza non si dà, e allora l'intero discorso economico si fonda su un presupposto

politico, tanto che è possibile pensare un sistema liberale con un ordinamento economico non di mercato (cfr. B. Croce, L. Einaudi, *Liberismo e liberalismo*, Riccardo Ricciardi, Milano-Napoli 1957). Secondo Röpke tale posizione è insostenibile perché è impossibile separare la libertà economica dalle altre libertà: «tutte sono ugualmente importanti e si condizionano a vicenda». Röpke stesso propone due argomenti per rifiutare la distinzione crociana: il primo è che la libertà economica «rappresenta essa stessa un importante settore della libertà» di cui gode un individuo, e se essa è limitata lo è anche l'individuo nelle sue scelte. Il secondo argomento, che evidenzia uno dei cardini del pensiero neoliberale, è che «venendo meno la libertà economica – la quale si sostanzia non solo nella libertà dei mercati, ma anche nella proprietà privata – la libertà spirituale e politica perde le sue vere basi» (W. Röpke, *L'errore di Croce: la distinzione tra liberalismo e liberismo*, in Id. *Umanesimo liberale*, Rubbettino, Soveria Mannelli 2000, pp. 121-127, qui p. 125). Si può notare come la seconda argomentazione tenda a ribaltare la gerarchia proposta da Croce evidenziando il fulcro della prospettiva neoliberale: non solo la libertà economica non è inferiore alle altre libertà, ma è anzi quella che, permettendo la libertà nell'esistenza materiale dell'individuo, garantisce la libertà spirituale e politica.

[3] La teoria della concorrenza trova la sua prima formulazione in L. Walras, *Elementi di economia politica pura* (1874), tr. it. di A. Bagiotti, Utet, Torino 1974.

[4] G. Akerlof, *The Market for 'Lemons': Quality Uncertainty and the Market Mechanism*, in «The Quarterly Journal of Economics», 84, n. 3, 1970, pp. 488-500. Quanto questo ambito di ricerca sia ancora dibattuto è dimostrato dal conferimento del cosiddetto premio Nobel per l'economia a Joseph Stiglitz, Michael Spence e allo stesso Akerlof per i loro studi sull'economia dell'informazione.

inesistente, oppure essa c'è in una qualche misura, ma nessuno, nel caso specifico, ha voluto controllare e guardare attraverso tale trasparenza. Ovviamente il discorso economico ha immediatamente optato per la seconda strada, tentando così di salvare il concetto di trasparenza, attraverso una piccola contorsione. Il movimento di ribaltamento della questione è stato il seguente: non è che la crisi non sia stata prevista perché il mercato non è trasparente, ma è la mancanza di trasparenza che ha causato la crisi di mercato. Detto in altre parole, la crisi non è stata endogena al mercato, per un suo cattivo funzionamento, ma esogena, è stata dovuta alle cattive informazioni su cui il mercato era costretto a processare[5].

La crisi del 2007 è stata dovuta all'asimmetria di informazioni tra i vari attori, e quindi alla mancanza di trasparenza nel mercato. Un esempio è dato dalla nascita di prodotti poco trasparenti, i *Collaterelised Debt Obligations*, che proprio a causa della loro poca chiarezza hanno dato vita alle *Credit Rating Agencies*, che valutano esse stesse tali prodotti con criteri non chiari: il tutto gestito prevalentemente dalle cosiddette banche ombra – nome che già da sé è poco trasparente. Ad aggravare tutto, secondo questa lettura, gli istituti regolatori non hanno avuto la capacità – se non la volontà – di intervenire in tempo. La mancanza di trasparenza, quindi, se forse non è stata da sé sola causa della crisi è certamente una delle sue cause rilevanti. Per evitare le crisi, di conseguenza, la soluzione è rendere il mercato trasparente: da qui la proposta, per esempio del Fondo monetario internazionale e del *Financial Stability Board*[6], di una migliore regolazione e supervisione, che da sé, rendendo il mercato trasparente, lo renderebbe pure sicuro e funzionante. Anche un partigiano del libero mercato come Richard Posner, tra le varie indicazioni per una migliore regolazione post-crisi, ha proposto l'istituzione di «un'agenzia di *intelligence* finanziaria» che abbia tra i suoi compiti quello di porre «attenzione alla raccolta e all'analisi delle informazioni»[7]. Come osserva criticamente Oliver Kessler, le soluzioni proposte dalle diverse istituzioni economiche e dai loro *board* di economisti sono la «concettualizzazione della questione dell'instabilità dei mercati finanziari globali come un problema informativo che può essere risolto con "più dati"»[8].

---

[5] Ovviamente questa non è l'unica spiegazione della crisi proposta dalla scienza economica. Per una lettura diversa, di stampo neo-keynesiano, cfr. S. Keen, *A monetary Minsky model of the Great Moderation and the Great Recession*, in «Journal of Economic Behavior & Organization», 86, 2013, pp. 221-235; per una critica generale al modello economico ortodosso cfr. E. Brancaccio, *Anti-Blanchard. Un approccio comparato allo studio della macroeconomia*, Franco Angeli, Milano 2021.

[6] Cfr. Financial Stability Board, The International Monetary Fund, *The Financial Crisis and Information Gaps*, 2009, disponible al link: https://www.imf.org/external/np/g20/pdf/102909.pdf; Financial Stability Board, International Monetary Fund, Bank for International Settlement, *Report to G20 Finance Ministers and Governance, Guidance to Assess the Systemic Importance of Financial Institutions, Markets and Instruments: Initial Considerations*, 2009, disponibile al link; https://www.imf.org/external/np/g20/pdf/100109.pdf.

[7] R. A. Posner, *La crisi della democrazia capitalista* (2010), tr. it. di M. Cupellaro, EGEA, Milano 2010, pp. 285-286.

[8] O. Kessler, S*leeping with the enemy? On Hayek, constructivist thought, and the current economic crisis*, «Review of International Studies», 38, pp. 275-299: 282.

Quanto questa visione che vede nella trasparenza la virtù che permette al mercato di funzionare sia ancora presente, ora che la crisi economica sembra alle spalle, lo dimostrano le parole del Presidente della *Federal Reserve* Jerome Powell del 25 marzo 2021, che, intervistato riguardo ai possibili cambiamenti nella politica monetaria statunitense, ha affermato che essi avverranno «molto gradualmente nel tempo e con grande trasparenza»[9]. Questa precisa affermazione, che potremmo definire la *summa* delle cosiddette politiche ortodosse neoliberali, è stata oggetto delle critiche, a causa della manipolazione centrale della moneta che comporta[10], di un'altra corrente neoliberale, quella austriaca, che trova in Mises e Hayek i due suoi esponenti principali nel XX secolo. Nei prossimi paragrafi si analizzerà la loro concettualizzazione della trasparenza prendendo come spunto di riflessione le loro diverse argomentazioni sulla possibilità del calcolo economico.

2. *Mises e la* Glasnost *del mercato*

La fama di Mises è sicuramente legata alla sua teoria della impossibilità del calcolo economico all'interno di una società socialista. La Rivoluzione d'Ottobre, e la conseguente nascita di un governo socialista permisero per la prima volta nella storia di provare su ampia scala un modo di produzione, cioè di allocazione efficiente delle risorse scarse date, che non fosse quello capitalista. Eppure, già nel 1920 Mises nel suo articolo *Die Wirtschaftsrechnung im sozialistischen Gemeinwesen*[11] affermava che tale modello di produzione era impossibile. La tesi, frutto del metodo *a priori* di Mises che parte dalla nozione fondamentale di "azione umana", è piuttosto semplice: il calcolo economico è possibile se ci sono prezzi, ma i prezzi si formano solo laddove c'è un mercato di beni che appartengono a qualcuno, cioè dove esistono diritti di proprietà privata. Infatti, dove viene a mancare la proprietà dei beni essi non possono essere né scambiati né prezzati, e senza prezzo non è possibile alcun calcolo economico e quindi alcuna allocazione razionale dei beni. La cosa importante da rilevare nella teoria di Mises sul calcolo economico è il ruolo chiave che riconosce alla moneta: egli afferma che «il calcolo economico può comprendere ogni cosa che venga scambiata contro moneta»[12]. La moneta, in quanto veicolo del calcolo economico, permette, grazie al prezzo, di valutare attraverso numeri cardinali quelle che sono preferenze soggettive ordinali[13]. Non va infatti dimenticato che Mises, come tutta l'economia neoliberale e

[9] National Public Radio, *Morning Edition*, 25 marzo 2021.
[10] Tra i molti articoli critici cfr. Fred Shostak, *Fed Transparency Won't Get Us out of the Mess the Fed Created*, disponibile al link: https://mises.org/mises-wire/fed-transparency-wont-get-us-out-mess-fed-created, 2021.
[11] L. von Mises, *Die Wirtschaftsrechnung im sozialistischen Gemeinwesen* in «Archiv für Sozialwissenschaft und Sozialpolitik», 47, 1920-21, pp. 86-121.
[12] L. von Mises, *L'azione umana* (1949), tr. it. di T. Bagiotti, Rubbettino, Soveria Mannelli 2016, p. 259; cfr. ivi p. 283; cfr. T. Bagiotti, *L'opera di L. von Mises, con alcune considerazioni sul determinismo e l'indeterminazione in economia*, in «Giornale degli Economisti e Annali di Economia», Nuova Serie, 17, n. 11/12, 1958, pp. 612-637.
[13] Cfr. L. von Mises, *L'azione umana*, cit., p. 247.

non solo, muove sempre da una concezione soggettiva del valore: il valore dei diversi beni è deciso dal soggetto, non esiste un marxiano valore oggettivo. Le diverse decisioni soggettive sul valore dei beni, le cosiddette preferenze del consumatore, sono esprimibili attraverso numeri ordinali, e in quanto tali non permetterebbero nessun calcolo, ma, attraverso il *medium* della moneta, tali preferenze ordinali possono trovare espressione in numeri cardinali, e così permettere il calcolo economico.

Da tutto questo deriva la possibilità del processo di mercato come ambito della sovranità dei consumatori: «il calcolo effettuato in termini di prezzi monetari è quello degli imprenditori che producono per i consumatori di una società di mercato»[14]. Detto in altri termini, i prezzi permettono di fare incontrare da un lato le preferenze dei consumatori e dall'altro i calcoli degli imprenditori che sono «tenuti a obbedire incondizionatamente agli ordini del capitano, che è il consumatore»[15]. Tutto il sistema si sostiene, come evidente, su una cosa: la moneta. Da qui l'intransigenza di Mises e dei suoi epigoni[16] verso qualsiasi trattamento o interferenza su questo *medium*, per tutelare non soltanto la trasparenza della moneta, quanto anche quella del mercato stesso e dei suoi calcoli.

Infatti, è del tutto evidente che un'alterazione del valore della moneta ha come effetto immediato non l'impossibilità di calcolo, ma l'impossibilità di un calcolo corretto, in quanto i dati non sono più validi: essi non trasmettono in maniera trasparente le preferenze ordinali dei consumatori. Il caso classico di tale interferenza è per Mises l'espansione o contrazione della moneta decisa dalle Banche centrali dei diversi Stati: una crescita della moneta circolante, falsando i calcoli, tenderà a permettere investimenti in attività non redditizie, e lo stesso effetto si ha con una contrazione della moneta[17]. Per evitare questa interferenza Mises propone da un lato un anacronistico ritorno al *Gold Standard*[18], sistema che permette di limitare le azioni dei singoli Stati sulle loro Banche centrali, dall'altro la nascita di un sistema di *free banking*[19], che permetterebbe che «il solo veicolo dell'espansione creditizia [sia] il credito di circolazione»[20].

È interessante notare come in Mises a sporcare il mercato, cioè a rendere il calcolo non corretto, sia sempre l'azione dello Stato[21]. Essa avviene non solo direttamente con l'espansione della moneta, ma anche indirettamente, per esempio attraverso l'emissione di titoli del tesoro, che sottraggono le risorse degli investimenti

---

[14] Ivi, p. 263.

[15] Ivi, p. 319.

[16] Sul tema sono presenti molti articoli nell'archivio del *Mises Institute*, raggiungibile al link: https://mises.org/.

[17] Cfr. L. von Mises, *L'azione umana*, cit., p. 588.

[18] Cfr. ivi, pp. 515-518.

[19] Cfr. ivi, pp. 489-493.

[20] Ivi, p. 480.

[21] Come afferma Richard Gonce per Mises «solo il libero mercato può creare prezzi concorrenziali e solo i prezzi concorrenziali sono qualificabili come prezzi razionali», cfr. R. A. Gonce, *Natural Law and Ludwig von Mises' Praxeology and Economic Science*, in «Southern Economic Journal», 39, n. 4, 1973, pp. 490-507: 504.

«alle leggi ferree del mercato e alla sovranità dei consumatori»[22], o la tassazione, definita «interferenza fiscale»[23]. La motivazione del perché sia proprio lo Stato a interferire nel mercato è dovuta al fatto che esso è l'unico attore economico che non deve rispettare le regole del mercato stesso per sopravvivere – per esempio, esso è sempre solvibile in quanto può fare uso della coercizione per l'imposizione fiscale[24].

La strenua battaglia di Mises contro l'interferenza possibile sul calcolo economico causata dall'intervento statale trova la sua giustificazione ultima non solo nel criterio di efficienza – se il calcolo non è corretto gli investimenti sono errati – ma soprattutto nel dato che «il risultato di questi sforzi [i molteplici calcoli economici] non è soltanto la struttura dei prezzi, ma anche la struttura sociale, l'attribuzione di specifici compiti ai vari individui»[25]. La trasparenza della moneta e la conseguente nascita di prezzi corretti non solo garantisce un mercato trasparente, e quindi capace di fare investimenti produttivi, ma permette anche una società trasparente, dove ognuno sta al posto che merita in quanto «assegnare a ognuno il suo posto nella società è compito dei consumatori»[26]. Così facendo, la trasparenza del prezzo si riverbera in tutte le sfere della vita umana. Da qui lo Stato come nemico, in quanto ente che falsando il prezzo opacizza tutte le sfere, e il mercato capitalista come amico, in quanto unico processo che rendendo possibili i prezzi rende possibile anche una società trasparente: «non esistono prezzi al di fuori del mercato, né essi possono essere creati artificialmente»[27].

Proprio sulla possibilità di creazione di prezzi artificiali nasce la celebre risposta di Abba Lerner e Oskar Lange [28] a Mises: essi affermano che anche in una economia socialista è possibile avere prezzi attraverso un processo per tentativi ed errori coordinato da un Ufficio centrale e questo permetterebbe di raggiungere tutte le virtù legate al prezzo corretto – sviluppo economico e società trasparente – evitando lo sfruttamento tipico di una economia capitalista. Proprio per evitare questa risposta, Hayek sviluppa un'argomentazione sulla impossibilità del calcolo economico in

[22] L. von Mises, *L'azione umana*, cit., p. 272.
[23] Ivi, p. 779.
[24] Cfr. ivi, p. 272.
[25] Ivi, p. 360.
[26] Ivi, p. 325.
[27] Ivi, p. 441
[28] A. Lerner, *Economy Theory and Socialist Economics*, in «Review of Economic Studies», 2, n. 1, 1934, pp. 51-61; O. Lange, *On the Economic Theory of Socialism*, in «Review of Economic Studies», 4, n.1, 1936, pp. 53-71 e 4, n. 2, 1937, pp. 123-142. Per una critica di stampo misesiano a tale proposta cfr. R. W. Garrison, *Mises and His Methods*, in J. M. Herbener (a cura di), *The Meaning of Ludwig von Mises: Contributions is Economics, Sociology, Epistemology, and Political Philosophy*, Kluwer Academic Publishers, Boston 1993, pp. 102-117. Riguardo al dibattito sul calcolo economico in una economia socialista J. Persky, *Retrospectives: Lange and von Mises, Large-Scale Enterprises, and the Economic Case for Socialism*, in «The Journal of Economic Perspectives», 5, n. 4, 1991, pp. 229-236; B. Jossa, *Socialismo e mercato: contributi alla teoria economica del socialismo*, ETAS, Milano 1978.

ambito socialista che, seppur spesso associata a quella di Mises in quanto per alcuni aspetti simile[29], si fonda su presupposti totalmente differenti.

3. Hayek: del fare senza sapere

L'argomentazione di Hayek sull'impossibilità del calcolo economico in ambito socialista, come evidenziano chiaramente i titoli dei suoi due articoli sull'argomento *Economics and Knowledge*[30] del 1937 e *The Use of Knowledge in Society*[31] del 1945, si fonda sul nesso tra economia di mercato e conoscenza[32]. Hayek, pur riprendendo da Mises la tesi relativa alla necessità della proprietà privata per consentire lo scambio e della moneta per rendere le preferenze ordinali del consumatore grandezze cardinali, apporta due notevoli novità. La prima riguarda lo statuto dei prezzi: in Hayek «le aspettative di prezzo e persino la conoscenza dei prezzi correnti costituiscono […] solamente una parte molto piccola del problema della conoscenza»[33]. Il sistema dei prezzi, come afferma anche Mises, attraverso il processo di monetarizzazione delle preferenze informa e dà conoscenza delle volontà dei consumatori, ma questa è solo *una parte* della conoscenza. Per capire cosa in realtà il mercato metta in moto con i prezzi, e qui troviamo la seconda novità rispetto a Mises, bisogna cambiare metodologia e superare l'approccio della logica delle azioni del consumatore e del produttore: «la logica delle azioni individuali è un *a priori*, ma nel momento stesso in cui si passa da questo all'interazione fra molte persone si entra in un campo empirico»[34]. Detto in altri termini, l'analisi di Mises del calcolo economico, escludendo un'analisi empirica di ciò che avviene nel mercato ma basandosi solamente sull'analisi *a priori* della fenomenologia del prezzo, da un lato non esclude una replica *à la* Lerner e Lange – in quanto anche loro propongono un modello di monetarizzazione – e dall'altro non permette di cogliere qual è l'essenza del processo di mercato: non semplicemente di trasmettere informazioni e conoscenza, ma di attivare o, ci pare meglio dire, *estrarre* conoscenza dagli individui.

La critica hayekiana riguardo all'impossibilità del calcolo economico in una società socialista è la seguente: se veramente gli esseri umani avessero a disposizione tutta la conoscenza, fossero onniscienti, il problema economico dell'utilizzo dei mezzi scarsi avrebbe una soluzione puramente logica (come vogliono tra l'altro Mises, Lerner e Lange); partendo però da un punto di vista empirico «la conoscenza delle circostanze di cui ci dobbiamo servire non esiste mai in forma concentrata o integrata,

---

[29] Cfr. R. M. Ebeling, *The Life and Works of Ludwig von Mises*, in «The Independent Review», 13, n. 1, 2008, pp. 99-109.

[30] F. A. von Hayek, *Economics and Knowledge*, in «Economica» 4, n. 13, 1937, pp. 33–54.

[31] F. A. von Hayek, *The Use of Knowledge in Society*, in «The American Economic Review», 35, n. 4, 1945, pp. 519-530.

[32] Cfr. M. Boccaccio, *Hayek. Teoria della conoscenza e teoria economica*, Laterza, Roma-Bari 1996.

[33] F. A. von Hayek, *Economia e conoscenza* (1937), in Id., *Competizione e conoscenza*, Rubbettino, Soveria Mannelli 2017, pp. 33-60: 50.

[34] F. A. von Hayek, *Autobiografia* (1994), tr. it. di E. Campani, Rubbettino, Soveria Mannelli 2011, p. 47.

ma solamente sotto forma di frammenti dispersi di conoscenza, incompleta e spesso contraddittoria, che gli individui posseggono separatamente»[35]. La questione del calcolo economico diviene allora la questione di come mettere in connessione, e quindi utilizzare, questa conoscenza dispersa[36]. Le soluzioni a tale problema sono due: o si trasmette tutta questa conoscenza a una istituzione pianificatrice, oppure è necessaria una istituzione che permetta l'utilizzo coordinato della conoscenza dispersa.

La realizzabilità di queste due soluzioni risiede nella nozione di conoscenza, che nel caso del mercato si declina non solo come conoscenza scientifica, per esempio del processo di produzione di X, ma nel senso di «conoscenza delle circostanze particolari di tempo e di luogo»[37]. Ora, come è evidente, più la conoscenza riguarda circostanze particolari, più è probabile che risieda negli individui e non in una istituzione centralizzata. Ma ciò non spiega ancora perché una pianificazione centralizzata sia poco efficiente; infatti, se questa informazione venisse trasmessa al pianificatore, egli avendo una conoscenza generale più ampia potrebbe farne un uso migliore – questa, in fondo è l'idea che sta alla base delle Banche centrali che decidono il tasso di interesse partendo dai dati che raccolgono dal mercato[38]. Ciò che Hayek afferma con l'argomento della conoscenza dispersa non è tanto la difficoltà di un tale sistema di comunicazione, quanto l'impossibilità della comunicazione della conoscenza stessa, in quanto nemmeno l'individuo sa di possederla. Ciò che caratterizza la conoscenza umana, infatti, è che essa è principalmente tacita: l'essere umano ha la capacità di fare senza sapere – il caso classico è il bambino che utilizza correttamente le regole grammaticali della sua lingua senza conoscerle[39]. Ora, ciò che l'istituzione mercato permette non è solamente la comunicazione di alcune informazioni mediante il sistema dei prezzi, ma anche l'attivazione di tale conoscenza tacita, incomunicabile, che ciascuno ha. Ancor prima di comunicare una conoscenza il mercato permette di estrarla dagli individui: «nessuno può comunicare ad un altro tutto ciò che egli conosce, perché molta dell'informazione che egli può usare verrà fuori soltanto mentre farà piani per l'azione»[40]. Ne consegue che la comunicazione non è possibile a causa dell'opacità del soggetto a sé stesso, e mancando tale

---

35 F. A. von Hayek, *L'uso della conoscenza nella società* (1945), in Id., *Competizione e conoscenza*, cit., pp. 61-74: 62.

36 Per una ricostruzione dell'evoluzione del pensiero di Hayek dai problemi economici a quelli epistemologici cfr. J. Birner, *La place de 'Sensory Order' dans l'oeuvre de F. A. Hayek*, in «Cahiers d'économie politique», n. 51, 2006, pp. 109-138.

37 F. A. von Hayek, *L'uso della conoscenza nella società*, cit., p. 64.

38 Sui problemi epistemologici di questa prospettiva cfr. S. Morris, H. Song-Shin, *Central Bank Transparency, and the Signal Value of Prices*, in «Brookings Papers on Economic Activity», n. 2, 2005, pp. 1-43.

39 Cfr. F. A. von Hayek, *Il primato dell'astratto* (1969), in Id., *Nuovo studi di filosofia politica, economia e storia delle idee* (1978), tr. it. di G. Minotti, Armando Editore, Roma 1988, pp. 45-59: 48-49.

40 F. A. von Hayek, *La presunzione fatale: gli errori del socialismo* (1988), tr. it. F. Mattesini, Rusconi, Milano 1997, p. 135.

comunicazione vengono a mancare i dati che permettono la possibilità di una pianificazione centrale.

L'argomentazione di Hayek a favore di una economia di mercato si fonda da un lato sul «punto che i prezzi rivelano una conoscenza collettiva di tutti gli agenti economici, aggregando le diverse informazioni che essi possiedono»[41], e dall'altro sul fatto che questa conoscenza non è possibile per il pianificatore non perché «il s*hopkeeper è* il meglio piazzato per osservare i fatti economici»[42], come propongono alcuni interpreti di Hayek, ma perché gli stessi attori economici meglio piazzati non sanno di possedere tale conoscenza, motivo per cui non possono comunicarla a un ente pianificatore. Come evidenzia giustamente Kessler, la prospettiva hayekiana mostra come in realtà ciò che serve non è una migliore disponibilità di dati, ma la capacità di dare senso, un contesto, a tali dati[43]. E tale senso – aggiungiamo, precisando – lo può dare l'istituzione mercato, non il soggetto, in quanto non è detto che egli sia sempre in grado di farlo.

L'opacità della conoscenza tacita che ognuno ha rende impossibile la pianificazione: il mercato si rivela quella istituzione che è in grado di estrarre informazioni da un soggetto opaco e successivamente comunicare. Tale soggetto agisce in un contesto esso stesso opaco, in quanto emergenza, frutto inintenzionale delle azioni individuali e della loro conoscenza dispersa: «le nostre tradizioni morali, come molti altri aspetti della nostra cultura si sono sviluppate contemporaneamente alla nostra ragione, e non come suo prodotto»[44].

Quanto sia distante la trasparenza nella prospettiva di Hayek è evidente se si riflette sulla sua visione per cui «la concorrenza è una procedura di scoperta»[45]. Come visto *supra* nella parte introduttiva, la trasparenza è uno degli aspetti essenziali all'interno del concetto classico di concorrenza, il quale prescrive che tutti gli attori abbiano la stessa conoscenza e non vi sia asimmetria informazionale. Tale opzione, come evidente, in Hayek non è sostenibile, in quanto dalla sua prospettiva non si dà mai una conoscenza completa: da qui la sua idea della concorrenza come scoperta, come forza del mercato che permette di acquisire nuova conoscenza. Per Hayek non si ha concorrenza quando il soggetto X decide di aprire una nuova hamburgheria in una città dove esistono cento ristoranti simili, ma quando X, in base a conoscenze che solo lui possiede, anche in maniera accidentale – lavora in città, va spesso fuori a cena, adora gli hamburger e nel fine settimana, tornando nel paese natale, si rende conto di non poter mangiare un hamburger – decide di aprire l'unica hamburgheria in un posto differente. La concorrenza è quel processo che permette di scoprire che in quel posto manca un servizio, conoscenza che nessuno può comunicare in quanto nemmeno il

[41] S. Morris, H. Song-Shin, *Central Bank Transparency, and the Signal Value of Prices*, cit., p. 15.
[42] *Ibid.*
[43] O. Kessler, S*leeping with the enemy? On Hayek, constructivist thought, and the current economic crisis*, cit. p. 298.
[44] F. A. von Hayek, *La presunzione fatale: gli errori del socialismo*, cit., p. 39.
[45] Ivi, p. 52. Su questo argomento cfr. D. Antiseri, L. Infantino, *Nota biografica*, in F. A. von Hayek, *Conoscenza, competizione e libertà*, Rubbettino, Soveria Mannelli 1998, pp. 31-49; cfr. I. M. Kirzner, *Competition and Entrepreneurship*, University of Chicago Press, Chicago 1978.

soggetto, fino a quando ha realizzato le azioni per aprire la propria attività sapeva di avere: è l'istituzione mercato che ha estratto dal soggetto tale conoscenza.

Per Hayek il tentare attraverso la pianificazione centrale, cioè attraverso lo Stato, di organizzare gli scambi ha pertanto come effetto un uso parziale della conoscenza disponibile per l'intera società. Tale sottodimensionamento dell'uso della conoscenza è dovuto al fatto che essa è tacita, non è conosciuta *a priori* nemmeno dal soggetto, che agisce senza sapere, in quanto sia egli che la società a cui dà vita inintenzionalmente sono opachi. Il tentativo di, per usare termini cari a Hayek, sostituire con una *taxis* un *cosmos*[46], cioè di pianificare centralmente e razionalmente per rendere meno convulso e perciò trasparente il mercato, ha come effetto il condannare l'umanità a un uso parziale della conoscenza che la informa: la trasparenza ha come costo l'ignoranza.

4. La trasparenza come grimaldello

In questo lavoro si è tentato di mostrare l'uso fatto nel discorso economico cosiddetto neoliberale del concetto di trasparenza. Ciò che è rilevante notare è che esso se da un lato ha una duplice interpretazione, virtù da perseguire per gli economisti cosiddetti ortodossi e per Mises, vizio da evitare per Hayek, dall'altro è utilizzato da tutte le posizioni come grimaldello per affermare e difendere l'istituzione mercato.

Nella versione del mercato trasparente qualsiasi azione esterna al mercato è di necessità un'interferenza, qualcosa che rende opaca la visuale e non permette di fare le giuste previsioni e valutazioni. Solo se il processo di mercato è lasciato libero di agire, non trova alcuna interferenza, esso creerà una società trasparente dove ognuno sta dove merita – come ci ricorda Mises, infatti il risultato dei molteplici calcoli economici «non è soltanto la struttura dei prezzi, ma anche la struttura sociale, l'attribuzione di specifici compiti ai vari individui»[47]. La trasparenza del mercato si riversa sull'intera società solo se forze esogene, quali lo Stato, non intervengono. La prospettiva di Hayek è diametralmente opposta[48]: la società si dà sempre come opaca, perché è l'emergenza di una conoscenza tacita sia dei soggetti sia delle istituzioni, e il processo di mercato è l'unico in grado di utilizzare tale conoscenza perché accetta questa opacità, anzi, è in grado di estrarre da essa della conoscenza. Viceversa, qualsiasi tentativo di pianificazione, in nome della trasparenza e per eliminare le contraddizioni del sistema, ha come risultato un sotto uso della conoscenza totale[49].

---

[46] Con il primo termine Hayek definisce tutte quelle organizzazioni che sono la costruzione volontaria di un soggetto, o di un gruppo di soggetti, e che quindi dipendono direttamente dalla conoscenza e dalla volontà dell'autore. Diversamente, con il termine *cosmos* Hayek descrive tutte quelle formazioni che sorgono spontaneamente. Cfr. F. A. von Hayek, *Nuovi studi di filosofia, politica, economia e storia delle*, cit., pp. 84-88.

[47] L. von Mises, *L'azione umana*, cit., p. 360.

[48] Cfr. B. Manin, *Le libéralisme radical de F. A. von Hayek*, in «Commentaire», 22, n. 4, 1983, pp. 328-336; G. Pecora, *Il liberalismo anomalo di Friedrich August von Hayek*, Rubbettino, Soveria Mannelli 2002.

[49] Cfr. J. Rodrigues, *Where to Draw the Line between the State and Markets? Institutional Elements in Hayek's Neoliberal Political Economy*, in «Journal of Economics Issues», 46, n. 4, 2012, pp. 1007-1033; J.

Ciò che accomuna queste concettualizzazioni della trasparenza del mercato è il loro tentativo di rendere quest'ultimo una istituzione incriticabile[50]. Dal lato della trasparenza come virtù, infatti, se esso non funziona è sempre a causa di forze esogene che non gli permettono di svolgere il suo compito, in quanto falsano i suoi dati oggettivi; dall'altro, quello della trasparenza come vizio, il tentativo di razionalizzare il mercato porta alla sua distruzione, perché alla sua base stanno un soggetto e una società opaca da cui estrarre conoscenza. Se una cosa è trasparente è che il mercato non si tocca.

Shearmur, *Hayek, Keynes and the State*, in «History of Economics Review», 26, n. 1, 1997, pp. 68-82. Sul diverso ruolo dello Stato e delle banche centrali tra Hayek e Keynes cfr. N. Wapshott*, Keynes o Hayek. Lo scontro che ha definito l'economia moderna* (2012), tr. it di G. Carlotti, Feltrinelli, Milano 2015; T. Hoerber, *Hayek vs Keynes: A Battle of Ideas*, Reaktion Books, London 2017; A. Burgin, *The Great Persuasion: Reinventing Free Markets since the Depression*, Harvard University Press, Cambridge (MA)-London 2012.

[50] Cfr. D. Hausknost, *Opacity and Transparency. On the 'Passive Legitimacy' of Capitalist Democracy*, in «Theoria», 177, 70, n. 4, 2023, pp. 26-53.

Jelson Oliveira, *Moeda sem efígie: a crítica de Hans Jonas à ilusão do progresso*, Curitiba, Kotter Editorial, 2023, 184 pp.[a]

Lenise Moura Fé de Almeida[*]

In un mondo in cui la crisi climatica è testimoniata da disastri naturali che si moltiplicano e occupano le pagine dei giornali, la questione ambientale è urgente e necessaria anche per la filosofia. Il libro di Jelson Oliveira, *Moneta senza effigie: la critica di Hans Jonas all'illusione del progresso* [traduzione mia], presenta un'importante riflessione sul tema del Progresso nel pensiero di Hans Jonas. Secondo Oliveira, nella modernità, il Progresso è diventato il destino dell'Occidente e, riconoscendo la centralità di questo problema, Hans Jonas ne parla con la "P" maiuscola, poiché sarebbe diffuso da una prospettiva privata e individuale a una sfera pubblica e collettiva[1], divenendo fine a sé stesso (il progresso per il progresso). Questa scelta di Jonas è seguita anche da Oliveira. Per l'autore il Progresso è una grande illusione, proprio come le "monete che hanno perso la loro effigie" e, insieme a Jonas, ci presenta un interessante percorso critico e una riflessione di grande attualità.

Il libro è composto da dieci capitoli, nel primo l'autore delimita e caratterizza il problema centrale dell'intera opera, ovvero il Progresso come credenza. Nel secondo capitolo, cerca di giustificare la rilevanza di tale questione come problema filosofico. Nel terzo capitolo, l'autore riflette sul rapporto tra Progresso, sviluppo economico e limiti della natura. Nel quarto, presenta il dibattito sulla "catastrofe imminente", annunciata dagli attuali problemi ambientali, e la funzione pedagogica e politica del catastrofismo metodologico presente nella proposta jonasiana della "euristica della paura". Dal quinto capitolo in poi, siamo condotti da Oliveira a una lettura esegetica, più centrata sulle opere jonasiane che, tuttavia, promuovono e ispirano «un atteggiamento etico e politico capace di invertire l'aporia imposta dal

[a] Recensione ricevuta in data 15/02/2024 e pubblicata in data 22/01/2025.

[*] Dottoranda, Universidade Federal de Minas Gerais, borsista, Fundação de Amparo à Pesquisa do Estado de Minas Gerais, email: almeida.lmf@hotmail.com.

[1] Questo argomento verrà chiarito dall'autore nel terzo capitolo, secondo un approccio storico-filosofico. La nostra intenzione è delimitare la scelta di Oliveira per utilizzare il termine "Progresso" con la "P" maiuscola, come ha fatto Hans Jonas nel suo testo *"Reflections on Technology, Progress and Utopia"*. Cfr. H. Jonas, *Reflections on Technology, Progress and Utopia*, in «Social Research», 48, 3, 1981, pp. 411-455.

dibattito stesso» (p. 18) dove natura e civiltà sembrano essere opposti inconciliabili. Di seguito riportiamo una sintesi critica di ogni capitolo.

Nel primo capitolo, Oliveira (pp. 21-36) sviluppa una serie di argomenti che sostengono l'idea del "progresso moderno" come credenza (una fede), basata sulla speranza di un miglioramento progressivo e inarrestabile della condizione storica della'umanità. Questa credenza porta in sé un contenuto profetico e/o utopico secondo cui «il domani sarà sempre migliore dell'oggi» (p. 21). Sebbene sia possibile affermare che nell'Antichità e nel Medioevo esistessero concezioni legate all'idea di Progresso, nella Modernità questo concetto è diventato uno dei presupposti teorici centrali, e potrebbe addirittura essere considerato un dogma proprio per la sua caratteristica di credenza che ha un'origine storica nell'Illuminismo. Per questa ragione il Progresso moderno è diverso da quello sperimentato nella storia della civiltà occidentale a partire dai Greci, poiché è diventato una sorta di fede (o una utopia) che ha invaso le sfere della storia, della conoscenza, della scienza, della tecnologia, dell'economia, del lavoro, dei valori, della politica, ecc. In queste pagine lo sforzo teorico di Oliveira si concentra nel dimostrare che il Progresso è diventato la credenza fondante (un assioma) della Modernità.

Nel secondo capitolo, Oliveira (pp. 37-45) spiega perché riflettere criticamente sull'idea moderna di Progresso è diventato un compito della filosofia (o dei filosofi). Poiché è possibile affermare che il Progresso moderno è una credenza, come è stato sostenuto nel primo capitolo, la filosofia trova la possibilità di realizzarsi come un modo di mettere in discussione ciò che è dato. Tuttavia, l'idea di Progresso moderno è così radicata nel modo di vivere contemporaneo che un atteggiamento critico nei suoi confronti diventa risibile, pessimistico o condannabile. Ecco perché la filosofia deve riprendere, oggi, «la sua funzione primordiale di analisi e critica di quei *miti* e di quelle *illusioni*» (p. 40) e pensare contro il "fanatismo" della fede nel Progresso. Per resistere alla seduzione di questa fede, tuttavia, «è necessario sollevare una 'scuola del sospetto'» (p. 42) (secondo il termine ricoeuriano), sostiene Oliveira.

Nel terzo capitolo, Oliveira (pp. 47-67) propone un approccio storico-filosofico al concetto di progresso per poi dimostrare i motivi per i quali esso viene oggi utilizzato come *moneta senza effigie*, cioè, come una moneta di scambio che mantiene solo un valore apparente senza, quindi, avere un fondo di valore reale, proprio perché promette sempre di più, ma si scontra con i limiti delle risorse finite della natura. All'inizio della sua analisi Olivera mette in luce il concetto aristotelico di progresso, inteso come "principio di movimento" come "essenza di tutte le cose naturali", che comprende un'idea di attuazione ciclica "nascita, crescita e decadenza" (p. 48). A sua volta, Sant'Agostino sostituisce l'idea di realizzazione finalistica (di tipo naturale) con quella escatologica, dove il concetto di Progresso è visto come la piena realizzazione di un disegno divino nel mondo. Questo cambiamento promosso dalla lettura agostiniana sostituisce l'idea di una realizzazione di cicli propri di ciò che è dotato di movimento, che comprende quindi la decadenza, all'idea di una realizzazione storica sempre orientata alla realizzazione del disegno divino (quindi del superamento e progresso). Questa lettura agostiniana fu reinterpretata

nell'Iluminismo, sostituendo l'idea di Dio con la Ragione. Questo movimento moderno, l'Iluminismo, inizia a vedere il Progresso come un orientamento verso il futuro, dove «gli esseri umani tendono a migliorarsi costantemente, trasformando il passato in una condizione per il proprio superamento» (p. 48). La Modernità è quindi una lotta della Ragione (delle scienze) contro ogni ostacolo al Progresso. Questa centralità dell'idea di Progresso, la fine del XIX secolo vide la nascita delle grandi "scuole del sospetto" con Marx, Nietzsche e Freud che presentarono una serie di critiche e considerarono il Progresso come un'illusione o un'utopia. Avanzando al XX secolo, Oliveira cita altri eventi storici come la Seconda Guerra Mondiale e la definitiva associazione dell'idea di Progresso alla crescita scientifica e tecnologica (per organizzare la barbarie e il caos lasciati dalla guerra). Questo percorso mostra che la critica dell'idea di Progresso era centrale e molto rilevante nel contesto intellettuale e culturale dal quale nasce *Il principio responsabilità*.

I punti successivi del terzo capitolo trattano del rapporto tra Progresso, economia e i problemi ambientali dell'attualità. Nel primo paragrafo, *I costi della crescita*, Oliveira ci presenta la tesi etico-economica di Acosta per un "buon vivere" come alternativa alla crescita economica illimitata basata sull'idea moderna di Progresso. Così come ne *I limiti della crescita* la proposta della "decrescita" viene sostenuta come un pensiero alternativo che ha già una certa portata teorica sviluppata fin dalla fine degli anni '70. Questa proposta non si traduce in un'imposizione del sottosviluppo globale, ma in un'organizzazione economica che sia guidata dai limiti imposti dalla natura stessa e dalle sue risorse esauribili (o che richiedono tempo per essere rinnovate). Per Oliveira, fondare un'organizzazione socioeconomica su questi limiti è una conquista tipica dell'uomo come *homo mensura*, cioè come quello che misura, che sa stabilire limiti. Nel paragrafo *Capitalismo e cambiamento climatico*, Oliveira ci ricorda insieme a Enrique Leff che «nessuna critica del Progresso può avvenire senza una critica del modello economico capitalista» e che criticare il capitalismo significa criticare il suo nucleo centrale: l'accumulazione che genera disuguaglianze. Con l'ultimo paragrafo, *Giustizia climatica*, Jelson Oliveira chiude il terzo capitolo tre includendo in maniera introduttiva una critica razziale del Progresso, identificato come «un progetto di uomini bianchi» le cui principali vittime sono «volti neri e indigeni». La giustizia climatica è un passo necessario per raggiungere la giustizia ambientale, poiché è più vicina alla critica del capitalismo che «non solo esaurisce le risorse naturali ma, soprattutto, genera vulnerabilità e sproporzionalità per affrontarle» (p. 67).

Nel quarto capitolo, Oliveira (pp. 69-85) esegue una lettura di Jonas basata su una classificazione proposta dai Larrère[2], i quali affermano che il pensiero etico-politica di Jonas può essere classificato come catastrofismo metodologico, in quanto opposto al catastrofismo retorico e alla collassologia. Il catastrofismo ontologico (o *collassologia*) è un tipo di narrativa sistematicamente deterministica, basata su un discorso di promessa che, proprio come nella narrativa del Progresso indefinito,

[2] Cfr. C. Larrère, R. Larrère, *Le pire n'est pas certain. Essai sur l'aveuglement catastrophiste*, Premier Parallèle, Paris 2020.

sarebbe destinato ad avvenire come un "processo naturale" lasciato a sé stesso senza la necessità dell'azione umana, incapace di guidarlo o di arrestarlo. Pertanto, la collassologia è un aspetto negativo dell'idea di progresso moderno che, carico della promessa apocalittica, sostituisce la promessa utopica e continua ad operare in modo deterministico. Alcuni studiosi accusano la collassologia di immobilità, poiché in questa prospettiva non ci sarebbe altro da fare per cambiare il destino necessario se non «adattarsi al continuo collasso» e cercare un ritorno all'interiorità, alle reti affettive, per vivere con «gioia, condivisione e fraternità» in mezzo al caos portato dal crollo. L'unica alternativa al "vivere bene" in tempi di collasso sarebbe quella di creare «spazi sicuri e/o controllati» dove il ritorno a questa interiorità e a questi valori sia possibile, come la «creazione di ecovillaggi, villaggi di transizione, economie alternative, agroecologia, comunità vegane, gruppi ecofemministi o queer, ecc.» (p. 73). Per Larrère e Larrère, bisognerebbe passare dai "collapsologi" ai "collassanauti". Anzi, i *catastrofisti retorici* e metodologici sono i difensori di un futuro aperto, cioè sono coloro che, indipendentemente dall'andamento negativo dei fatti, credono nella possibilità che potrà accadere qualcosa di sorprendente. Così, mentre per i catastrofisti retorici è sufficiente l'adempimento di una funzione pedagogica, per i *catastrofisti metodologici* è importante «'profetizzare' la catastrofe per convincere che bisogna fare di tutto per evitarla» (p. 76), cioè la funzione pedagogica deve essere al servizio di un cambiamento reale degli orizzonti politici.

Seguendo questa classificazione proposta dei Larrères, Oliveira elenca e aggiorna alcune proposte jonasiana presentate ne *Il principio responsabilità* che corroborano l'idea che bisogna profetizzare il peggio per insegnarci ciò che dovremmo temere, ma, soprattutto, per produrre una forma di esistenza che impedisca o crei ostacoli al male profetizzato (affinché non diventi reale). Secondo Oliveira, il catastrofismo si presenta nel pensiero jonasiano come un'euristica che mira a evitare l'immobilità e il fatalismo, lasciando aperta la possibilità che accada qualcosa di sorprendente. Nelle sue parole, «la catastrofe ha bisogno di essere tematizzata per poterla affrontare politicamente. La consapevolezza dei rischi deve quindi portare all'impegno per evitarli». (p. 78). Pertanto, di fronte ai dibattiti attuali come, ad esempio, quello relativo all'interrogativo se stiamo vivendo non solo una crisi ambientale, ma anche una nuova era geologica chiamata "Antropocene" (per alcuni teorici) o "Capitalocene" (per altri), non dovremmo paralizzare noi stessi come se fossimo di fronte a un male radicale dinanzi al quale non c'è più nulla da fare. Al contrario, comprendere l'Antropocene (e/ o Capitalocene) come l'era delle catastrofi significa, in primo luogo, comprenderlo come l'era dei limiti, afferma Oliveira (p. 80). Per definire quali limiti debbano essere imposti alla "utopia del Progresso tecnologico", che provoca una tale catastrofe, Hans Jonas propone una "euristica della paura" che deve essere supportata da una "scienza della futurologia" e guidata da un nuovo valore morale, vale a dire, la predizione.

Nel quinto capitolo, Oliveira (pp. 87-100) visita alcune riflessioni di Jonas dopo la pubblicazione de *Il principio responsabilità*, riunite in una raccolta di otto interviste e un discorso, tradotti dal tedesco e pubblicati in francese con il titolo *Une éthique pour*

la nature[3] nel 2000. Tali interviste presentano commenti, discorsi più specifici e perfino complementari a quanto sviluppato nell'*opus magnum* dell'autore. Oliveira utilizza alcuni di questi testi per evidenziare la centralità della critica al Progresso moderno e rafforzare l'inevitabile ricorso al sentimento di paura per una pratica della responsabilità jonasiana e, anche, l'urgenza di nuovi limiti compatibili con la libertà umana ampliata dalla tecnologia. Secondo Oliveira, in un'intervista rilasciata originariamente nel 1992 alla rivista tedesca *Der Spiegel*, Hans Jonas mostra un grande scoraggiamento di fronte all'aggravarsi della crisi ambientale e all'incapacità umana di cambiare l'ordine delle cose, ma riconosce anche che la conoscenza generale sul problema ambientale è aumentata. In questo senso, Jonas ci avverte dell'urgenza di attuare misure politiche planetarie in grado di porre un freno all'illusione del Progresso che storicamente ha sponsorizzato uno stile di vita non più compatibile con i limiti della natura, data la crescita demografica e il successo tecnologico. Jonas ritiene che la vera consapevolezza e, quindi, le decisioni politiche e le azioni concrete dovrebbero arrivare non attraverso la diffusione di informazioni o conoscenze su proiezioni catastrofiche, ma attraverso il sentimento prodotto da una "pedagogia della catastrofe". In questo modo si può dire che Jonas scommette sulla capacità mobilitante della "paura" che, trattandosi di una questione di psicologia, introduce elementi di volontà necessari per un'azione morale che richiede "sacrifici". Un esempio concreto di "pedagogia della catastrofe" è stato menzionato nel discorso *Tecnica, libertà e obbligo* quando Jonas, a partire dall'evento catastrofico reale di Chernobyl, afferma che «la morte delle foreste ha avuto sulla maggior parte di noi un impatto maggiore di quello di una qualsiasi previsione che, sebbene penetrante, fosse astratta» (Jonas *apud* Oliveira, *ibid.*, p. 89). Nonostante sia controversa, la "euristica della paura" è una delle proposte centrali in *Il principio responsabilità* che non deve essere confusa, per esempio, con una sorta di appello autoritario. Anzi, Jonas insiste sul fatto che si tratta di un appello *contro* il rischio di una "dittatura ambientale" in cui la libertà politica verrebbe limitata o addirittura liquidata in nome della sopravvivenza della specie. La mobilitazione del sentimento necessario all'azione morale richiesta dal principio responsabilità è una via possibile per evitare azioni estreme imposte dall'emergere della catastrofe ecologica. A tal fine, Oliveira afferma che Jonas propone un cammino di rinunce volontarie, per evitare un male maggiore, che egli definisce "l'era dei sacrifici". Questa via deve essere una proposta alternativa di critica e di limiti al Progresso moderno, alla crescita, al consumo e all'individualismo, poiché la vera libertà «non sarebbe nel *fare ciò che si vuole* o *tutto ciò che si può*, ma nella *scelta volontariamente di non fare ciò che si può e ciò che si vuole fare* quando ciò mette a rischio il bene più grande dell'esistenza dell'umanità» (p. 94). Ecco perché, in questo stesso passo, l'euforia del Progresso può essere contrapposta alla cautela richiesta dalla responsabilità come sostenuto nell'ultimo paragrafo del quinto capitolo. Oliveira afferma che l'euforia «deve cedere il passo prima alla paura e al timore e poi alla responsabilità stessa» (p. 98), dove l'imposizione dei limiti è il suo primo obbligo.

[3] Cfr. H. Jonas, *Une éthique pour la nature*, Desclèe de Brouwer, Paris 2000.

Nei capitoli sei e sette, Oliveira (pp. 101-118; 119-138) ci presenta un ampio dibattito su un aspetto particolare del pensiero jonasiano, cioè, relativo alla critica dell'idea di Progresso nel senso della storia umana: da quello insito nel programma baconiano di sviluppo tecnico/scientifico all'idea escatologica dell'ascensione del "vero uomo", liberato dal lavoro e pienamente sviluppato nelle sue facoltà morali, intellettuali e sensibile presenti nell'utopia marxista. Questo aspetto non è meno importante è, anzi, essenziale per caratterizzare i veri motori del principio responsabilità, libero da ogni ideale utopico. Per questo dibattito, Oliveira recupera la lettura, soprattutto dei capitoli 5 e 6 dell'*opus magnum* di Jonas, oltre a presentarci un testo dell'Archivio Hans Jonas di Konstanz, intitolato *Reflections on Technology, Progress and Utopia*. Secondo Oliveira, analizzando il rapporto tra tecnologia, progresso e utopia, è possibile svelare un concetto di "utopia del Progresso tecnologico" che rappresenta la caratteristica principale della società moderna e che si trasforma nel principio generale che dá senso alla storia umana (p. 103). Questa "utopia del Progresso tecnologico" è presente sia nelle concrete società liberali e/o capitaliste che nelle promesse utopistiche della teoria marxista; entrambe confondono una possibile evoluzione morale dell'uomo che può avvenire in una sfera privata e individuale con una conseguente evoluzione storica dell'umanità che dovrebbe avvenire in una sfera pubblica e collettiva. Perciò, i due capitoli trattano lo stesso problema, nel sesto capitolo esso viene affrontato a partire dalle società liberali e/o capitaliste concrete mentre nel settimo il problema è esaminato a partire dalla teoria marxista e dei suoi sviluppi.

Nell'ottavo capitolo, Oliveira (pp. 139-152) affronta l'idea di Progresso espresso anche sotto forma di miglioramento umano che può essere incluso nell'attuale dibattito sul transumanesimo e sul postumanesimo. Inizialmente, Oliveira si concentra sulla critica jonasiana dell'ontologia del "non-ancora" presente nell'utopia marxista di cui il principale rappresentante, direttamente criticato, è Ernst Bloch. Per Oliveira, Jonas considera l'utopia marxista erede del programma baconiano, espressione quindi della "utopia del Progresso tecnologico", tradotta nella sua massima potenza dalle promesse di far esistere il "vero uomo", un nuovo uomo mai esistito prima. La speranza di un uomo perfetto, in senso ultimo, sarebbe quella di un uomo senza "ambivalenza" e, quindi, un uomo privato della sua condizione di libertà (un *homunculus*). Secondo Jonas, la negazione dell'ambivalenza dell'uomo costituisce un errore ontologico nella filosofia di Ernst Bloch. Un altro aspetto di questa stessa speranza riguarda l'errore antropologico che, negando la formazione storica dell'essere umano, respinge la prospettiva di riformare o trasformare l'uomo basandosi sul suo passato e presente, appellandosi esclusivamente alla sua completa negazione. Ossia, il passato è riconosciuto come una somma di errori che devono essere negati, poiché il "vero uomo" nella proposta di Bloch è espresso nell'equazione "S non è ancora P", cioè "il soggetto non è ancora il suo predicato". Proseguendo il suo discorso, l'autore passa dall'aspetto etico a quello bioetico nella critica jonasiana del miglioramento umano. Oliveira ci presenta gli scritti che fanno parte dell'opera *Technik, Medizin und Ethik* de Jonas, che sono diventati un riferimento nei dibattiti

applicati sullo sviluppo tecnico negli esperimenti umani per controllare il comportamento, estendere la longevità e la manipolazione genetica. Per Oliveira è importante evidenziare che Hans Jonas non è un tecnofobo, cioè non è contrario allo sviluppo tecnologico, ma a favore di «politiche di prevenzione e preoccupazioni anticipatrici che mirano a 'conoscere in tempo opportuno ciò che stiamo per fare'» (p. 146). Questa è la prospettiva dell'ultimo paragrafo dell'ottavo capitolo, *Transumanesimo e Progresso*, dove Oliveira dimostra come, nel pensiero di Jonas, il Progresso nell'ambito della medicina, della biotecnologia e, soprattutto, dell'ingegneria genetica, deve passare attraverso un vaglio etico e politico come quello proposto dal principio responsabilità.

Nel nono capitolo, Oliveira (pp. 153-162) afferma che la grande sfida etica e politica del principio responsabilità nel nostro tempo è quella di «trasformare l'entusiasmo per l'utopia in un entusiasmo per l'austerità» (p. 154). Questo perché sia il capitalismo che il socialismo sono stati sequestrati dall'idea di Progresso e, secondo Jonas, nessuno dei due sarebbe in grado di gestirlo e di imporre limiti al suo incontrollabile sviluppo. Oliveira porta quindi al centro della questione del Progresso la difficoltà per qualsiasi sistema politico ed economico di imporre – di fronte agli attuali problemi ecologici – dei limiti, sostituendo il successo con la moderazione, l'eccesso con la frugalità, l'utopia con l'austerità. Il sequestro (per l'idea di Progresso) di entrambi i sistemi, capitalista e socialista, li rende incapaci di affrontare in modo opposto la "dinamica del successo" che si impone nella stessa misura in cui il successo biologico ed economico si alimentano a vicenda indefinitamente. Ossia, in altre parole, quanto maggiore è il successo economico nella produzione dei beni, nell'aumento del benessere sociale e nella riduzione dell'orario di lavoro, tanto maggiore è il successo biologico e il "corpo metabolizzante" – la crescita della popolazione attraverso il tasso di natalità e la longevità (e il sovrasfruttamento delle risorse naturale per mantenerlo). Oliveira ci conduce poi alla proposta politica jonasiana del "potere sul potere" che sarebbe capace di esigere una conoscenza di natura morale, in grado di controllare il Progresso. E, infine, nel paragrafo *Un falso sogno*, ancora una volta, Oliveira ritorna su quello che è il primo orientamento del principio responsabilità, cioè la critica dell'utopia come primo e fondamentale tassello per l'esercizio del principio etico nella civiltà tecnologica.

Nel decimo capitolo, Oliveira (pp. 163-175) amplia l'argomentazione di Jonas sulla necessità di offrire un'alternativa alla cultura dell'eccesso, con valori nuovi e più realistici come precauzione, frugalità e modestia. Per l'autore, «l'appello di Jonas è alla saggezza e alla moderazione nell'uso dei poteri e alla responsabilità globale che rappresentano» (p. 165). Ciò significa che è possibile ipotizzare un modello di progresso tecnologico regolato dai cauti precetti dell'etica della responsabilità. Agli eccessi dell'utopia bisogna quindi contrapporre "la modestia dei fini". A questo punto, Oliveira afferma che Jonas propone, con mezzi pacifici, misure come il controllo demografico della popolazione mondiale e, anche, uno stile di vita che non imiti quello della "minoranza globale dispendiosa", ma che, in cambio, si ispiri all'"appello a fini modesti". Pertanto, Oliveira sostiene che Jonas è vicino alle tesi *preservazioniste* «perché

ritiene che la prudenza e la responsabilità [...] portino a una critica dell'idea di umanizzazione o di ricostruzione della natura, comune nei discorsi utopici» (p. 166) come quelli trasmessi dalle idee di sostenibilità ambientale esplorate nella cultura consumistica del capitalismo, per esempio. Per Oliveira, la via indicata da Hans Jonas è quella della *frenata volontaria e della frugalità*, che deve svolgere anche un importante ruolo educativo di fronte alla *grande biforcazione* tra restare dove siamo oggi o seguire questa nuova direzione di armonia con la natura e rispetto della vita. Pertanto, la ricerca della frugalità proposta da Jonas «sostituisce l'espansione eccessiva con il rispetto dei limiti della natura» (p. 173), rielaborando l'economia classica in un nuovo modello che potremmo chiamare *bioeconomia*. In questo nuovo cammino possiamo imparare molto dalle "culture della gratitudine" che «intendono la natura come un dono da venerare e rispettare e non un mero oggetto da esplorare» (p. 175), conclude Oliveira indicando come chiave di lettura il concetto di "futuro ancestrale" sviluppato dal filosofo brasiliano Ailton Krenak[4].

[4] Cfr. AKrenak, *Futuro ancestral*, Cia das Letras, São Paulo 2022.

Eleonora Piromalli *L'alienazione sociale oggi. Una prospettiva teorico-critica*, Carocci, Roma 2023, 256 pp.[a]

Annaflavia Merluzzi[*]

In questo libro Eleonora Piromalli analizza un tema, oggi più che mai, fondamentale: quello dell'alienazione sociale. Il fine dell'autrice è di riportarlo al centro del dibattito filosofico, innanzitutto evidenziandone la natura eminentemente politica. Nel far ciò prende le mosse dalla delimitazione concettuale di questo termine, operazione necessaria in virtù della centralità e versatilità che esso ha assunto nel corso della filosofia moderna e contemporanea. Riportando, nelle prime pagine, le accezioni che il termine "*alienazione*" ha assunto nei densissimi secoli che ci separano dalla sua prima apparizione, l'autrice enuclea le problematiche relative alla trattazione recente di tale concetto – che hanno condotto a una sua graduale esclusione dal dibattito filosofico – quali: l'opacità e vaghezza dei confini teorici di esso, che portarono nel secolo scorso ad una sua applicazione eccessivamente estesa e poco rigorosa; l'associazione diretta dell'alienazione al marxismo, che ha destinato tale categoria a subire i contraccolpi del crollo del socialismo reale; la mancata risposta, da parte del marxismo teorico, alle critiche che erano state mosse nei confronti del concetto di alienazione (a questo proposito l'autrice ci ricorda che le principali critiche si concentravano sulle seguenti tre obiezioni: essenzialismo, paternalismo e riduzionismo).

Al fine di giungere a una definizione il più possibile univoca, ben determinata ma al contempo sufficientemente ampia della categoria di alienazione sociale, Piromalli propone di identificare quest'ultima come «un fenomeno pratico, socialmente causato, caratterizzato dal farsi estraneo di ciò che è proprio» (p. 33): la specificità di questa categoria sta nel fatto che i termini in essa coinvolti non divengono semplicemente separati l'uno dall'altro. Essi continuano, in molti modi, ad essere reciprocamente legati e ad appartenersi a vicenda, sebbene tale relazione si presenti come manchevole o deficitaria; questo costituisce per l'autrice il nucleo propriamente *filosofico* dell'idea di alienazione sociale. Questa definizione, nell'essere chiaramente delimitata rispetto a fenomeni di diversa natura ma che sono stati nei secoli assimilati all'alienazione stessa (come isolamento, esclusione, avversione, anomia, ecc.), permette però di considerare entrambi i piani tradizionalmente associati

[a] Recensione ricevuta in data 10/02/2024 e pubblicata in data 22/01/2025.
[*] Università degli Studi "La Sapienza" di Roma.

a tale categoria: ossia, l'alienazione sociale intesa come fenomeno sociale di larga scala e come fenomeno relativo all'interiorità del singolo.

Per dimostrare la necessità di considerare entrambi i livelli ("macro" e "micro"), Piromalli analizza l'alienazione sociale seguendo una struttura tripartita, come un "concetto a tre determinazioni" interne: queste sarebbero, nello specifico, l'alienazione sovraindividuale, l'alienazione soggettiva pratica, e l'alienazione soggettiva psicologica. L'alienazione sovraindividuale si verifica quando i membri di un dato gruppo sociale percepiscono le forme di organizzazione, da essi stessi generate e perpetuate, come esterne e intrasformabili, regolate da leggi che prescindono dal loro controllo. In secondo luogo, l'alienazione soggettiva pratica, ossia il lato soggettivo dell'alienazione sovraindividuale, viene inconsapevolmente generata a partire dalle prestazioni routinarie dei membri della società in questione. L'alienazione soggettiva psicologica, infine, si costituisce come una «scissione interna al soggetto, riguardante il suo rapporto con la propria interiorità, e, attraverso la mediazione di quest'ultima, con gli altri individui o con il mondo circostante» (p. 47).

Nei primi capitoli, dunque, l'autrice illustra le cause di queste tre determinazioni interne all'alienazione sociale, insieme con la loro composizione e correlazione. Negli ultimi anni si è discusso molto intorno all'intersezionalità, tema cui il concetto di alienazione soggettiva pratica, qui enucleato da Piromalli, può dare un notevole contributo (su questo punto è inoltre rilevante il paragrafo 7.2 del volume). Gli assi lungo i quali si può costituire il fenomeno dell'alienazione sociale, infatti, sono molti, come molti sono gli assi di oppressione socio-politica, e spesso è necessario agire in relazione alla loro totalità – senza mai assumerla come un tutto astrattamente omogeneo ma, piuttosto, come una molteplicità poliedrica ravvisabile tanto sul piano sociale quanto individuale. Come il soggetto si presenta in forma molteplice, così sono molteplici le affezioni che subisce e al contempo apporta all'organizzazione sociale cui appartiene.

Piromalli presenta poi una disamina di alcuni concetti spesso confusi con l'alienazione, distinguendoli da essa e fornendo gli strumenti teorici per evitare ambiguità. Da questo punto di vista, la confusione più frequente si dà tra alienazione e dominio sociale; concetti che, sebbene si rafforzino e perpetuino a vicenda, è importante mantenere distinti. L'alienazione sociale, infatti, non è di per sé un fenomeno di coercizione, come può essere il dominio, bensì consiste nella rappresentazione estraniata di forze create da un soggetto – o da una collettività – che si presentano a esso stesso come una datità intrascendibile ed esterna. Si potrebbe dire, tutt'al più, che si tratta di una percezione che porta a considerare come esterne e coercitive delle forze che in realtà il soggetto ha il potere di governare; creando, così, un'impasse circa la possibilità dell'azione politica, impasse che, naturalmente, si rivela ancella del mantenimento del potere sociale vigente. Su questo punto, il libro di Eleonora Piromalli ci fornisce degli strumenti fondamentali per demistificare le rappresentazioni su cui si regge, ad esempio, il dominio economico-politico del sistema capitalistico, aprendo lo spazio teorico per una decostruzione ed un ripensamento dell'assetto socio-economico-politico in cui viviamo. Nei capitoli 3 e 4,

l'autrice inquadra l'alienazione in ottica relazionale, dovendo a tal fine presentare una caratterizzazione del soggetto; in quest'ottica, e distanziandosi per alcuni versi da Rahel Jaeggi, delinea una stratificazione della costituzione soggettiva su tre livelli, guardandosi sempre dal rischio di sfociare in una prospettiva essenzialistica. Tuttavia, sarebbe forse stato opportuno dedicare una parentesi di natura teoretica più approfondita alla questione del soggetto, come realtà molteplice costituentesi in relazione all'altro, al contesto circostante e alla percezione di sé, che andasse oltre alla generale divisione per specie e per individuo dei livelli presentati da Piromalli. Va da sé, però, che questa sede non era probabilmente la più opportuna, vista la densità e complessità dei temi qui affrontati.

La seconda parte del testo è dedicata alle sfere in cui concretamente si può ravvisare oggi il fenomeno dell'alienazione sociale: economica, politica, ideologica. Attraverso una minuziosa scomposizione di tali sfere entro le circostanze in cui si determina l'alienazione sociale, l'autrice fornisce gli strumenti necessari ad immaginare soluzioni politiche che agiscano sulle cause di tali occorrenze, piuttosto che limitarsi a colpire gli epifenomeni. Possiamo notare che si contribuisce, così, a decostruire l'idea – funzionale essa stessa a mantenere in auge le politiche sociali occidentali messe in atto a partire dalla seconda metà del secolo scorso – che si debba agire solo sul fenomeno e non sui rapporti sociali, politici ed economici che lo causano (come avviene ad esempio nella lotta al fanatismo religioso, trattato quasi come fosse una psicosi collettiva, invece che il risultato della perpetuazione di uno stato di vulnerabilità e oppressione in cui versa una determinata frazione della popolazione mondiale). Questa impostazione teorica, che l'autrice presenta facendo attenzione a non cadere in atteggiamenti paternalistici, ci permette di andare oltre il singolo caso, e di orientare l'intervento politico al fine di prevenire il darsi dell'alienazione sociale, invece che dover gestire *in medias res* i risultati spesso tragici di essa.

La vera e propria *pars construens* del libro di Piromalli si incontra, poi, nell'ultimo capitolo, dove l'autrice formula i princìpi teorici che, nella sua prospettiva, possono fungere da veicoli di disalienazione al fine della costruzione di una società il più possibile priva di rapporti di dominio, oppressione, ed ogni fonte di alienazione sociale. Il primo e più importante di tali princìpi risiede nella libera comunicazione e interazione, da realizzarsi istituzionalmente integrando la democrazia rappresentativa con forme di democrazia partecipativa e deliberativa, quest'ultima individuata dall'autrice come modello in grado di mettere al centro «i soggetti come autori e creatori delle forme della propria società» (p. 217); tuttavia, Piromalli aggiunge che una compiuta ed efficace democrazia deliberativa può costituirsi solo attraverso un ulteriore momento: il riconoscimento intersoggettivo e infrasoggettivo, che acquisisce, in questa prospettiva, una valenza al contempo individuale, relazionale e politica, e che deve darsi innanzitutto entro i rapporti pratici della società per potersi tradurre in sede istituzionale. Una deliberazione che risulti equa, infatti, necessita che i partecipanti al processo godano di paritarie e adeguate condizioni materiali, occupazionali, di inclusione e uguaglianza giuridica; è attraverso le lotte per il riconoscimento che si rivendicano le condizioni di partenza di processi democratici

giusti. La conclusione cui l'autrice giunge, dunque, è una coimplicazione necessaria tra democrazia deliberativa e riconoscimento – di cui descrive la composizione interna e le condizioni di possibilità –, al fine tanto di innescare un processo di disalienazione quanto di realizzare una società giusta.

Che si concordi o meno con le soluzioni politiche proposte da Piromalli, *L'alienazione sociale oggi* è un testo che si propone di riportare al centro del dibattito filosofico un tema di prima importanza, fornendo a tal proposito degli utili strumenti di critica, teorica e sociale, tanto per un'analisi del fenomeno quanto per un approccio propositivo di risoluzione dello stesso.

Micheal Oliver, *Le politiche della disabilitazione. Il Modello Sociale della disabilità*, Ombre Corte, Verona 2023, 175 pp.[a]

Alessia Molisso[*]

Risale al 1990 il rigoroso saggio del sociologo britannico Micheal Oliver, *The politics of disablement*, per la prima volta tradotto in italiano nell'edizione di Ombre Corte, *Le politiche della disabilitazione*, curata da Enrico Valtellina, pubblicata e raccolta nella collana *Culture* nel 2023.

Il volume di Micheal Oliver potrebbe essere considerato una sorta di manifesto del Modello Sociale inglese della disabilità, ossia il primo modello interpretativo della condizione di persone con disabilità che sia stato essenzialmente politico, nella misura in cui si concentra non sulle diverse conformazioni fisiche delle persone disabili, evidenziando le varie possibilità di "funzionamento" corporeo e di azione, bensì sulle modalità in cui le società, le scuole, gli enti pubblici e privati, le organizzazioni abbiano risposto e rispondano alla presenza di queste persone nella rete dei rapporti sociali. In più, si tratta di un modello, nell'orizzonte dei *Disability Studies*, in cui sono in particolare le persone disabili stesse a operare la ricerca, al fine di renderla fattualmente emancipativa. Del resto, la cosiddetta *emancipatory research* ha condotto a risultati concreti nell'alimentare l'attivismo sui temi della disabilità e sulle rivendicazioni delle persone disabili.

Lo stesso Micheal Oliver, per un incidente in piscina durante l'adolescenza, era paraplegico. All'interno della medesima costellazione teorica e di attivismo politico è doveroso menzionare Paul Hunt, affetto da distrofia muscolare, e Vic Finkelstein, psicologo clinico sudafricano, esiliato dal regime dell'apartheid per la sua opposizione politica militante, utilizzava anch'egli la sedia a rotelle a causa di un incidente sportivo.

In verità, un manifesto del Modello Sociale inglese esisteva già, e in esso confluivano le voci appena ricordate: si tratta dei *Fundamental principles of disability* (1976), risultante da una discussione pubblica tenuta il 22 novembre 1975 e frutto dell'incontro tra l'UPIAS (Union of the Physically Impaired Against Segregation), associazione sindacale nata su proposta di Hunt, e la Disability Alliance. Il portato teorico principale del modello si esprime nella differenza concettuale tra i termini di

[a] Recensione ricevuta in data 08/08/2024 e pubblicata in data 22/01/2025.
[*] Dottoressa in Scienze filosofiche, email: molisso.a98@gmail.com.

menomazione e disabilità, *impairment* e *disability*. Se da un lato la menomazione è un dato naturale, biologico, in quanto dacché esistono società umane si sono ripetutamente presentati fenomeni di deformazioni fisiche, sensoriali, relazionali, cognitive ecc., dall'altro lato la disabilità è un fenomeno sociale, connotato dal valore inferiorizzante che la società ha attribuito tramite la segregazione agli individui con *impairment*. Così, il processo di disabilitazione è da intendere come una determinata forma di oppressione che Oliver non tarda a definire "multidimensionale". Egli non può infatti trascurare l'importanza di adottare uno sguardo intersezionale e di constatare, oltre al peso del dato reale dell'*impairment* su alcuni approcci teorici prevaricanti, che la maggior parte delle persone disabili è attraversata da ulteriori oppressioni, di natura ideologica (di genere e razziale).

Tra le forze sociali il principale agente produttore della disabilità come problema (del singolo individuo) è il modello medico, che non esita a fare della condizione di menomazione una tragedia personale. Da quest'ultimo modello, attualmente egemone nella società capitalistica, discende l'idea secondo cui sia l'individuo disabile a doversi adattare a un mondo ritagliato pedissequamente secondo le esigenze dei normodotati. È evidente che, nel ribaltare totalmente i termini della questione e puntando al contrario a un'adeguazione delle società alle attese criticamente riconosciute e affermate dalle persone disabili, il Modello Sociale affondi le radici dei propri argomenti nel sostrato del materialismo, segnatamente gramsciano e althusseriano.

Nel volume di Oliver, accuratamente scandito in tutte le sue parti, è possibile rinvenire vari nodi concettuali attorno a cui si agglomerano le sue riflessioni: medicalizzazione, normalità, lavoro, dipendenza. Ma il filone ermeneutico, funzionale a dirimere la concezione fondativa da cui si dipartono le tesi dell'autore, è l'oscillazione fra costruttivismo sociale e creazionismo sociale, risolta a favore del secondo. Sebbene il costruttivismo sociale scalzi dall'ideologia individualista il problema della discriminazione, esso rimane ancorato all'idea secondo cui l'abilismo si situi nelle *menti* dei singoli individui e si esprima nei loro *atteggiamenti* interpersonali. Si tratta dello stesso motivo per cui, ad avviso di Oliver e Finkelstein, la nozione di stigma, elaborata da Goffman, risulta esplicativa ma insufficiente in chiave anti-individualistica. Al contrario, il creazionismo sociale colloca l'emarginazione entro il quadro delle *pratiche istituzionali* che intessono attivamente le relazioni sociali, "producendo" la trama sociale così come le sue nozioni egemoni. Secondo quest'ultimo punto di vista, sarebbe infatti la società medesima a creare la categoria di disabilità.

A questo proposito, sulla scorta del suo orizzonte di studi, uno dei riferimenti principali dell'argomentazione di Oliver è dichiaratamente Gramsci, per il quale le idee, lungi dall'avere una natura accidentale o dall'applicarsi contingentemente alle menti dei singoli, rendendo solo *a posteriori* il conto della natura della società come mera sommatoria; sono *forze materiali* capaci di creare l'ideologia. Dunque, la prospettiva creazionista riesce a oltrepassare i limiti di una critica costruttivista e a giungere alla nozione di discriminazione istituzionalizzata, in grado di evidenziare

l'assoluta negligenza delle organizzazioni di potere e delle istituzioni nello svolgere il ruolo di attori sociali attivi nella lotta alle disuguaglianze sostanziali e strutturali.

Per quanto concerne l'idea di dipendenza, invece, Oliver sostiene che, in omogeneità disgiuntiva con il suo negativo (l'indipendenza), è il prodotto delle pratiche istituzionalizzate di una società abilista come quella contemporanea capitalistico-industriale. Si tratta di una dicotomia retoricamente affermata dalla pratica politica per ammantare l'interdipendenza reciproca che realmente intride la nostra società, a vantaggio della promozione di una certa tendenza competitiva che investe tutta la popolazione, indistintamente.

Attorno al concetto di dipendenza non può che articolarsi il tema del lavoro, fondamentale nelle società industriali. Come ha mostrato Foucault, altro riferimento esplicito di Oliver, dalla cosiddetta "età classica" (nella partizione foucaultiana: l'epoca che si dispiega dopo il Medioevo fino agli anni della Rivoluzione francese) pullularono istituti a funzione internante, tra cui le *Zuchthäuser* tedesche o le *houses of correction* e *workhouses* inglesi, ove l'obbligo del lavoro divenne uno strumento di controllo morale, di normazione e sanzione, che profilava il configurarsi del vizio capitale dell'etica borghese: l'inoperosità. Tale atmosfera non tardò a influenzare negativamente le vite delle persone disabili, non facilmente immettibili nei nuovi circuiti lavorativi, soprattutto con l'industrializzazione tra il XVIII e il XIX secolo.

Oliver pone l'attenzione sul fatto che anche attualmente sussiste per le persone disabili il medesimo rischio di esclusione «dalla forza lavoro a causa della percezione a priori di una loro incapacità, e quindi si riproduce la dipendenza» (p. 114). Questa percezione non è altro che tale, quindi manifestamente storica, cioè determinata da uno specifico assetto economico, sociale e politico; non da una presunta naturalità della condizione di disabilità, ossia della menomazione, come vorrebbe un certo "acritico riduzionismo sociologico". Anche le politiche che incentivano l'inclusione dissimulano un'esclusione, dal momento che sono orientate solamente all'offerta di lavoro, cioè a rendere più appetibili gli aspiranti lavoratori disabili per i datori di lavoro, puntando a favorire un adattamento dei disabili stessi alle richieste del mercato lavorativo (senza mettere in discussione quest'ultimo). Corroborando così il suo approccio creazionista, Oliver pone l'accento sull'urgenza per cui nuove politiche del lavoro, orientate a «creare ambienti di lavoro senza barriere» (*ibidem*), dovrebbero partire dal governo e non dagli agenti produttori, tendenzialmente non sostenuti a «progettare macchinari o strumenti che siano utilizzabili da tutti, indipendentemente dalle loro abilità funzionali» (*ibidem*).

Oliver, nel volgersi della sua puntuale argomentazione, pur riconoscendo il carattere globale del fenomeno disabilità/disabilitazione nella società contemporanea e capitalistica, esamina in maniera più attenta e circoscritta il caso della Gran Bretagna. Per quanto invece concerne l'Italia è emblematico avvedersi della medesima introiezione, a livello istituzionale, della soglia percettiva rispetto alle persone disabili, soprattutto considerando che la discriminante tradizionale in merito è il tema del lavoro (basti riflettere sugli articoli 3 e 38 della nostra Costituzione, da cui emergono le problematiche equazioni concettuali tra persona e lavoratore, e inabile al lavoro e

disabile). L'idea stessa del *welfare*, come deposito di risorse in caso di incapacità di sostentamento della famiglia, si può dire derivi storicamente dal divieto di mendicare sancito da Luigi XIV e dal conseguente internamento dei soldati tornati invalidi, a seguito di campagne militari, in strutture come l'Hôtel des Invalides. Da questa breve ricostruzione è possibile evincere la genesi storica del rapporto sociale tutt'oggi presente: una struttura verticale formata da un'istanza detentrice del potere, presuntamente benevola, dispensatrice di sussidi, e un polo negativo, rappresentato dagli invalidi, per la loro condizione aiutati economicamente, eppure in qualche modo recisi fuori dal resto della comunità (dei normodotati e dei lavoratori). In ogni caso, l'analisi di Oliver ci spinge a precisare che vi sono state delle variazioni, dal momento che attualmente anziché l'"umanitarismo benevolo", l'ideologia di base del *welfare* «riflette piuttosto il peso che si presume siano le persone disabili non produttive e l'influenza del realismo monetarista» (p. 109).

Tornando quindi al saggio, Oliver sottolinea che le precedenti considerazioni economiche non sono esaustive, nella misura in cui, oltre a non tener conto del dato che «la maggioranza delle persone disabili in età lavorativa ha un lavoro, e quindi è economicamente produttiva» (p.115), trascurano il ruolo del consumo, che è invece dominante nell'economia tardocapitalistica.

Così, l'etichetta di dipendenza, affibbiata in chiave economicistica alle persone disabili, deriva in primo luogo dal significato sociale, dall'ideologia creata dall'orizzonte politico, preoccupato di fornire assistenza a esse mediante svariati servizi (strutture residenziali e diurne, trasporto con mezzi specializzati, protesi, equipaggiamenti automobilistici ecc.). Tali servizi assistenziali, secondo Oliver, realizzano l'istituzionalizzazione delle persone disabili, rendendole dipendenti dal potere decisionale di una serie di professionisti amministratori delle risorse (esigue) destinate alla loro riabilitazione. Ciò dà luogo a una struttura gerarchica professionista-cliente (o "utente", "consumatore", la sostanza non cambia), che riecheggia il rapporto di potere medico-paziente descritto dal Foucault di *Storia della follia* in merito al manicomio, in cui era riscontrabile una distribuzione politica piramidale, discendente dalla posizione apicale del medico fino a quelle di aiutanti, infermieri, sorveglianti guardiani o inservienti.

Oliver, ad ogni modo, rimarca che, per il loro lavoro stipendiato, in verità «sono i professionisti che dipendono dalle persone disabili» (p. 118). Le relazioni fondate sulla dipendenza, che irretiscono in un medesimo vincolo sia il personale riabilitativo sia le persone disabili, possono essere scompaginate solo attraverso una risemantizzazione del concetto di indipendenza già proposto come soluzione dalle due parti in campo. Se i professionisti intendono tale nozione nell'accezione dell'autocura riferita esclusivamente a bisogni fisico-pratici, le persone disabili sostengono una visione più ampia, più prossima al concetto di autodeterminazione, legata dunque alla realizzazione di processi decisionali nella propria vita, su un piano psicologico e sociale insieme. L'indipendenza dovrebbe avere a che fare con la qualità della vita, così come asserisce il sociologo disabile Irving Zola, la cui citazione seguente, riportata da Oliver, esprime paradigmaticamente un compito inaggirabile:

«il personale della riabilitazione deve cambiare il modello di servizio dal fare qualcosa *a* qualcuno al pianificare e creare servizi *con* qualcuno» (p. 119, corsivo mio).

A confluire nell'idea sociale di dipendenza vi sono i pregiudizi di immaturità e isolamento diffusi dalla percezione dei normodotati, sui quali si appoggiano anche le organizzazioni di volontariato tradizionali, volte ad alimentare una certa sensibilità caritatevole che rievoca il cosiddetto modello caritativo-religioso di retaggio cristiano, al quale, così come al modello medico, i *Disability studies* hanno reagito polemicamente. Lo scopo di tali organizzazioni è «massimizzare le entrate, indipendentemente dall'immagine presentata» (p. 120), istituendo, per ricalcare un'espressione di Paul K. Longmore, un vero e proprio *business* della carità, nell'ottica dell'accumulazione capitalistica, la stessa per cui è redditizio che «le persone disabili possono svolgere una funzione economica come parte della riserva lavorativa e una funzione ideologica nell'essere mantenute nella loro posizione di inferiorità» (p. 98).

Ad ogni modo Oliver mostra che «le persone disabili non sono trattate come inferiori in tutte le società o in tutti i momenti storici» (p. 98), sottolineando l'importanza di non tendere a una lettura naturalizzante di qualcosa che è storico, come le costruzioni ideologiche della disabilità alimentate dall'individualizzazione e dalla medicalizzazione. Alla stessa maniera, non bisogna obliare che il concetto di normalità è culturale, pertanto transeunte, sebbene definizioni ufficiali come quella dell'OMS, rimarca l'autore, giungano a una reificazione dell'idea di normalità.

Oliver si preoccupa di smascherare il carattere ideologico della presunta normalità del corpo abile, la stessa che alimenta il concetto della riabilitazione e che è affermata dall'individualismo, dilagante anche in ambito pedagogico oltre che medico, come è evidente dalla popolarità dell'"educazione conduttiva". La stessa terminologia negativa ("*dis-abile*") tradisce il pregiudizio delle società contemporanee, contenente un'ingiunzione al fare, all'essere abili a svolgere determinate attività (produttive) e in un certo modo (modo di produzione capitalistico). Come ci autorizza a dire Oliver, si tratta di una questione di "pratiche discorsive", di foucaultiana memoria, le quali, impregnate delle ideologie dominanti, producono l'esperienza individuale della disabilità. L'oggetto è dunque il tema dell'identità disabile, che «non si forma semplicemente attraverso processi psicologici interni, ma può essere imposta dall'esterno» (p. 105). L'intreccio tra le forze esterne dominanti è ben visibile se si considera che il discorso medicalizzante si è consolidato non solo per il successo della medicina basata sugli ospedali, ma anche e soprattutto poiché essa «è nata dal bisogno di classificare e controllare la popolazione e di distinguere tra lavoratori e non lavoratori all'interno del nuovo ordine sociale capitalista» (p. 82). Benché i corpi siano attualmente plasmati dall'ideologia individualista, e sebbene la pratica medica abbia favorito le aspettative di vita delle persone disabili, queste ultime rivendicano la necessità di una forma di vita che sia di effettiva qualità, un vivere bene (al di là degli aspetti clinico-terapeutici), Finkelstein parla così della disabilità nei termini di "modo di vivere". A dispetto del controllo biopolitico sui corpi in funzione della loro utilizzabilità e utilità, della produzione e riproduzione sociale di essi, della divisione del lavoro fondativa del sistema capitalistico, il Modello Sociale della disabilità offre

gli strumenti epistemologici e politici per la lotta volta alla riacquisizione del controllo sulle vite delle persone disabili, alla riappropriazione del proprio spazio di autodeterminazione.

Un *empowerment* autentico per le persone disabili, ossia sociale, non può trascurare di conoscere il «processo storico che ha portato alla formazione delle immagini culturali delle persone disabili» (p. 104) e di affermare in replica ad esso un nuovo "processo di formazione dell'identità". Quest'espressione del sociologo sembra suggerire la possibilità di un nuovo processo di soggettivazione, che tenga conto di altri fattori strutturanti come la razza e il genere, che sia consapevole del proprio portato storico in quanto processo e al contempo prescinda da una corporeità presuntamente normale. In questa direzione pensiamo possa essere utile integrare la visione del "tardo" Foucault sulla cura di sé intesa come auto-costituzione del sé, come pratica di libertà e non di assoggettamento.

Uno scoglio da arginare è rappresentato dal dato che «le lotte all'interno del terreno ideologico generato dall'oppressione non avvengono solo tra gli oppressori e gli oppressi, ma anche tra gli oppressi stessi» (*ibidem*). Pertanto, a giocare un ruolo decisivo nello sviluppo di consapevolezza della propria socializzazione verso una identità critica è la nascita dei nuovi movimenti sociali, come appunto il movimento dei disabili, in cui si rileva la dimensione portante dell'Independent Living Movement. Il Modello Sociale afferma così l'attivismo disabile contro la posizione di *expertise*, cioè del (presunto)sapere-potere, di chi si pone a capo di organizzazioni *per* disabili e non *di* disabili. In sostanza, per Oliver, il movimento dei disabili deve "allearsi" con lo Stato per assicurarsi le risorse adeguate ai propri bisogni e al contempo rimanere autonomo rispetto ad esso per non appiattire il proprio modello dinamico alla visione monetarista e paternalistica dell'*establishment*. Sebbene le rivendicazioni del movimento, per le condizioni materiali attuali, siano inefficaci a rovesciare lo *status quo*, «è il loro potenziale contro-egemonico, non le loro realizzazioni effettive, ad essere significativo nel tardo capitalismo» (p. 156).

In conclusione, la recente pubblicazione italiana dell'opera di Oliver colma un vuoto trentennale, arricchendo ampiamente il panorama di studi e politico del nostro Paese. Essa, invero, come complessivamente i contributi del Modello Sociale inglese, è in grado di aiutare a far pensare nuove forme di soggettivazione, individuale e collettiva, in vista della resistenza all'oppressione disabile, affinché un vento di cambiamento soffi anche su questa terra.