TEORIA
T
Rivista di filosofia
fondata da Vittorio Sainati
XLII/2022/2 (Terza serie XVII/2)

New challenges in decision-making processes

Nuove sfide nei processi di decisione

Edizioni ETS

TEORIA

Rivista di filosofia
fondata da Vittorio Sainati
XLII/2022/2 (Terza serie XVII/2)

Iscritto al Reg. della stampa presso la Canc. del Trib. di Pisa n° 10/81 del 23.5.1981

Direzione e Redazione
Dipartimento di civiltà e forme del sapere dell'Università di Pisa, via P. Paoli 15, 56126 Pisa,
tel. (050) 2215400 - www.cfs.unipi.it

Direttore Responsabile
Adriano Fabris

Periodico semestrale

Abbonamento (cartaceo, privato): Italia e UE € 40,00; extra UE € 50,00
Abbonamento (cartaceo, istituzionale): Italia e UE € 50,00; extra UE € 60,00
PDF (online, stampabile): privato (accesso individuale) € 40,00
PDF (online, stampabile): istituzionale (accesso con riconoscimento IP) € 40,00

Bonifico bancario intestato a
Edizioni ETS - Banca C.R. Firenze, Sede centrale, Corso Italia 2, Pisa
IBAN IT 21 U 03069 14010 100000001781
BIC/SWIFT BCITITMM
causale: abbonamento «Teoria»

«Teoria» è indicizzata ISI Arts&Humanities Citation Index e SCOPUS, e ha ottenuto la classificazione "A" ANVUR per i settori 11/C1-C2-C3-C4-C5. La versione elettronica di questo numero è disponibile sul sito: www.rivistateoria.eu

L'indice dei fascicoli di «Teoria» può essere consultato all'indirizzo: www.rivistateoria.eu
Qui è possibile acquistare un singolo articolo o l'intero numero in formato PDF, e anche l'intero numero in versione cartacea.

I numeri della rivista sono monografici. Gli scritti proposti per la pubblicazione sono double blind peer reviewed. I testi devono essere conformi alle norme editoriali indicate nel sito.

Palazzo Roncioni - Lungarno Mediceo, 16, I-56127 Pisa
info@edizioniets.com - www.edizioniets.com

Distribuzione
Messaggerie Libri SPA - Sede legale: via G. Verdi 8 - 20090 Assago (MI)

Promozione
PDE PROMOZIONE SRL - via Zago 2/2 - 40128 Bologna

ISBN 978-884676525-3
ISSN 1122-1259

Contents / Indice

Veronica Neri
Premessa / Premise, p. 5

Antonio Da Re
Il giudizio clinico integrato e la responsabilità del medico
al tempo di Covid-19, p. 17

Silvia Dadà
Autonomia in telemedicina. L'*e-patient*
tra indipendenza e coinvolgimento, p. 33

Francesca Marin
Decisioni di fine vita, dipendenza e vulnerabilità, p. 53

Federico Zilio
Tough Decisions in Unclear Situations.
Dealing with Epistemic and Ethical Uncertainty
in Disorders of Consciousness, p. 67

Mario De Caro, Massimo Marraffa
Consciousness and responsibility, p. 81

Marco Menon
L'esternalizzazione dei processi di decisione
nella società postindustriale.
Vilém Flusser e il funzionario nell'apparato, p. 97

Benedetta Giovanola, Simona Tiribelli
Equità e decisioni algoritmiche, p. 117

Sofia Bonicalzi
A matter of justice. The opacity of algorithmic decision-making
and the trade-off between uniformity and discretion
in legal applications of artificial intelligence, p. 131

Veronica Neri
Intelligenza artificiale e scelte di consumo:
l'immaginazione come antidoto
ai processi di *behavioural bias*, p. 149

Francesca Pongiglione
Trust, experts, and the potential side effects
of critical thinking, p. 163

Sarah Songhorian, Francesca Guma,
Federico Bina, Massimo Reichlin
Moral Progress: *Just* a Matter of Behavior?, p. 175

Francesca Guma
Becoming Better Moral Agents by Strengthening Free Will.
A Possible Prospect?, p. 187

Federico Bina
Models of moral decision-making:
Recent advances and normative relevance, p. 201

T

Veronica Neri

Premessa / Premise

Il presente fascicolo di «Teoria» trae origine dal convegno «Nuove sfide nei processi di decisione. Bioetica, Neuroetica, Etica dell'Intelligenza Artificiale», svoltosi presso l'Università di Pisa il 7 e 8 aprile 2022 e promosso dal programma PRIN 2019/2022 su «Etica & tecnologia. Nuove sfide per l'etica applicata», progetto che ha visto come Coordinatore scientifico nazionale il prof. Mario De Caro (Università Roma Tre) e come Responsabile scientifico dell'unità di ricerca pisana il prof. Adriano Fabris. Il convegno si è incentrato su una riflessione relativamente alle nuove sfide nei processi di decisione aperti dall'uso sempre più massivo delle tecnologie via rete nella società. In particolare si sono indagati alcuni aspetti propri della bioetica, della neuroetica e dell'etica dell'intelligenza artificiale. Il tema della decisione ha del resto sempre caratterizzato il dibattito eticofilosofico richiedendo l'esame di alcuni concetti cardine della filosofia morale quali la responsabilità nelle sue molteplici declinazioni, l'autonomia del soggetto e il concetto di libero arbitrio. Si affiancano a queste fondamentali problematiche nuove sfide legate agli sviluppi contemporanei in ambito scientifico e tecnologico.

Con questo fascicolo si intende dunque aprire alcune considerazioni sull'impatto di tali sviluppi nei processi di decisione, ponendo attenzione in particolare a cinque nodi concettuali: le nuove forme di vulnerabilità derivanti dall'utilizzo delle tecnologie nelle pratiche di cura, la responsabilità del medico e la formulazione del giudizio clinico nell'epoca dell'intelligenza artificiale, la ridefinizione dei processi decisionali, in particolare dell'agire morale, della libertà e dell'autonomia umane alla luce degli sviluppi neuroscientifici, le conseguenze dell'intelligenza artificiale, degli algoritmi e dei modelli statistico-predittivi a livello individuale e sociale e, infine, le tecno-

 DOI: 10.4454/teoria.v42i2.159

logie intese come "agenti artificiali morali" coinvolti nell'ambito dei processi di decisione. Si tratta di uno sguardo trasversale che parte dalla filosofia morale, ma che intende promuovere un confronto tra diversi ambiti – come quello medico, scientifico-tecnologico e meramente filosofico –, che sempre più, anche in ragione della recente pandemia, appaiono interrelati.

Le tematiche premesse sono state affrontate in tre sessioni, la prima delle quali incentrata sulla bioetica. In particolare il saggio di apertura del convegno, di Antonio Da Re, dal titolo «Il giudizio clinico integrato e la responsabilità del medico al tempo di Covid-19» apre ad alcune questioni sollevate dal *triage* ospedaliero, emerse durante la pandemia da Covid-19, che riguardano da una parte la «rivincita del principio di giustizia», dall'altra la «limitazione del principio di autonomia, anche in nome del riconoscimento dell'interdipendenza di tutti gli esseri viventi». Il *triage* assume connotati diversi e, come sottolinea l'autore, diventa 'estremo' nella misura in cui non si tratta più di stabilire l'ordine di priorità temporale nella cura, «ma chi può essere curato e chi invece no», sulla base dei criteri di urgenza clinica e della possibilità di sopravvivenza del paziente. L'autore, facendo un accurato confronto anche con documenti e direttive, ritiene dunque fondamentale un ripensamento sul concetto di giudizio clinico attraverso l'analisi di aspetti più specificatamente clinici con quelli maggiormente legati al principio di giustizia.

Il saggio subito successivo, di Silvia Dadà, «Autonomia in telemedicina. L'*e-patient* tra indipendenza e coinvolgimento» analizza il ruolo dell'autonomia del paziente (*e-patient*) ai tempi delle ICT in campo medico, una autonomia che sembra aumentare grazie alla telemedicina. Dopo una prima ricostruzione del dibattito sull'argomento tra chi pensa che la telemedicina sia un incentivo all'autonomia del paziente e chi la interpreta come uno strumento di controllo e dominio sul corpo dei pazienti si esplora il concetto di autonomia in senso esecutivo, valorizzato – con i limiti che ne possono emergere – dalla telemedicina. L'autrice propone infatti una prospettiva più relazionale, che possa favorire il coinvolgimento tra medico e paziente, la tutela delle dipendenze e delle vulnerabilità presenti.

Francesca Marin analizza, invece, il dibattito italiano sul suicidio assistito (PAS). Con il saggio «Decisioni di fine vita, dipendenza e vulnerabilità» mostra in che modo il PAS sia caratterizzato da un misconoscimento delle differenze mediche e bioetiche tra i vari trattamenti e procedure, con il conseguente rischio di adottare un approccio riduttivo a ciò che afferisce alle c.d. "decisioni di fine vita". Dopo una prima parte incentrata sul les-

sico relativo al fine vita – che equipara pratiche diverse in termini di procedure e obiettivi –, la seconda parte del contributo sostiene che equazioni improprie possono essere rilevate anche analizzando sia le argomentazioni della Corte Costituzionale sia il recente tentativo di estendere l'espressione «trattamenti di sostegno vitale».

Questa prima sessione si conclude con l'intervento di Federico Zilio intitolato «Scelte difficili in situazioni poco chiare. Il processo decisionale di fine vita nel contesto dei disturbi di coscienza». Dopo una prima definizione dei disturbi della coscienza (DoC), caratterizzati da una perdita compromessa o completa della consapevolezza di sé e dell'ambiente, si presentano alcune questioni epistemiche e metodologiche che caratterizzano i disturbi della coscienza stessa: l'errore diagnostico, l'incertezza prognostica, la comunicazione con la famiglia e gli operatori sanitari e il valore performativo del linguaggio clinico. L'incertezza epistemica che ne emerge è, secondo l'autore, profondamente intrecciata con l'incertezza etica, in specie nel caso di decisioni cliniche che possono portare alla morte di persone in uno stato di coscienza non chiaro. L'autore suggerisce, in questi casi, come possibile via la necessità della prudenza epistemica ed etica, formulando un equilibrio tra i due principi del rischio induttivo, per scongiurare decisioni, interpretazioni e comunicazioni medico-familiare errate.

La seconda sessione, dedicata alla neuroetica, è aperta dal contributo di Mario De Caro con Simone Marraffa dal titolo «Coscienza e responsabilità». Dopo aver esposto la tensione tra le neuroscienze cognitive e alcune visioni etiche basate sulla visione ordinaria del mondo, gli autori sostengono l'adozione di una posizione intermedia tra quella sostenuta dagli eticisti tradizionali (che attribuiscono un primato assoluto al pensiero cosciente nell'agire morale) e quella sostenuta da neuroscienziati e filosofi cognitivi (secondo i quali la mente cosciente è davvero epifenomenale). Il modello alternativo proposto, sulla scia di Levy, Carruthers e King, si fonda sul fatto che le scoperte della neuroscienza cognitiva, piuttosto che mostrare che la mente conscia sia epifenomenale, richiedono un'articolazione più fine e imparziale della dialettica tra elaborazione inconscia e riflessione conscia.

Segue il contributo di Marco Menon «L'esternalizzazione dei processi di decisione nella società post-industriale. Vilém Flusser e il funzionario nell'apparato». L'autore osserva la presenza sempre più invasiva e trasformativa dell'intelligenza artificiale e degli algoritmi nei processi decisionali alla luce del concetto di apparato flusseriano. Se, da un lato, si offre un contributo alla ricostruzione del concetto di apparato flusseriano, d'altra

parte, si propone di utilizzare le categorie di Flusser per interpretare alcuni fenomeni emergenti della società postindustriale come forme di esternalizzazione dei processi decisionali. Quest'ultimo aspetto comporta una crescente deresponsabilizzazione degli agenti morali, e, secondo l'autore, si tratta di un fenomeno che può essere inquadrato ricorrendo alla nozione di "funzionario". Eppure, secondo Flusser, un processo decisionale libero ed esistenzialmente significativo è possibile anche nel mondo degli apparati, in ragione della capacità di astrazione umana.

Il saggio successivo, di Benedetta Giovanola e Simona Tiribelli, «Equità e decisioni algoritmiche» si concentra su uno dei rischi contemporanei più urgenti dell'intelligenza artificiale, e più specificamente, del processo decisionale algoritmico (ADM), ovvero il rischio di essere ingiusti. Nella prima sezione le autrici forniscono una panoramica del concetto di equità in ADM e ne mostrano le carenze; nella seconda parte perseguono un'indagine etica sul concetto di equità e ne individuano le dimensioni e le componenti principali, traendo spunto da una rinnovata riflessione sul rispetto (anche per particolari individui). Nella terza ed ultima parte le autrici mostrano come la nostra rielaborazione concettuale dell'equità può aiutare a identificare i criteri che dovrebbero guidare la progettazione etica dei sistemi basati su ADM per renderli realmente equi.

La sessione si chiude con il contributo di Sofia Bonicalzi «Una questione di giustizia. L'opacità del processo decisionale algoritmico e il compromesso tra uniformità e discrezionalità nelle applicazioni giuridiche dell'intelligenza artificiale». L'autrice osserva come negli ultimi anni, le decisioni in materia di giustizia distributiva e retributiva sono state sempre più esternalizzate a sistemi automatizzati (AI) e sono progressivamente emerse nuove sfide etiche. Rispetto agli arbitri umani, i sistemi basati sull'IA presentano vantaggi concreti in termini di efficienza e uniformità delle prestazioni. Tuttavia, la ricerca dell'uniformità può avere anche costi considerevoli. Questo contributo mira a concentrarsi su una sfida specifica – il difficile compromesso tra uniformità e discrezione nelle applicazioni giudiziarie dell'intelligenza artificiale – sullo sfondo degli attuali dibattiti in filosofia, scienze cognitive e intelligenza artificiale. Eludere le peculiarità del ragionamento umano potrebbe avere alcuni effetti dannosi sull'equità dell'amministrazione della giustizia.

La terza ed ultima sessione del convegno è dedicata, infine, all'etica dell'intelligenza artificiale. Il saggio di apertura della sessione, di Veronica Neri, dal titolo «Intelligenza artificiale e scelte di consumo: l'immagina-

zione come antidoto ai sistemi di *behavioural bias*» ha l'obiettivo di indagare quali decisioni deleghiamo (in)consapevolmente all'IA in un contesto di scelte di consumo e di acquisto (di informazioni, immaginari, beni e servizi) e quali margini di autonomia può avere l'individuo cercando di preservare la propria capacità valutativa. Se ormai l'idea di neutralità dell'IA appare superata, è necessario esplorare se e come l'IA possa chiudere in *clustering* o, al contrario, consentire un maggiore coinvolgimento morale. Il documento analizza le strategie di micro-targeting utilizzate dall'IA per incoraggiare consumi e la profilazione dei nostri comportamenti online, prestando particolare attenzione alla pubblicità comportamentale e ai sistemi di *bias* comportamentale. In conclusione, l'autrice riflette sulle modalità tramite le quali l'individuo può arginare tale potere di orientamento degli algoritmi, in particolare attraverso la capacità di immaginazione, propria dell'essere umano, che può controbilanciare i meccanismi di induzione al consumo razionalmente pianificati attraverso l'IA.

Segue il saggio di Francesca Pongiglione sulla relazione tra «Fiducia, esperti e la potenziale deriva del critical thinking». I nostri doveri epistemici come cittadini del mondo globale ci impongono di cercare informazioni per garantire che le nostre azioni non danneggino gli altri o noi stessi. Quando integriamo queste informazioni, non dovremmo accettare passivamente ciò che ci viene detto, senza garantire che le fonti su cui facciamo affidamento siano affidabili. Questo evitare una fiducia eccessiva è il consiglio di un atteggiamento epistemicamente vigile. Tuttavia, l'intenzione di esercitare il pensiero critico si traduce talvolta nell'eccesso opposto: sfiducia estesa anche a esperti riconosciuti sia dalla comunità scientifica che dagli stessi individui. Sia un atteggiamento passivo che eccessivamente critico rischiano di indurre gli individui in errore. Occorre dunque ridefinire secondo l'autrice il ruolo degli esperti per stabilire con loro un rapporto che non sia né di subordinazione passiva né di sfiducia. L'autrice mostra come un corretto rapporto con gli esperti passi anche attraverso l'esercizio di una particolare virtù epistemica: l'umiltà intellettuale.

Sarah Songhorian, Francesca Guma, Federico Bina e Massimo Reichlin affrontano, invece, il tema del «Progresso morale: solo questione di comportamento?» in cui si sostiene come il miglioramento morale sia un prerequisito per il progresso morale e che dovrebbe essere inteso in termini procedurali (e non sostanziali). Gli autori indicano quindi un resoconto procedurale delle capacità richieste per ragionare e per giustificare le proprie azioni e convinzioni come primo passo necessario per comprendere il contributo del miglioramento morale individuale al dibattito sul progresso

morale. Infine, gli autori sostengono che nessun resoconto motivazionale conta come una forma adeguata di giustificazione morale.

Di libero arbitrio si tratta, invece nel saggio di Francesca Guma, «Rafforzare il libero arbitrio per diventare agenti morali migliori: una prospettiva possibile?». Il saggio riguarda se e come sia fattibile il miglioramento morale individuale. Assumendo la presenza ineliminabile di condizioni che rendono difficile all'agente il controllo della propria azione e scelta, si sostiene il forte legame tra decisioni, azioni e questione del libero arbitrio. L'autrice presenta due possibili approcci per raggiungere il miglioramento morale individuale. Una proposta sostiene spinte e suggerimenti per migliorare i giudizi morali delle persone, mentre l'altra identifica modi per aumentare l'agency del soggetto. L'autrice conclude ritenendo che lo sviluppo di procedure in grado di rafforzare il libero arbitrio del soggetto permette di pensare a miglioramenti morali chiari e stabili perché genera miglioramenti non in determinati comportamenti esteriori ma nell'atteggiamento morale generale dell'individuo.

Si conclude, infine, la terza e ultima sessione del convegno con il contributo di Federico Bina dal titolo «Modelli di decisione morale: recenti progressi e rilevanza normativa». Negli ultimi decenni, la ricerca in psicologia cognitiva e neuroscienze ha alimentato un ricco dibattito sui principali meccanismi alla base del processo decisionale umano (morale) e la loro affidabilità. In questo articolo l'autore afferma che in tali processi la distinzione emozione/ragione dovrebbe essere messa da parte a favore di una struttura a doppio processo per il processo decisionale morale informato da modelli computazionali di apprendimento per rinforzo. L'autore considera infine alcune implicazioni normative di questa ricerca, sottolineandone la natura procedurale.

In conclusione i risultati di questo numero speciale di «Teoria» propongono un interessante confronto tra diversi ambiti – come quello medico, scientifico, tecnologico e filosofico – che, anche in ragione della recente pandemia, appaiono necessariamente sempre più interconnessi.

New Challenges in Decision-Making Processes: Bioethics, Neuroethics, Ethics of AI

This issue of «Teoria» results from the Conference «New Challenges in Decision-Making Processes», held at the University of Pisa on 7th and 8th April 2022 and sponsored by Project PRIN 2019/2022 on «Ethics & Tech-

nology. New Challenges for Applied Ethics». The National Scientific Coordinator of the Project is Professor Mario De Caro (University Roma Tre) and the person in charge of the Research Unit of Pisa is Professor Adriano Fabris.

The issue focuses on the new challenges in decision-making processes presented by the increasingly massive use of technologies in the network society. In particular, some aspects of bioethics, neuroethics and the ethics of artificial intelligence are investigated. The problem of "decision" has always characterized the ethical-philosophical debate, requiring the examination of some key concepts of moral philosophy such as responsibility in its various declinations, autonomy and free will. In addition to these fundamental issues, new challenges emerge today, related to contemporary developments in science and technology. This issue of «Teoria» reflect on the impact of these developments in decision-making processes, focusing mainly on the following aspects: new forms of vulnerability resulting from the use of technologies in care practices, the responsibility of the physician and the formulation of clinical judgment in the age of artificial intelligence, the redefinition of decision-making processes, in particular of moral agency, human freedom, and autonomy, in light of the latest neuroscientific developments, the effects of artificial intelligence, algorithms, statistical-predictive models, and, finally, technologies as «moral artificial agents» involved in decision making.

These topics were addressed in three sessions, the first of which focused on bioethics. In particular, the opening essay of the conference is Antonio Da Re's «Integrated clinical judgement and the physician's responsibility at the time of Covid-19». This essay tackles some of the issues raised by the hospital Triage during the Covid-19 pandemic. Such issues regard, on one hand, the «revenge of the principle of justice» and, on the other hand, «the limitation of the principle of autonomy», also on the basis of a recognition of the interdependence of all living beings. The Triage has different features and, as the author emphasizes, becomes «extreme» to the extent that its task is no longer toestablishthe order of priority in treatment, «but who can be treated and who cannot». It proceeded on the basis of criteria of clinical urgency and on the patient's survival chances, also by referring to documents and directives. The concept of clinical judgment must be thought anew by means of an analysis of more specifically clinical aspects and furtherconsiderations related to the principle of justice.

The article of Silvia Dadà, «Autonomy in telemedicine. The e-patient between independence and involvement», analyzes the patient autonomy

(e-patient), which seems to increase thanks to telemedicine. In the first part, the author considers the debate on the subject between those who think that telemedicine is an incentive for patient autonomy and those who see it as a tool for control and domination over patients' bodies. In the second part, the autor explores the concept of autonomy in an executive sense, which seems to be enhanced by telemedicine. After showing the limits of this idea, the conclusion proposes a relational perspective, which may favour the involvement between doctor and patient and the protection of dependencies and vulnerabilities.

The third paper is Francesca Marin's «End-of-life-decisions, dependence and vulnerability», which analyzes the Italian debate on Physician-Assisted-Suicide (PAS), showing that it is characterized by a misrecognition of the medical and bioethical differences between the various treatments and procedures, with the consequent risk of adopting a reductive approach to the so-called end-of-life decisions. The first part of the article concentrates on the current lexicon of end-of-life practices, which equates different practices in terms of procedures and objectives. The second part argues that improper undue equations can also be noticed by analyzing both the arguments of the Constitutional Court (provided in Order n. 207/2018 and Judgment n. 242/2019) and the recent attempt to extend the expression "life-sustaining treatments".

Finally, this first session ends with Federico Zilio's «Tough Decisions in Unclear Situations. Dealing with Epistemic and Ethical Uncertainty in Disorders of Consciousness». After an initial definition of disorder of consciousness (DoC), characterized by a compromised or complete loss of self-awareness and environmental awareness, some epistemic and methodological issues arise that characterize the disturbances of consciousness. Such are diagnostic error, prognostic uncertainty, communication with family and health workers, and the performative value of clinical language. The epistemic uncertainty emerging from these problems is deeply intertwined with ethical uncertainty, especially when it comes to clinical decisions that can lead to the deathof people whose states of consciousness (and desires) are not entirely clear. The need for epistemic and ethical prudence is suggested as a possible way, through the formulation of a balance between the two principles of inductive risk, avoiding hasty end-of-life decisions, cases of incorrect interpretation, and family-doctor communication.

The second session on neuroethics is opened by Mario De Caro and Simone Marraffa's contribution, «Consciousness and responsibility». Af-

ter exposing the disagreement between cognitive neuroscience and many ethical views based on the ordinary worldview, the authors defend the adoption of an intermediate position between the one held by traditional ethicists (who keep attributing an absolute primacy to conscious thought in moral agency) and the one held by cognitive neuroscientists and philosophers (who venture to claim that the conscious mind is indeed epiphenomenal). They argue that an alternative and more promising model may be built by referring to some suggestions by Levy, Carruthers, and King. In this light, the authors claim that cognitive neuroscience's findings – rather than showing that the conscious mind is epiphenomenal – require that we offer a fine-grained and unbiased articulation of the dialectic between unconscious processing and conscious reflection.

The following contribution is Marco Menon's «The Externalization of Decision-Making Processes in the Postindustrial Society. Vilém Flusser and the Functionary within the Apparatus». The author reads the increasingly invasive and transformative presence of artificial intelligence and algorithms in decision-making processes in the light of the concept of apparatus as developed by Flusser. On one hand, it offers a contribution to the reconstruction of the concept of *apparatus*. On the other hand, it proposes to use Flusser's categories to interpret some emerging phenomena of postindustrial society as forms of externalization of decision-making processes. The latter entails an increasing de-responsibilization of moral agents and is a phenomenon that can be framed by resorting to the notion of "functionary." Still, according to Flusser, free and existentially meaningful decision-making is possible even in the world of apparatuses, given the specifically human capacity of abstraction.

The paper of Benedetta Giovanola e Simona Tiribelli, «Equity and algorithmic decisions», focuses on one of the most urgent risks of artificial intelligence, and more specifically of algorithmic decision-making (ADM), that is, the risk of being unfair. In the first section the authors provide an overview of the discussion on fairness in ADM and show its shortcomings; in the second section they pursue an ethical inquiry into the concept of fairness, and identify its main dimensions and components, drawing insight from a renewed reflection on respect (for particular individuals too). In the third and last section the authors show how our conceptual re-elaboration of fairness can help identify the criteria that ought to steer the ethical design of ADM-based systems to make them really fair.

The session ends with Sofia Bonicalzi's «A matter of justice. The opacity of algorithmic decision-making and the trade-off between uniformity and

discretion in legal applications of artificial intelligence». The author observes how, in the last few years, decisions about matters of distributive and retributive justice have been more and more outsourced to automated systems (A.I.), and ethical challenges have progressively emerged. As compared to human adjudicators, A.I.-based systems present concrete advantages in terms of efficiency and uniformity of performance. However, striving for uniformity may also have some sizeable costs. This paper aims to focus on a specific challenge – the difficult trade-off between uniformity and discretion in judicial applications of artificial intelligence – against the backdrop of current debates in philosophy, cognitive science, and artificial intelligence. The author argues that sidestepping the peculiarities of human reasoning might have some detrimental effects on the fairness of justice administration.

The last session of the conference concentrates on the ethics of artificial intelligence (AI). The opening essay is Veronica Neri's «Artificial intelligence and consumer choices: imagination as an antidote to the processes of behavioural bias». This contribution investigates which decisions we (un-)consciously delegate to AI in a context of consumption and purchase choices (of information, imaginaries, goods and services) and what margins of autonomy the individual may still have, while trying to preserve their own evaluative capacity. While today the idea of AI neutrality appears to be outdated, it is necessary to explore whether and how AI can entangle us in clusterings or, on the contrary, allow greater moral involvement. The paper analyzes the micro-targeting strategies used by AI to encourage consumption and purchases and the profiling of our online behaviours, paying attention to behavioural advertising and behavioural bias systems. In conclusion, the author reflects on the ways the individual can stem this power of orientation of the algorithms, in particular through the capacity of imagination, typical only of human beings, which can counterbalance the mechanisms of induction to consumption rationally planned through AI.

The paper of Francesca Pongiglione addresses the relationship between «Trust, experts, and the potential side effects of critical thinking». Our epistemic duties as citizens of the global world require us to seek information to ensure that our actions do not harm others or ourselves. As we integrate that information, we should not passively accept everything we are told without thinking it through—without ensuring, at the very least, that the sources we rely on are reliable. This avoidance of excessive trust is the counsel of an epistemically vigilant attitude. However, the intention

to exercise critical thinking sometimes translates into the opposite excess: distrust improperly extended even to experts recognized as such by the scientific community and by the individuals themselves. If a passive or compliant attitude risks leading individuals into error, so does an excessively critical attitude. It is necessary to redefine the role of experts in order to establish a relationship with them that is neither one of passive subordination nor one of distrust. The author shows how a correct relationship with experts also passes through the exercise of a particular epistemic virtue—intellectual humility.

Sarah Songhorian, Francesca Guma, Federico Bina and Massimo Reichlin instead address the theme of «Moral progress: *Just* a Matter of Behavior?». The aim of this paper is to argue that moral improvement is a prerequisite for moral progress and that it should be understood in procedural (rather than substantive) terms. Thus, the authors defend a procedural account of the abilities required to reason and to justify one's actions and beliefs as the first necessary step to understand the contribution individual moral improvement offers to the debate on moral progress. Finally, they consider a challenging objection to their account, arguing that not any reason-giving account counts as a proper form of moral justification.

Francesca Guma's paper, «Becoming Better Moral Agents by Strengthening Free Will. A Possible Prospect?», asks whether and how individual moral improvement is feasible. Assuming the ineradicable presence of conditions that make it difficult for the agent to control her action and choice, the author discusses the strong relationship between decisions, actions and the question of free will. The author presents two possible approaches to achieve individual moral improvement. One proposal advocates nudges and suggestions to enhance people's moral judgments, whereas the other identifies ways to increase the subject's agency. The author concludes by arguing that developing procedures that can strengthen the subject's free will makes it possible to think of genuine and stable moral improvements, because the enhancements so generated do not concern any specific outward behaviours but the individual's general moral attitude.

The third session, and therefore the conference, ends with Federico Bina's contribution, «Models of moral decision-making: Recent advances and normative relevance». The author argues that in the last decades, research in cognitive psychology and neuroscience fostered a rich debate about the main mechanisms underlying human (moral) decision-making and their reliability. In this paper, the author first makes clear that the emotion/reason distinction should be set aside, although this does not imply casting doubt

on dual-process models in general. To support this idea, he discusses a dual-process framework for moral decision-making informed by computational models of reinforcement learning. Finally, he considers some normative implications of this research, stressing their procedural nature.

In conclusion, the results of this special issue of «Teoria» propose an interesting comparison between different fields – medical, scientific, technological, and philosophical – which, also due to the recent pandemic, necessarily appear increasingly interconnected.

T

Antonio Da Re

Il giudizio clinico integrato e la responsabilità del medico al tempo di Covid-19

1. *Oltre il paradigma dell'autonomia: questioni di giustizia e* triage *estremo*

«Nulla sarà più come prima»: è un'affermazione che abbiamo sentito ripetere più volte da quando, a fine febbraio 2020, ha cominciato a diffondersi il contagio da Sars-CoV-2. Già l'11 marzo 2020 l'Organizzazione Mondiale della Sanità dichiarava ufficialmente lo stato di pandemia. Da allora le nostre esistenze hanno sperimentato una radicale cesura, contraddistinta da malattia, morte, distanziamento, isolamento, stravolgimento della vita sociale, economica, educativa. Davvero si può sostenere che la pandemia abbia tracciato un solco profondo nelle nostre storie personali e collettive, così da poter dire che c'è un prima e, auspicabilmente, un poi della pandemia, e che appunto nulla sarà più come prima. In qualche misura questa logica di radicale cesura sembra aver colpito anche la bioetica e il suo statuto epistemologico, per molti versi incerto e ancora in fieri. Forse anche a proposito della bioetica si potrebbe quindi dire che «nulla sarà più come prima».

Alcuni segnali in tal senso provengono dalla discussione scientifica sviluppatasi negli ultimi tempi, sulla scorta dell'esperienza pandemica. Sin da subito, nel 2020, alcuni studiosi sottolineavano quanto risultasse inadeguato l'approccio individualistico che a loro dire avrebbe caratterizzato la bioetica, sin dal suo costituirsi negli anni settanta dello scorso secolo[1].

[1] J.P. Kahn-A.C. Mastroianni-S. Venkatapuram, *Bioethics in a Post-COVID World: Time for Future-Facing Global Health Ethics*, in *COVID-19 and World Order: The Future of Conflict, Competition, and Cooperation*, ed. by F.J. Gavin, H. Brands, Johns Hopkins University Press, Baltimore 2020, pp. 114-132.

 DOI: 10.4454/teoria.v42i2.153

Preso di mira era soprattutto il cosiddetto paradigma principialistico e l'enfasi riposta nel principio di autonomia intesa come autodeterminazione[2]. Per tale motivo gli autori proponevano una sorta di ritorno al paradigma della bioetica globale, già teorizzato da Van Rensselaer Potter, a cui va ascritto il merito di aver coniato in epoca contemporanea il neologismo '*bioethics*'. La caratterizzazione della bioetica in senso autonomistico e la sua declinazione in termini quasi esclusivamente biomedici avrebbero infatti comportato una serie di conseguenze problematiche quali: l'adozione di un approccio antropocentrico con la corrispettiva inadeguata considerazione delle strette interrelazioni che la forma di vita umana intreccia con altre forme di vita; una visione riduttiva della salute, intesa per lo più come bene individuale; la poca attenzione prestata alle questioni di giustizia sociale e globale che si riflettono potentemente sulle condizioni di salute; una sottovalutazione del ruolo della politiche sanitarie nella prevenzione della malattia e nella promozione della salute.

Una diagnosi simile a quella testé fornita è stata avanzata anche da chi sostiene che le questioni bioetiche più urgenti esigano soluzioni globali a lungo termine e che sia quindi indispensabile considerare la rilevanza dei determinanti sociali e ambientali della salute, superando nel contempo una prospettiva tutta incentrata sull'etica medica[3]. Ebbene, le questioni sollevate da Covid-19, riguardino esse – come si vedrà – l'organizzazione del *triage* oppure le misure di distanziamento e di lockdown o la politica vaccinale, hanno tutte a che vedere con problemi principalmente di equità e di sicurezza. A tale proposito si è parlato efficacemente di «rivincita del principio di giustizia», al quale farebbe da *pendant* la «limitazione del

[2] Il riferimento è ovviamente all'opera, così influente, di T.L. Beauchamp, J.F. Childress, *Principles of Biomedical Ethics*, Oxford University Press, New York-Oxford 2019, 8ª ed. (1ª ed. 1979). Ad onore del vero va precisato che il principialismo di Beauchamp e Childress sosteneva espressamente che nelle valutazioni e nelle scelte di ambito bioetico andavano considerati quattro principi basilari (non maleficenza, beneficenza, autonomia, giustizia) e che di essi nessun principio – neppure quindi quello di autonomia – poteva vantare a priori un qualche primato rispetto agli altri; semmai erano le questioni bioetiche specifiche a richiedere, a seconda del contesto, di privilegiare ora l'uno ora l'altro principio. Tuttavia, nonostante questa sorta di pariteticità tra i principi asserita sul piano teorico, non vi è dubbio che nelle pratiche effettive il discorso bioetico sia stato fortemente contrassegnato da una sovradeterminazione del principio di autodeterminazione (basti pensare alla rilevanza etico-giuridica attribuita nel tempo all'istituto del consenso informato).

[3] Cfr. P. Vineis, L. Savarino, *La salute del mondo. Ambiente, società, pandemie*, Feltrinelli, Milano 2021, pp. 131 s.

principio di autonomia, anche in nome del riconoscimento dell'interdipendenza di tutti gli esseri viventi»[4].

Tra i diversi temi bioetici collegati all'esperienza pandemica, che a vario titolo hanno attinenza con questioni di giustizia, un posto di prim'ordine va senz'altro riservato al problema del *triage*, che nell'occasione ha assunto connotati del tutto speciali. A cavallo tra fine febbraio e inizio marzo 2020, risultò subito chiaro che il sistema sanitario e ospedaliero delle regioni e dei paesi maggiormente colpiti dalla diffusione di Sars-CoV-2 di lì a poco si sarebbe trovato in una situazione di emergenza estrema. Di ciò si aveva avuto nelle settimane precedenti qualche avvisaglia attraverso le notizie piuttosto frammentate che giungevano dalla Cina e dalla città di Wuhan. La consapevolezza della gravità della situazione emerse in modo chiaro prima di tutto in Italia, quale primo paese occidentale duramente colpito dalla pandemia e specialmente nelle province della Lombardia e tra queste in particolare a Bergamo, e poi a seguire in altri paesi, in Europa e nel mondo. L'emergenza era dovuta alla grave carenza di risorse sanitarie di cui si poteva disporre per far fronte al dilagare del contagio: mancavano in misura adeguata le apparecchiature mediche quali i ventilatori, scarseggiavano i posti letto nei reparti ospedalieri, specie quelli di terapia intensiva (d'ora in poi TI) e sub-intensiva, e la carenza colpiva anche il personale medico e sanitario in servizio, spesso obbligato a turni di lavoro massacranti, che mettevano a rischio la loro incolumità.

L'organizzazione del *triage* in tali condizioni emergenziali non può certo essere paragonato al normale *triage*, solitamente applicato nei reparti di pronto soccorso ospedalieri: in questo caso si tratta di stabilire, sulla base del criterio dell'urgenza clinica e quindi ridimensionando il principio '*first come, first served*', chi dovrà essere curato per primo. Diverso è il *triage* che potremmo definire 'estremo', quale quello a cui si è dovuto far ricorso durante la pandemia: qui non si tratta più di stabilire chi deve essere curato prima e chi dopo, ma chi può essere curato e chi invece no, chi ha

[4] *Ivi*, p. 137. Mi permetto qui di rinviare a un mio saggio, progettato durante la prima ondata della pandemia e pubblicato a metà del 2020. In esso approfondivo alcune questioni etico-normative legate al *triage*, che ora riprendo in queste pagine, con però un armamentario concettuale più articolato; oltre al tema del *triage*, in quella sede sollevavo l'interrogativo se la giustizia non andasse interpretata come «nuovo paradigma della bioetica», quasi a voler proporre la sostituzione del paradigma autonomistico, sinora dominante; cfr. A. Da Re, *Il dilemma del triage. La deliberazione medica tra apriorismo e giudizio clinico*, in L. Napolitano, C. Chiurco (a cura di), *Senza Corona. A più voci sulla pandemia*, QuiEdit, Verona-Bolzano 2020, pp. 69-95 (specie 69-75).

la possibilità di accedere alla TI e chi invece ne rimane escluso, perché mancano i posti letto. Si può poi introdurre un'altra distinzione, quella tra *triage* diretto e indiretto. Durante la pandemia i medici hanno adottano un *triage* normale o, purtroppo, in alcuni frangenti estremo, e l'hanno fatto in modo diretto, sulla base di alcuni criteri assunti, come quelli dell'urgenza clinica e della possibilità di sopravvivenza del paziente. Di fatto però si è assistito, in modo non sempre pienamente consapevole, anche alla messa in opera di una sorta di *triage* indiretto; le strutture sanitarie, specie dei mesi di febbraio-aprile 2020, hanno concentrato gli interventi terapeutici sui malati Covid-19, rinviando in molti casi la diagnostica, le sedute chemioterapiche, gli interventi chirurgici rivolti a pazienti affetti da altre gravi patologie. Anche a proposito della forma indiretta del *triage* si è quindi posto un problema di giustizia, nell'allocazione di risorse sanitarie carenti per varie ragioni.

2. *Apriorismo, algoritmo e criterio extraclinico*

La forma del *triage* estremo rientra nell'ambito della cosiddetta medicina delle catastrofi. Il ricorso a tale forma può essere richiesto da eventi bellici particolarmente distruttivi, terribili catastrofi naturali, epidemie e pandemie molto aggressive, come accaduto con Sars-CoV-2. La domanda che immediatamente affiora è quale sia il criterio o l'insieme di criteri che giustificano il *triage* estremo, per esempio quando si presenti come '*triage* in emergenza pandemica', per riprendere un'espressione coniata dal CNB[5]. Per rispondere a tale domanda la *Società italiana di anestesia, analgesia, rianimazione e terapia intensiva* (SIAARTI) pubblicava il 6 marzo 2020 una serie di *Raccomandazioni di etica clinica sui criteri di ammissione in terapia intensiva*[6]. Il tempismo della pubblicazione faceva sì che il documento SIAARTI diventasse ben presto un testo di riferimento indispensabile per la riflessione bioetica sia in Italia che nel mondo; esso delineava,

[5] Parere del Comitato Nazionale per la Bioetica (CNB), *Covid 19: la decisione clinica in condizioni di carenza di risorse e il criterio del "triage in emergenza pandemica"*, 8 apr. 2020, pp. 1-11, https://bioetica.governo.it/media/4248/p136_2020_-covid-19-la-decisione-clinica-in-condizioni-di-carenza-di-risorse-e-il-criterio-del-triage-in-emergenza-pandemica.pdf.

[6] SIAARTI, *Raccomandazioni di etica clinica per l'ammissione a trattamenti intensivi e per la loro sospensione, in condizioni eccezionali di squilibrio tra necessità e risorse disponibili*, 6 marzo 2020, https://www.flipsnack.com/siaarti/siaarti_-_covid19_-_raccomandazioni_di_etica_clinica_-2/full-view.html

senza infingimenti e false illusioni, la situazione drammatica che di lì a qualche giorno si sarebbe determinata. Soprattutto esso poneva i medici di fronte a delle responsabilità tremende, con l'inevitabile scelta di chi tra i pazienti avrebbe potuto accedere alla TI e di chi invece ne sarebbe rimasto escluso. Di qui la necessità di delineare i criteri di giustificazione del *triage*, evitando che la deliberazione medica fosse lasciato al caso, all'arbitrio individuale o all'estemporaneità.

L'obiettivo primario dichiarato era quello di derogare dal principio 'first come, first served', che a ben vedere neppure nel *triage* normale viene del tutto seguito. Nella premessa del documento venivano formulate alcune indicazioni condivisibili su come elaborare una decisione clinica in una situazione di grave mancanza di risorse sanitarie; tale decisione doveva tener conto della proporzionalità delle cure e cercare di garantire i trattamenti intensivi ai «pazienti con maggiori possibilità di successo terapeutico». Veniva poi proposto di «privilegiare la 'maggior speranza di vita'», un criterio già di per sé più discutibile, che sembrava anticipare il criterio dell'età presente nella Raccomandazione n. 3, oggetto di innumerevoli discussioni tanto in Italia, che nel mondo. Tale Raccomandazione dichiarava espressamente che «può rendersi necessario porre un limite di età all'ingresso in TI»; si proponeva poi di «riservare risorse che potrebbero essere scarsissime a chi ha *in primis* più probabilità di sopravvivenza e secondariamente a chi può avere più anni di vita salvata, in un'ottica di massimizzazione dei benefici per il maggior numero di persone» (SIAARTI 2020, p. 5).

Non è qui il caso di ritornare sulle molte critiche che immediatamente apparvero sulla stampa quotidiana all'uscita del documento SIAARTI: esse sostenevano in sintesi che il diritto costituzionale alle cure non veniva garantito, dal momento che il limite d'età poteva fungere da possibile criterio discriminatorio nei riguardi di una ben precisa categoria di pazienti, gli anziani. Ad onore del vero va però anche precisato che in molti casi tali critiche sembravano rimuovere il problema e non avere piena contezza del fatto che le risorse disponibili erano del tutto insufficienti a coprire le richieste di cura di tutti coloro che ne avrebbero avuto necessità oltre che ovviamente diritto. D'altro canto la formulazione del criterio dell'età risultava essere problematico, anche perché ad esso veniva attribuito un peso determinante, preminente rispetto alla valutazione della comorbilità e dello status funzionale del paziente (così si esprimeva la Raccomandazione 4), quando sul piano clinico sarebbe stato più ragionevole rovesciare il rapporto e stabilire la priorità di questi due ultimi elementi rispetto all'età.

Problematica risultava essere anche la giustificazione in chiave meramente utilitaristica (si parlava di «massimizzazione dei benefici»), senza quindi sviluppare una logica etico-normativa maggiormente articolata, di cui si dirà più avanti.

Da un punto di vista etico e deontologico-professionale, oltre che clinico, il limite del criterio dell'età consiste proprio nel suo carattere aprioristico, che qualora fosse applicato in modo automatico non consentirebbe di discernere tra i differenti contesti clinici dei vari pazienti considerati. Tenendo conto del parametro della comorbilità e in presenza di patologie che hanno all'incirca la medesima gravità, probabilmente il quadro clinico di due malati affetti da Covid-19 – uno piuttosto anziano e l'altro giovane – sarebbe assai differente e le chance di guarigione del secondo sarebbero superiori. Ma non si può neppure escludere che un anziano possa avere maggiori chance di guarigione di un giovane, nel caso questi presenti parametri di comorbilità e patologie più gravi. Naturalmente il poter contare su un criterio decisionale ben preciso quale quello del limite d'età ha i suoi vantaggi: in condizioni di grande stress emotivo, con un carico di lavoro abnorme, con il rischio di infettarsi e di mettere a repentaglio la salute e persino la vita, i medici possono contare su una direttiva ben precisa, non soggetta a interpretazioni discrezionali. Il prezzo da pagare però è alto, perché in tal modo, ai fini della deliberazione, si assume un criterio rigido, predeterminato, non in grado di cogliere le differenze tra le diverse condizioni cliniche dei malati.

La logica aprioristica in qualche misura veniva radicalizzata nell'approccio algoritmico propugnato in alcuni documenti internazionali. A titolo d'esempio si possono menzionare le linee guida del *National Institute for Health and Care Excellence* (NICE): i parametri presi in considerazione per stabilire chi poteva accedere alla TI non riguardavano solo l'età, se superiore o inferiore ai 65 anni, ma anche la presenza o meno di disabilità persistenti a lungo termine – come la paralisi cerebrale – e persino di disturbi di apprendimento (*learning disabilities*) e di autismo; la valutazione doveva in ogni caso stimare il peso delle comorbilità e delle sottostanti condizioni di salute. Si individuavano poi due categorie generali di pazienti: per una, comprendente gli adulti sopra i 65 anni e senza disabilità, si proponeva di seguire le misurazioni proposte dalla cosiddetta Scala di Fragilità Clinica (*Clinical Frailty Scale*, CFS). Invece per la categoria dei pazienti sotto i 65 anni oppure di qualsiasi età, ma con disabilità, si invitava a lasciar da parte la CFS e a promuovere una valutazione individualizzata della fragi-

lità[7]. Veniva così elaborato un *Critical Care Admission Algorithm*, il cui obiettivo era evidentemente quello di fornire un sostegno al medico per decidere in tempi brevi se un paziente necessitava di essere ricoverato in un reparto normale o di TI o se invece si dovevano già attivare le cure previste per il fine vita. Tuttavia le linee guida del NICE vennero immediatamente accusate, specialmente dalle associazioni di disabili, di fondarsi su presupposti discriminatori; di qui la decisione di rivedere in modo considerevole il contenuto di tali linee, anche per evitare un possibile rischio di tal genere[8].

Ritornando al tema dell'apriorismo sulla base del limite d'età, esso è stato anche qualificato come un criterio di tipo extraclinico, che troverebbe giustificazione nel contesto straordinario ed emergenziale della pandemia[9]. Tale qualificazione, però, non può non sollevare degli interrogativi, se non altro per il fatto che essa non sarebbe definibile *prima facie* come clinica eppure dovrebbe valere ai fini della deliberazione medica, e quindi in ambito clinico. L'assunzione a priori di un criterio esterno, applicato in modo rigido, non può che comportare una svalutazione della specificità della deliberazione del medico e della sua correlativa responsabilità etica, deontologica e giuridica. In tal senso è ben strano che sia stata una società scientifica medica a proporre di adottare il criterio extraclinico dell'età, come se fossero i medici stessi a dover rinunciare alla propria specificità professionale e deontologica nella formulazione di un giudizio circonstanziato rispetto alla diversità delle situazioni cliniche dei vari pazienti in cura. V'è da chiedersi poi se sia legittimo che una società scientifica detti delle linee guida che hanno stretta attinenza con la deontologia professionale; questa però è appannaggio dell'intera comunità dei medici, da cui è regolata, e non pertiene quindi solamente a una sua componente (rianimatori e anestesisti), per quanto più direttamente chiamata in causa nelle

[7] NICE, *Covid-19 rapid guideline: critical care in adults*, 20 marzo 2020, nr. 159, https://www.ncbi.nlm.nih.gov/books/NBK566886/pdf/Bookshelf_NBK566886.pdf.

[8] Il documento originario venne rivisto già il 27 marzo 2020 e in seguito ripetutamente implementato; vd. NICE, *Covid-19 rapid guideline: managing Covid-19*, 13 apr. 2022, https://www.nice.org.uk/guidance/ng191/resources/covid19-rapid-guideline-managing-covid19-pdf-51035553326.

[9] M. Mori, *Posizione di minoranza*, postilla al parere CNB, *Covid 19: la decisione clinica*, cit., pp. 12-15, https://bioetica.governo.it/media/4248/p136_2020_-covid-19-la-decisione-clinica-in-condizioni-di-carenza-di-risorse-e-il-criterio-del-triage-in-emergenza-pandemica.pdf Si veda anche S. Camporesi, M. Mori, *Ethicists, doctors and triage decisions: who should decide? And on what basis?*, in «Journal of Medical Ethics» 47 (2021), in https://jme.bmj.com/content/early/2020/07/10/medethics-2020-106499.

problematiche deontologiche concernenti i criteri di accesso alla TI. Vi è poi una questione di legittimità ancor più importante, che chiama in causa le prerogative di competenza, in democrazia, delle istituzioni costituzionalmente stabilite[10].

Per tutti questi motivi il criterio extraclinico non può ragionevolmente qualificarsi come criterio esclusivo e neppure prioritario della deliberazione medica, nemmeno quando questa sia sottoposta a inevitabili scelte terribili e tragiche imposte dall'emergenza. Infine va ricordato che altri studiosi hanno sostenuto la validità dell'adozione di un criterio extraclinico, individuato però più che nell'età, nella qualità della vita[11], e anche qui non mancano gli interrogativi, specie se il giudizio sulla qualità della vita futura non è formulato dal paziente stesso, ma da altri (il medico) sulla base di parametri oggettivi, o presunti tali, quali potrebbero essere per esempio le misurazioni fatte valere dal modello algoritmico del NICE. Il rischio di un esito discriminatorio è forte, ed è per giunta accompagnato da una possibile deriva paternalistica.

3. Il giudizio clinico e il giudizio clinico integrato

La difesa della tesi della preminenza del giudizio extraclinico, identificato con il criterio aprioristico dell'età, presuppone che il giudizio clinico non sia adatto a guidare le scelte dei medici, per lo meno in tempi straordinari, al contrario di quanto aveva sostenuto il Comitato Nazionale per la

[10] Va però anche aggiunto che in successivi documenti, sottoscritti con altre società medica, la SIAARTI ha rivisto in modo radicale le proprie tesi. Si veda a tal proposito il documento congiunto FNOMCeO (Federazione Nazionale degli Ordini dei Medici Chirurghi e Odontoiatri)–SIAARTI, *Scelte terapeutiche in condizioni straordinarie*, 30.10.2020, https://portale.fnomceo.it/scelte-terapeutiche-in-condizioni-straordinarie-approvato-il-documento-congiunto-fnomceo-siaarti-frutto-di-un-lavoro-condiviso-supportera-il-medico-di-fronte-a-decisioni-drammatiche/ Si veda inoltre il documento condiviso da SIAARTI-SIMLA (*Società Italiana di Medicina Legale e delle Assicurazioni*), *Decisioni per le cure intensive in caso di sproporzione tra necessità assistenziali e risorse disponibili in corso di pandemia di COVID-19*, 13.1.2021, pubblicato nel portale dell'Istituto Superiore di Sanità (https://snlg.iss.it/?p=2706%20%C2%A0). Rispetto alla definizione molto netta delle *Raccomandazioni* del 6 marzo 2020, dove il criterio dell'età veniva proposto come prioritario, fondamentalmente sulla base di considerazioni economiche attinenti l'allocazione di risorse scarse, nel documento sottoscritto con la SIMLA si sostiene che «l'età deve essere considerata nel contesto della valutazione globale della persona malata e non sulla base di cut-off predefiniti»; tale considerazione fra l'altro fa seguito alla determinazione di ben precisi criteri clinici di valutazione diagnostica e prognostica.

[11] Cfr. L. Lo Sapio, *Sars-CoV-2. Questioni bioetiche*, Tab edizioni, Roma 2021, pp. 112 s.

Bioetica (CNB)[12]. In questo testo, pur in assenza di critiche esplicite verso il documento SIAARTI, emergeva come il primato del giudizio clinico comporti indirettamente una presa di distanza dalla logica dell'apriorismo e dell'algoritmo. Si affermava infatti che «ogni paziente va visto nella globalità della sua situazione clinica, tenendo in considerazione tutti i necessari fattori di valutazione», tra i quali in via prioritaria il grado di urgenza e poi la gravità del quadro clinico in atto, la comorbilità, la condizione di terminalità a breve, ecc. Si precisava quindi che l'età è un parametro importante per la valutazione clinica, anche prognostica, puntualizzando, però, che esso «non è l'unico e nemmeno quello principale». Il giudizio clinico individualizzato, inoltre, non escludeva un'analisi più ampia, riferita all'insieme dei pazienti; in tale prospettiva, si sosteneva che la priorità non andava ricercata a priori nel criterio dell'età, bensì «valutando, sulla base degli indicatori menzionati, i pazienti per cui ragionevolmente il trattamento può risultare maggiormente efficace, nel senso di garantire la maggiore possibilità di sopravvivenza».

Il parere del CNB riarticolava poi il giudizio clinico secondo la scansione dell'appropriatezza e della proporzionalità del trattamento. Con il primo elemento, quello dell'appropriatezza, ci si riferisce ad aspetti più oggettivi, che hanno attinenza, appunto con cure appropriate ed efficaci sulla base di protocolli e di linee guida consolidate; con la proporzionalità l'attenzione si focalizza su aspetti più soggettivi, riguardanti la condizione particolare del paziente: la preoccupazione è che il trattamento proposto risulti ben tollerato e non troppo gravoso o frutto di un'ostinazione irragionevole, per riprendere la terminologia della L. 219/2017.

In questo modo si stabiliva il primato dell'approccio medico fondato sul giudizio clinico. A tale riguardo c'è da chiedersi tuttavia se – come dire – esista un giudizio clinico 'puro': può il medico deliberare adeguatamente, senza tener conto della grave carenza di risorse che evidentemente limitano fortemente la sua azione? Anche se il parere del CNB non ha espressamente messo a fuoco la problematicità insita in questa dinamica tra la dimensione clinica e quella extraclinica, esso però l'ha indirettamente colta nel momento in cui ha individuato due livelli della deliberazione medica. Il primo livello è quello riferito al singolo paziente: qui l'obiettivo è formulare un giudizio clinico individualizzato, che tenga conto di molteplici fattori, *in primis* quello dell'urgenza; il secondo livello è riferito alla 'comunità dei pazienti', il che comporta una valutazione dei pazienti con mag-

[12] Parere del CNB, *Covid 19: la decisione clinica*, cit.

gior chance di vita, per i quali il trattamento (le cure intensive) possono risultare maggiormente 'efficaci'. In altri termini, si ribadisce la centralità del criterio clinico *patient centered*; ma realisticamente si tiene conto dei vincoli imposti da quelle situazioni estreme riconducibili alla cosiddetta medicina delle catastrofi.

Si potrebbe qui allora parlare con maggiore precisione di 'giudizio clinico integrato'[13]: nel *triage* in emergenza pandemica è ancora necessario il giudizio clinico elaborato sulla base dell'appropriatezza clinica e della proporzionalità, e tuttavia non è sufficiente; esso deve essere *integrato* da una valutazione che prenda in esame le condizioni di una pluralità di pazienti e la disponibilità di risorse assai limitate. Detto altrimenti, il giudizio clinico integrato deve tener conto del fatto che la decisione medica non può più solo riferirsi al singolo paziente; essa dovrà tener conto della comunità dei pazienti interessati e sarà costretta a far valere dei criteri di priorità, in cui peraltro l'elemento clinico continuerà a rimanere determinante, sia pure in una prospettiva più ampia.

Sulla base di tali criteri, e badando bene ad escludere esiti discriminatori verso persone anziane, disabili e particolarmente vulnerabili, la deliberazione medica sarà chiamata ad un compito comunque arduo nello stabilire a chi indirizzare le scarse risorse disponibili. Essa dovrà basarsi su un giudizio clinico, integrato con altri elementi essenziali ai fini della decisione; in tal senso il *triage* in emergenza pandemica mostra come sia illusorio immaginare che si possa formulare un giudizio puramente clinico. A ben vedere un giudizio clinico 'puro' non si dà neppure nel *triage* in condizioni normali e persino nella quotidiana pratica medica, sebbene qui gli elementi integrativi abbiano un minore impatto. È quindi un'astrazione ipotizzare un giudizio esclusivamente clinico; ugualmente illusorio è immaginare che si possa dare un giudizio esclusivamente extraclinico, con l'aggravante in questo caso che si misconosce la specificità della dimensione medica, consegnandola sin da subito a logiche (organizzative, economiche, politiche, ecc.) di altro genere. La dialettica tra clinico ed extraclinico è sempre presente, in contesti per così dire normali come pure straordinari; legittimare però una sorta di pariteticità tra i due momenti o addirittura la supremazia dell'extraclinico rappresenterebbe uno svilimento del ruolo

[13] A. Da Re, A. Nicolussi, *Hard Choices in the Pandemic and Guidelines. Ethical and Juridical Remarks on Medical Responsibility and Liability*, in E. Hondius, M. Santos Silva, A. Nicolussi et al. (eds.), *Coronavirus and the law in Europe*, Intersentia, Cambridge-Antwerp-Chicago 2021, pp. 411-438.

e delle finalità della medicina e la subordinazione della deliberazione clinica verso modelli standardizzati e uniformanti. Vi è una specificità del clinico che va salvaguardata, pur nella consapevolezza del condizionamento a cui esso è sottoposto da fattori vari, per esempio la carenza di risorse. L'espressione 'giudizio clinico integrato' ci dice che vi è un indubbio primato della valutazione clinica, che contraddistingue in modo speciale la deliberazione medica[14]; questa però deve inevitabilmente tener conto anche di altri aspetti, non strettamente clinici e che tuttavia incidono sul piano più strettamente clinico, e lo deve fare soprattutto (ma non solo) in condizioni di particolare emergenza.

4. *Un approccio etico-normativo plurale: di fronte all'emergenza e contro la logica dell'eccezione*

Un'analisi circostanziata dei numerosissimi documenti pubblicati da società scientifiche, ordini professionali e organismi istituzionali dei paesi colpiti dalla pandemia, fa affiorare in molti casi il loro carattere piuttosto disomogeneo, in cui indicazioni cliniche si mescolano – e a volte si confondono – con suggerimenti di natura deontologica, bioetica, giuridica, non sempre esplicitati in modo rigoroso. Carente risulta essere anche la

[14] Questa sorta di primato del giudizio clinico è riconosciuto anche nel documento comune FNOMCeO-SIAARTI, ove al punto d.1 (*Aspetti considerati in stato di assoluta necessità emergenziale*) si parla addirittura di una valutazione clinica 'caso per caso', che deve tener conto di molteplici fattori, inclusa l'età. È interessante notare al riguardo che qui si parla di età biologica e non anagrafica, precisazione di cui invece non si trova traccia nelle Raccomandazioni Siaarti del 6.3.2020 al n. 4. A proposito dell'approccio clinico, è frutto di un fraintendimento la tesi di Lo Sapio, *op. cit.*, p. 112: in quanto espressione del tradizionale paradigma ippocratico, tale approccio perseguirebbe «l'idea del medico che cura, allo stesso modo, tutti i pazienti». Si può ribattere che in linea di principio tutti, nessuno escluso, hanno diritto a essere curati e al meglio delle possibilità disponibili, ma non tutti andranno curati allo stesso modo, perché le condizioni cliniche sono ovviamente assai differenziate e perché le risorse, specie se insufficienti, andranno allocate in modo ugualmente differenziato. La deriva uniformante e indifferenziata non si produce certo nell'approccio clinico, semmai in quello aprioristico dell'età, ove un singolo paziente, per il fatto di rientrare in un determinato gruppo, verrebbe escluso dalle cure, per esempio dall'accesso alla TI. Diverso è poi il criterio aprioristico della qualità, propugnato dallo stesso autore, che sembra sottrarsi all'esito uniformante, al prezzo però di affidare (paternalisticamente) al medico il giudizio sulla qualità di vita del paziente, con conseguenti rischi discriminatori. Condivisibili sono invece le osservazioni critiche di A. Blasimme, M. Canevelli, F. Rufo, *Covid-19 e criteri etici per l'accesso alle cure intensive: un esame critico*, in «Bioetica», 28 (2020), nn. 2-3, pp. 295-314, che individuano nell'aspettativa di guarigione a breve il criterio da seguire nelle decisioni riguardanti l'accesso alla TI.

giustificazione più direttamente etico-normativa delle scelte difficili che avrebbero dovuto assumere i medici[15]. Alcuni di tali limiti sono rinvenibili anche nel documento SIAARTI, in cui non è chiaramente delineato quel duplice livello normativo che da un lato dovrebbe prevedere l'ancoraggio al diritto universale alle cure e al principio di giustizia; dall'altro dovrebbe interrogarsi su come concretamente rispettare tali principi e dar loro seguito, in un contesto di emergenza estrema. Nel documento era stata ben individuata la domanda a cui si doveva trovare risposta e che può essere così riassunta: qual è la modalità migliore per curare il maggior numero possibile di pazienti, tenendo conto della grave carenza di risorse? La risposta a tale domanda però presupponeva un riconoscimento previo e dichiarato dell'indispensabilità dei principi basilari di giustizia, equità, universalità, diritto alle cure, dovere di solidarietà (Cost., art. 2), dovere di cura da parte del medico in base alla sua posizione di garanzia, ecc.: solamente a partire da questa necessaria presupposizione poteva svilupparsi coerentemente la riflessione su quali criteri adottare concretamente ai fini della deliberazione. Il limite del documento SIAARTI è stato quello di non aver messo a fuoco con chiarezza l'importanza di questi due livelli che si implicano reciprocamente e che non possono essere scorporati neppure in condizioni di emergenza[16]. Anzi, è proprio in tali condizioni estreme che è necessario ribadire la rilevanza di questa reciproca implicazione.

Facendo ricorso a una terminologia tecnica il duplice livello normativo dell'etica si articola su un piano deontologico, con riferimento ai principi fondamentali richiamati poc'anzi, e su un piano teleologico o consequenzialistico, per il quale si potrebbe anche per semplicità adoperare la categoria dell'etica della responsabilità. L'interrogazione sulle possibili conseguenze che potranno prodursi non esclude, ma anzi presuppone il riconoscimento che in linea di principio tutti, nessuno escluso, hanno diritto alle cure, indipendentemente dall'età, dallo status sociale, dall'appartenenza a un gruppo piuttosto che a un altro e così via; il riferimento ai principi (il

[15] Cfr. S. Jöbges, R. Vinay, V.A. Luyckx, *Recommendations on COVID-19 Triage: International Comparison and Ethical Analysis*, in «Bioethics», 34 (2020), n. 9, pp. 1-12; H.-J. Ehni, U. Wiesing, R. Ranisch, *Saving the Most Lives. A comparison of European Triage Guidelines in the Context of the COVID-19 Pandemic*, in «Bioethics», 35 (2021), pp. 125-134.

[16] La duplicità di piani etico-normativi è invece opportunamente considerata nel successivo documento SIAARTI-SIMLA, *Decisioni per le cure intensive*, cit., in cui il punto 5.1 s'intitola significativamente «Principi e responsabilità»; all'elencazione dei «principi etici e giuridici» implicati si accompagna pertanto l'analisi di come debbano essere applicati e tra loro bilanciati nella concreta situazione emergenziale.

piano deontologico) è imprescindibile anche – o forse soprattutto, verrebbe da precisare – nelle condizioni di emergenza in cui ci si trova a scegliere. La straordinarietà indubbia del contesto deliberativo, quale quella che ha contraddistinto i tempi più duri e bui della pandemia, non può giustificare una sorta di straordinarietà dell'etica; semmai è proprio nelle condizioni estreme e impreviste che emerge ancor più la rilevanza del riferimento a dei principi etici basilari, che pure poi nel concreto dovranno trovare un loro bilanciamento, nell'ottica dell'etica della responsabilità.

Misconoscere o anche solo sottovalutare il fondamento dei principi del diritto alle cure, dell'equità, dell'universalità, adducendo come motivazione la condizione emergenziale che condiziona pesantemente il proprio agire, potrebbe rappresentare un pericoloso precedente e giustificare un processo analogo di misconoscimento o sottovalutazione in tempi di normalità. Più in generale, non è tanto la condizione emergenziale che dovrebbe orientare il normale; vale piuttosto la logica inversa ovvero che quanto più ci si sarà abituati nel quotidiano ad agire e a deliberare nel segno della buona prassi, tanto più si sarà in grado di affrontare quelle situazioni impreviste ed estreme, che pure si presentano. Ciò è importante anche per salvaguardare l'autonomia etica e deontologica del medico, che per potersi esercitare nell'emergenza ha bisogno – per così dire – di consolidarsi giorno per giorno nella normalità.

A tale proposito, quasi a completamento di quella duplice struttura etico-normativa dell'agire sopra delineata, merita qui menzionare un'altra tradizione etica, quella dell'etica delle virtù, che concentra la propria attenzione su colui che è responsabile dell'agire, quindi sull'agente e sui tratti del suo carattere; e l'agente sarà capace di deliberare anche in situazioni particolarmente difficili e straordinarie, quanto più sarà stato in grado nell'ordinarietà della propria esperienza di vita e professionale a scegliere e ad agire con saggezza. Non si vuole con questo minimizzare la rilevanza degli aspetti strutturali e organizzativi, in capo alle politiche sanitarie, di cui nel caso vanno denunciati i limiti, ma riaffermare la centralità dell'autonomia responsabile del medico, in tempi di ordinarietà e anche di straordinarietà; ciò richiede di non indulgere a una estremizzazione della logica dell'emergenza sino al punto da stravolgere l'ordinarietà, perché dovrebbe essere piuttosto quest'ultima a porre le condizioni per affrontare, se necessario, la straordinarietà dell'emergenza.

È possibile qui rintracciare un parallelismo con il dibattito più generale in merito al timore che, attraverso la straordinarietà di misure approntate durante la pandemia, s'instauri in ambito sanitario e non solo il cosiddetto

stato di eccezione (*Ausnahmezustand*), per riprendere l'espressione di Carl Schmitt, rilanciata da Giorgio Agamben[17]. Chiarificatrici al riguardo sono le distinzioni proposte da Roberto Esposito, per il quale lo stato di emergenza nasce da eventi imprevisti, quale per esempio il diffondersi del contagio da Sars-CoV-2[18]; esso quindi è stato in qualche modo imposto da una necessità e per questo motivo non può certo essere paragonato a un colpo di stato prodotto dalla volontà del sovrano; inoltre l'obiettivo dello stato di emergenza è quello di ricostituire la normalità interrotta. Lo stato di eccezione intende invece infrangere la normalità, a favore di un diverso ordinamento che conduce verso forme più o meno decise di dittatura. È quindi sbagliato – secondo Esposito – confondere le due espressioni, perché di fronte a eventi catastrofici lo stato di emergenza si propone di proteggere i diritti dei cittadini, anche se questo può comportare una limitazione temporanea della libertà, per esempio per salvaguardare la salute propria e altrui dagli esiti letali del contagio.

Certo, le misure emergenziali devono essere proporzionate, limitate nel tempo, pienamente rispettose delle garanzie costituzionali; e bisogna vigilare affinché la logica emergenziale non venga estremizzata sino al punto da poter scivolare pericolosamente verso eccezioni *contra legem*. Per questo è importante predisporre per tempo programmi di preparazione all'eventualità di un'emergenza quale quella pandemica (è la cosiddetta *preparedness*), come pure promuovere politiche sanitarie di potenziamento dei servizi di cura territoriali ed ospedalieri. È importante però anche valorizzare l'autonomia professionale dei medici, la cui specificità sul piano del giudizio clinico non viene meno neppure quando si trovino nella necessità di integrare tale giudizio con valutazioni di altro genere, richieste in modo speciale dal trovarsi in condizioni di particolare emergenza.

English title: Integrated clinical judgement and physician responsibility at the time of Covid-19

[17] Si vedano al riguardo i suoi interventi pubblicati nel blog https://www.quodlibet.it/una-voce-giorgio-agamben.

[18] Cfr. R. Esposito, *Immunità comune. Biopolitica all'epoca della pandemia*, Einaudi, Torino 2022, pp. 160-165. Si veda anche G. Zagrebelsky, *Introduzione* a J. Habermas, *Proteggere la vita*, il Mulino, Bologna 2022, pp. 40-43.

Abstract

The essay deals with ethical and professional issues concerning access to intensive care units during the Covid-19 pandemic, when available healthcare resources were very scarce. The article criticises the a priori criterion of the age limit proposed by SIAARTI (Italian Society of Anaesthesia, Analgesia, Resuscitation and Intensive Care), and the a priori criterion of the algorithm for determining which patients should be admitted to intensive care units. As an alternative to the extra-clinical criteria of the age limit and the algorithm, the essay proposes the integrated clinical criterion, which is more respectful of the physicians' professional responsibility.

Keywords: pandemic emergency triage; age limit criterion; algorithm criterion; extra-clinical criterion; integrated clinical criterion; medical responsibility

Antonio Da Re
Università di Padova
antonio.dare@unipd.it

T

Silvia Dadà

Autonomia in telemedicina. L'*e-patient* tra indipendenza e coinvolgimento

1. *Introduzione*

All'epoca della cosiddetta «quarta rivoluzione»[1], l'identità di ognuno di noi, nei suoi diversi ruoli e contesti, ha subìto una ridefinizione alla luce della nuova dimensione tecnologica e informazionale in cui si trova immersa. Tra questi spazi di ridefinizione della propria identità, anche la nostra esperienza della malattia, e quindi il nostro essere-pazienti, è stato notevolmente modificato dalle tecnologie dell'informazione e della comunicazione (TIC). L'*e-patient*, rispetto al suo antenato, si trova davanti a una realtà nuova, impensata. Il nostro accesso alle cure è mediato e la dimensione clinica è estesa oltre la sola relazione terapeutica in presenza. Sempre più spesso, infatti, ci serviamo di dispositivi per il monitoraggio personale dei nostri parametri e del nostro stato di salute, ci affidiamo a applicazioni per regolare il nostro stile di vita, oppure svolgiamo visite mediche a distanza, da remoto attraverso video.

L'impiego delle TIC in questo ambito ha dato vita alla cosiddetta «telemedicina», un settore ampio e variegato, che comprende tutti quegli interventi e quelle pratiche eseguite a distanza (come suggerisce il prefisso greco *tele-*). Riprendendo la definizione proposta dall'OMS:

La telemedicina è l'erogazione di servizi sanitari, quando la distanza è un fattore critico, per cui è necessario usare, da parte degli operatori, le tecnologie dell'in-

[1] Dopo quella copernicana, quella di Darwin e quella psicanalitica di Freud, la rivoluzione informatica di Alan Turing è la quarta rivoluzione, che priva l'essere umano della sua unicità anche rispetto alla capacità intellettiva e razionale: L. Floridi, *La quarta rivoluzione. Come l'infosfera sta trasformando il mondo*, Cortina, Milano 2017.

 DOI: 10.4454/teoria.v42i2.155

formazione e delle telecomunicazioni al fine di scambiare informazioni utili alla diagnosi, al trattamento ed alla prevenzione delle malattie e per garantire un'informazione continua agli erogatori di prestazioni sanitarie e supportare la ricerca e la valutazione della cura[2].

Negli ultimi anni assistiamo a una crescente diffusione di questo settore[3], che comporta anche una progressiva differenziazione interna, in base alle funzioni, alle tecnologie di cui si serve e alle sue applicazioni[4]. Ciò si rispecchia anche da un punto di vista terminologico. Al generico «telemedicina», infatti, si aggiungono col passare degli anni altre sottocategorie o specificazioni a seconda che si tratti dell'applicazione delle TIC alla prevenzione e alla gestione pubblica della salute («telehealth»)[5], alla gestione in rete delle pratiche mediche («e-Health»)[6], oppure, alla crescente diffusione della gestione personalizzata della salute tramite la tecnologia mobile collegata agli smartphones e alle apps («m-Health»)[7]. Ancor più recentemente, poi, la cosiddetta Intelligenza Artificiale e l'utilizzo dei Big Data stanno ulteriormente contribuendo ad ampliare lo spazio di applicazione tecnologica alla medicina[8].

Per quanto, quindi, la WMA affermi che «face-to-face consultation between physician and patient remains the gold standard of clinical care»[9],

[2] World Health Organization, *A Health Telematics Policy in Support of WHO's Health-For-All Strategy for Global Health Development: Report of the WHO Group Consultation on Health Telematics.* 1998, p. 10 (consultabile online: http://apps.who.int/iris/bitstream/10665/63857/1/WHO_DGO_98.1.pdf).

[3] Nel caso dell'Unione Europea, dal 2004 con l'e-Health Action Plan (rinnovato nel 2012) si susseguono una serie di politiche finalizzate alla diffusione di servizi medici online e alla loro interconnessione nel territorio europeo (documento ufficiale consultabile al link https://ec.europa.eu/health/publications/ehealth-action-plan-2012-2020_it).

[4] Per una tassonomia completa e ben articolata che tiene conto di tutti questi aspetti si rimanda a R. Bashur, G. Shannon, E. Krupinski, J. Grigsby, *The Taxonomy of Telemedicine, in* «Telemedcine and e-Health», (2011), 17, 6, pp. 484-493.

[5] S.N. Gajarawala, J.N. Pelkowski, *Telehealth Benefits and Barriers*, in «The Journal for Nurse Practitioners», (2021), 17, pp. 218-221.

[6] G. Eysenbach, *What is e-health*, in «Journal of Medical Internet Research», (2021), 3, 2, e20.

[7] H. Oh, C. Rizo, M. Enkin, A. Jadad, *What is mHealth (3): A systematic review of published definitions*, in «J Med Internet Res», (2005), 7, 1 (consultabile online : www.jmir.org/2005/1/e1).

[8] J. Morleya, C.C.V. Machadoa, C. Burra, J. Cowlsa, I. Joshid, M. Taddeo a,b,c, L. Floridi, *The ethics of AI in health care: A mapping review*, in «Social Science & Medicine», (2020), 260; S. Dash, S.K. Shakyawar, M. Sharma, S. Kaushik, *Big data in healthcare: management, analysis and future prospects*, in «Journal of Big Data», (2019), 6, 54.

[9] WMA, *WMA Statement on the Ethics of Telemedicine* (adottato dalla 58th WMA General Assembly, Copenhagen, Denmark, October 2007 e emendato dalla 69th WMA General Assembly,

l'applicazione delle telecomunicazioni e la preferenza per forme virtuali di visita o di trasmissione di dati rappresentano un servizio spesso preferito a quello standard.

Questo vasto panorama di possibilità aperte dall'uso medico delle nuove tecnologie, quindi, è innegabilmente un fattore positivo da vari punti di vista. Oltre agli odierni vantaggi legati al contenimento della pandemia, se ne possono rintracciare altri soprattutto di ordine economico e gestionale, quali la diminuzione dell'ospedalizzazione e conseguente riduzione dei costi del sistema sanitario, la maggior accessibilità delle cure anche da luoghi difficilmente raggiungibili, la personalizzazione delle cure attraverso informazioni mirate rispetto alle proprie preoccupazioni e bisogni.

Oltre ai suddetti vantaggi, un aspetto che è stato spesso posto in evidenza è la potenzialità da parte di questa pratica medica di contribuire al rafforzamento e all'ampliamento dell'autonomia del paziente: il maggior accesso alle informazioni sulla propria salute, la gestione personale delle proprie cure e la possibilità di svolgerle nel proprio ambiente domestico, sembrano in effetti tutti elementi che contribuiscono a rendere il processo di cura maggiormente legato alla volontà attiva del paziente. L'idea di *empowerment*[10] trova in questo contesto un grande impiego, proprio ad indicare questa accresciuta capacità del paziente di svolgere in modo autonomo il percorso di cura.

In effetti la diffusione delle TIC in ambito medico e sanitario ha favorito il superamento di un modello di paziente *compliant* tipico di una relazione paternalistica di cura a favore di un rapporto di *partnership*, in cui il paziente più informato e interessato, ha sempre maggior controllo sulla propria salute.

Tuttavia, la convinzione riguardo questo accrescimento dell'autonomia del paziente merita di essere indagata approfonditamente per essere compresa in modo adeguato. Per quanto infatti molta letteratura ponga l'accento sul potenziale positivo, non mancano casi in cui il fenomeno è interpretato piuttosto in senso opposto, come esempio di controllo politico del cittadino-paziente e della medicalizzazione totale della sua esistenza. Nel nostro intervento ci dedicheremo ad un'analisi di questo dibattito, per mostrare i limiti di entrambe le posizioni. Ci sembra infatti che sia opportuno,

Reykjavik, Iceland, October 2018) (consultabile online https://www.wma.net/policies-post/wma-statement-on-the-ethics-of-telemedicine/).

[10] K. Veitch, *The government of health care and the politics of patient empowerment: New Labour and the NHS reform agenda in England*, in «Law & Policy», (2010), 32, 3, pp. 313-331.

per comprendere se effettivamente la telemedicina nelle sue molteplici forme favorisca o meno l'autonomia del paziente e il suo *empowerment*, risalire a comprendere in che modo debba essere intesa l'idea di «autonomia» nel contesto degli ambienti digitali. L'*e-patient* non presenta infatti le stesse caratteristiche del paziente classico, soprattutto per quanto riguarda la sua capacità di prendere decisioni e di esercitare la propria volontà. Egli si muove in uno spazio mediato, in cui acquisisce nuove competenze e responsabilità, perdendo invece altre caratteristiche, delegate ai dispositivi, alle infrastrutture tecnologiche e al medico.

Dopo aver preso in esame il dibattito sul tema, ci soffermeremo su uno specifico senso di autonomia, che chiameremo *esecutiva* o *gestionale*. Ci sembrerà che questo senso di autonomia si trovi potenziato nel contesto di telemedicina sebbene in modo problematico, lasciando invece in secondo piano l'aspetto decisionale. Ricondurremo questo limite alla necessità di superare l'equazione tra autonomia e indipendenza, favorendo una prospettiva relazionale nella pratica di cura. L'incontro tra medico e paziente nella telemedicina ci apparirà veicolo di potenziamento dell'autonomia solo nei casi in cui vada di pari passo con una relazione personale basata su fiducia, coinvolgimento e cura delle vulnerabilità.

2. *Tra utopia e distopia*

Passiamo, quindi, all'analisi di questo dibattito, che vede da un lato dei convinti sostenitori della telemedicina come risorsa per l'accrescimento per l'autonomia e l'*empowerment* del paziente, dall'altro visioni assai più pessimistiche sulla diffusione di questi sistemi come strumenti di sorveglianza e controllo. Come sottolinea giustamente Tamar Sharon nella sua attenta ricostruzione[11], la discussione risulta fortemente polarizzata, andando a contrapporre modi differenti di intendere nozioni etiche fondamentali, quali solidarietà, autenticità e, appunto, autonomia.

I saggi di Eric Topol rappresentano dei chiari esempi dell'atteggiamento entusiastico nei confronti dell'innovazione digitale, che a suo parere co-

[11] L'autrice identifica tre principali polarizzazioni: 1. *empowerment* vs. sorveglianza; 2. miglioramento della salute generale vs. deterioramento della responsabilità statale; 3. accrescimento dell'auto-conoscenza vs. falsa imparzialità dei numeri. Per i nostri fini sarà soprattutto la prima polarizzazione a interessarci. T. Sharon, *Self-Tracking for Health and the Quantified Self: Re-Articulating Autonomy, Solidarity, and Authenticity in an Age of Personalized Healthcare*, in «Philosophy and Technology», (2017), 30, pp. 93-121.

stituisce un evento rivoluzionario di «distruzione creativa»[12]. Essa è una completa ridefinizione della nostra esistenza, paragonabile per portata ed impatto all'invenzione della stampa da parte di Gutemberg. Così come quest'ultima, infatti, ha favorito una democratizzazione della conoscenza, allo stesso modo lo sviluppo delle TIC ha reso possibile una distribuzione maggiormente egualitaria delle informazioni sulla salute. Se in passato il medico, circondato da un'aura divina, era l'unico possessore della conoscenza e il paziente dipendeva esclusivamente dalla sua volontà per poter guarire, Internet e i dispositivi medici di automonitoraggio e misurazione hanno notevolmente ampliato la capacità del malato di gestire i propri dati e informazioni personali, contribuendo attivamente alla ricerca di una soluzione personalizzata del proprio stato. L'asimmetria di conoscenze tra medico e paziente non deve, a parere dell'autore, corrispondere a una disimmetria di informazioni: contro il silenzio del medico che non comunica e quello del paziente che non può quindi esprimersi[13], si sviluppa una nuova prospettiva, più egualitaria, di scambio reciproco, paragonabile alle relazioni di *partnership* in cui il paziente diventa dirigente e ha pieno controllo e responsabilità sulla sua salute, mentre il medico è piuttosto un amministratore delegato[14].

A conclusioni analoghe giunge anche Melanie Swan, riferendosi in particolare alle TIC legate all'auto-misurazione e monitoraggio dei parametri biologici, delle opportunità per il paziente di assumere in modo consapevole la responsabilità per la propria salute possedendo nuovi strumenti per poter comprendere se stessi e i propri bisogni[15]. L'applicazione di queste tecnologie permette di sviluppare nuovi orizzonti della pratica di cura e di prevenzione, la cosiddetta «4p medicine»: una medicina predittiva, personalizzata, preventiva e partecipativa[16]. Secondo queste proposte, quindi, la

[12] E. Topol, *The Creative Destruction of Medicine*, Basic Books, New York 2012.

[13] J. Katz, *The silent Word of Doctor and Patient*, Johns Hopkins University Press, Baltimore 2002.

[14] E. Topol, *The patient Will See You Now. The Future of Medicine in Your Heand*, Basic Books, New York 2015, p. 12.

[15] M. Swan, *Emerging Patient-Driven Health Care Models: An Examination of Health Social Networks, Consumer Personalized Medicine and Quantified Self-Tracking*, in «International Journal of Environmental Research and Public Health», (2009), 6, p. 513.

[16] M. Swan, *Health 2050: the realization of personalized medicine through crowdsourcing, the quantified self, and the participatory biocitizen*, in «Journal of Personalized Medicine», (2012), 2, p. 94. Sull'espressione «4p medicine» si veda in particolare L. Hood, J.R. Heath, M.E. Phelps, B. Lin, *Systems biology and new technologies enable predictive and preventative medicine*, in «Science», (2004), 306 (5696).

medicina, nel suo divenire "informazionale" permette al paziente di divenire maggiormente autonomo, suggerendo che maggiore è l'informazione, migliore è l'efficienza del servizio sanitario, maggiori sono le opportunità di esercizio di controllo su di sé e di libertà del paziente.

Sul versante opposto troviamo invece chi, come Deborah Lupton, pensa che proprio queste nuove tecniche di telemedicina, monitoraggio e autogestione dei propri dati ed esami siano espressione di una più subdola e sottile forma di sottomissione a un controllo biopolitico[17] da parte delle istituzioni, per interessi di tipo economico e di ordine pubblico. La ricostruzione ottimistica data dai primi, secondo questa visione, non tiene infatti sufficientemente presente alcune centrali ambiguità, in particolare legate alla medicalizzazione dell'ambiente domestico e alla pervasività della questione della salute in tutti gli ambiti della vita del paziente. Incentivata soprattutto per questioni economiche al fine di far fronte alla riduzione delle risorse nella sfera sanitaria, la telemedicina risulta funzionale all'idea di un paziente-consumatore, frutto del modello neoliberale che baratti la sua individuale autosufficienza con minori tutele da parte dello stato. L'*e-patient*, chiamato dai sostenitori della salute digitale «Smart Patient»[18], «Emancipated consumer»[19] non ottiene, alla luce della sua maggior conoscenza un maggior controllo di sé, bensì si trova sempre più oppresso da una normativizzazione del suo stile di vita, che deve sottostare a criteri sempre più alti di efficienza e funzionalità. Come dei moderni Panopticon[20], i sistemi di misurazione, monitoraggio e autogestione della cura fanno sì che il corpo venga quantificato e regolamentato[21], e lo standard di salute possa essere utilizzato per fini di controllo e disciplinamento intrecciando la sfera della cura con altri ambiti della vita dell'individuo esercitando su di essi una decisiva influenza. Pensiamo, ad esempio, alla possibilità, per un'agenzia

[17] Il riferimento filosofico principale è al pensiero di Michel Foucault (M. Foucault, *Nascita della clinica. Una archeologia dello sguardo medico*, Einaudi, Torino 1998). Sul tema si veda D. Lupton, *Foucault and the medicalisation critique,* in A. Petersen, R. Bunton (eds.), *Foucault, Health and Medicine*, Routledge, London 1997, pp. 94-110. Per una ricostruzione chiara e completa sul dibattito biopolitico si veda R. Esposito, *Bios. Biopolitica e filosofia*, Einaudi, Torino 2004.

[18] T. Topol, *The Creative Destruction of Medicine*, cit.

[19] E. Topol, *The patient Will See You Now,* cit.

[20] L'idea di Jeremy Bentham è resa nota su un piano politico e legata alle pratiche di controllo e dominio sui corpi da parte di Michel Foucault (M. Foucault, *Sorvegliare e punire. Nascita della prigione*, Einaudi, Torino 2014).

[21] D. Lupton, *Quantifying the body: Monitoring, performing and configuring health in the age of mHealth technologies.,* in «Critical Public Health», (2013), 29, 3, pp. 393-404.

assicurativa, di fornire i suoi servizi in base allo stile di vita del richiedente; oppure la possibilità, d'altra parte non così immaginaria, di controllare i ritmi di vita del lavoratore, le sue abitudini, tramite i sensori e gli altri dispositivi digitali di misurazione. Questi aspetti costituiscono chiaramente degli ostacoli all'autonomia del paziente-cittadino, tanto che nel contesto di telemedicina, «'empowerment' becomes a set of obligations»[22].

Le due prospettive, tra loro apparentemente opposte, sembrano da un lato propendere per uno scenario futuro utopico, segnato da opportunità di emancipazione e di padronanza totale da parte del paziente; dall'altro invece descrivono una realtà dai tratti orwelliani, in cui la sorveglianza e il dominio costituiscono gli orizzonti ultimi di senso.

Tuttavia, per quanto sembrino opposte, le due posizioni, a ben vedere, non sono così incompatibili, e questo per due principali ragioni.

La prima è che il controllo e il possesso delle proprie informazioni e quindi la possibilità di utilizzarle e interpretarle non esclude a priori la possibilità che anche altri enti, quali il potere statale, oppure gruppi economici, possano possederle. L'essere un paziente emancipato, in questo senso, non esclude la possibilità di essere un cittadino, un lavoratore, o, più genericamente, un soggetto sorvegliato (e, viceversa, la libertà sul piano civile e politico non assicura a priori l'assenza di un modello medico di tipo paternalistico). Il controllo sui propri dati e sul proprio corpo non implica quindi necessariamente l'*esclusività* di tale rapporto. A ben vedere, quindi, i due discorsi possono convivere, affrontando la questione da punti di vista differenti, quello interno alla salute e quello esterno, di tipo politico-sociale.

La seconda ragione, centrale per il nostro discorso sull'autonomia, è legata al fatto che entrambe le prospettive ragionano su questo concetto in termini di potere e di rapporti di forza. Nel caso dei sostenitori, infatti, questa innovazione è percepita come una sorta di emancipazione: «we now have a formula for freedom, for relative autonomy from the canonical medical community that forced patients to be subservient and dependent. No longer»[23]. Allo stesso modo, nel caso degli oppositori, il controllo esercitato sul cittadino-paziente costituisce la limitazione da parte dei dispositivi di dominio[24]. Che si tratti, come nel primo caso, del medico, o del sistema

[22] D. Lupton, *The digital engaged patient: self-monitoring and self-care in the digital era*, in «Social Theory and Health», (2013), 11, 3, p. 261.

[23] E. Tupol, *The patient Will See You Now*, cit. p. 276.

[24] Potremmo parlare di un approccio «esternalista», che individua le limitazioni dell'autonomia in impedimenti esterni alla volontà e all'intenzione dell'individuo, riconducibili all'ambiente

di potere di riferimento, nel secondo, in entrambe le situazioni l'*e-patient* determina la sua autonomia interna ai rapporti di forza in cui è coinvolto. L'autonomia è percepita come una liberazione dal medico e dal suo potere attraverso il possesso dei propri dati, una liberazione dalla relazione, la quale è vista come un conflitto tra forze contrapposte, e non animate da un comune intento, ossia il bene del paziente[25]. Secondo entrambe queste letture, quindi, l'autonomia è da intendersi come *non-interferenza* e *indipendenza*, per cui la relazione tra medico e paziente è interpretata come un rapporto di limitazione di tale autonomia, ed è attraversata dal perenne tentativo da parte del medico del controllo sul paziente (in termini, quindi, paternalistici). La realizzazione di un processo decisionale autonomo, in quest'ottica, è raggiungibile attraverso un rapporto univoco ed esclusivo tra il paziente e il possesso dei propri dati.

Per quanto siano innegabili sia i vantaggi apportati dall'avvento della telemedicina in fatto di maggior indipendenza del paziente e per quanto, dall'altro lato, anche le preoccupazioni di chi vede un possibile strumento di dominio sui corpi siano giustificate; tali considerazioni non risolvono la nostra questione, quanto piuttosto aprono a una più profonda riflessione sul concetto di autonomia. Ancor prima di chiedersi se l'autonomia venga accresciuta o meno nel nuovo orizzonte della telemedicina, ciò su cui si deve fare chiarezza è che cosa si intenda per «autonomia». La visione di autonomia proposta da queste due posizioni è, infatti, *una* specifica opzione, e la scelta di questa piuttosto che di un'altra comporta importanti conseguenze etiche. È quindi necessario a questo punto indagare il concetto in altre direzioni, al fine di comprendere quale di esse è maggiormente presente in telemedicina e per valutarne le criticità, cercando di proporne un senso differente da valorizzare.

e al contesto in cui vive (L. Chiapperino, M. Annoni, P. Maugeri, G. Schiavone, *What Autonomy fot Telecare? An Externalist Approach,* in «The American Journal of Bioethics», (2012), 12, pp. 55-57).

[25] Il riferimento è qui al principio di beneficità, e alla centralità che occupa nel pensiero di E.D. Pellegrino, D.C. Thomasma, *Per il bene del paziente. Tradizione e innovazione nell'etica medica*, Paoline, Cinisello Balsamo 1992. Sul tema si veda anche F. Marin, *Il bene del paziente e le sue metamorfosi nell'etica biomedica*, Mondadori, Milano 2016.

3. *Quale autonomia?*

Nei principali testi bioetici l'idea di autonomia è considerata in un senso multidimensionale, che tiene conto sia dell'idea di indipendenza che di quella di auto-determinazione, intesa come coincidente con la stessa capacità decisionale. Come si legge infatti nella seconda edizione alla *Encyclopedia of Bioethics*:

> The concept of autonomy in moral philosophy and bioethics recognizes the human capacity for self-determination, and puts forward a principle that the autonomy of persons ought to be respected [...] There are three elements to the psychological capacity of autonomy: *agency*, *independence*, and *rationality*. Agency is awareness of oneself as having desires and intentions and of acting on them. . . . Independence is the absence of influences that so control what a person does that it cannot be said he or she wants to do it[26].

Ciò appare ancora più evidente nella definizione del principio di rispetto dell'autonomia fornito da Beauchamp e Childress in *Principles of Biomedical Ethics*[27] in cui gli autori chiariscono che essa debba essere considerata, nel loro modello, non tanto come autonomia della *persona*, prospettiva che legherebbe la riflessione a specifiche teorie filosofiche e morali, quanto come autonomia della *decisione*. È infatti facilmente riscontrabile che anche persone autonome possono trovarsi a compiere scelte che non sono tali, e viceversa. Per quanto riguarda quindi le decisioni, esse possono rientrare nel terreno dell'autonomia quando sono 1) intenzionali, 2) comprese appropriatamente 3) compiute senza costrizione (interna o esterna al soggetto). A causa degli ultimi due punti, si possono avere diversi gradi di autonomia, e, non potendo pretendere sempre una completa e assoluta comprensione e assenza di qualche influenza, un livello medio di «substantial autonomy» è considerato sufficiente per decretare un'azione come autonoma[28]. La possibilità di compiere le proprie decisioni in modo volontario, senza costrizioni o influenze esterne e potendo pianificarle secondo i propri obiettivi rappresenta, secondo questa proposta, il senso più completo di autonomia.

[26] B. Miller, *Autonomy*, in W.T. Reich (ed.), *Encyclopedia of Bioethics*, 2nd ed., Macmillan, New York 1995, pp. 215-216.

[27] T.L. Beauchamp, J.F. Childress. *Principles of Biomedical Ethics*, Oxford University Press, New York 2013.

[28] Anche nel caso del modello proposto da Beauchamp e Faden vediamo tornare questi tre elementi: 1) *Understanding*, 2) *Intentionality* e 3) *Vouluntariness*.

Ancora più nettamente, essa è valorizzata da Hugo Tristram Engelhardt, in cui l'autonomia, poi rinominata «principio del permesso», è vista come la base per un'etica liberal-libertaria, l'unico possibile punto d'incontro in un contesto pluralistico[29]. Per quanto differenti in tutte queste interpretazioni del tema dell'autonomia, e in altre, quali ad esempio quella di Dworkin[30], essa è collegata strettamente all'autodeterminazione e all'indipendenza come centro della capacità decisionale.

In un articolo uscito nel 2009 sul «The American Journal of Bioethics», dal titolo *Patient Autonomy for the Management of Chronic Conditions: A Two-Component Re-Conceptualization*[31], Naik e colleghi sostengono che l'incremento notevole, negli ultimi decenni, delle malattie croniche e il connesso sviluppo di quesiti bioetici oltre la dimensione «di frontiera»[32], ha portato a un progressivo spostamento del modo di intendere l'autonomia oltre lo spazio della decisione puntuale, tipica dei casi di patologie acute. In questo senso viene proposta l'idea di una *executive autonomy*, intendendo con essa «the capacity to perform complex self-management tasks, especially those related treatment planning and implementation»[33]. Tale dimensione dell'autonomia, che potremmo definire anche gestionale (visto il suo legame con l'idea di *self-management*)[34] a parere degli autori, permette di porre maggior attenzione sull'intenzionalità del paziente, sulla sua capacità non soltanto di prendere decisioni, ma anche di pianificare la propria terapia, gestirla nella quotidianità e portarla a termine. In molti casi, infatti, soprattutto legati a patologie croniche, l'interruzione di una terapia non corrisponde a un rifiuto delle cure o alla volontà di non condurla da parte della persona curata, bensì a un limite rispetto a questo aspetto gestionale. La capacità di attivare dei processi di decisione sembra in questo contesto perdere il suo carattere puntuale e diventare invece un aspetto dinamico, che riguarda le intenzioni, capacità e le competenze del paziente in modo più diretto.

[29] H.T. Engelhardt, *Manuale di bioetica*, Il Saggiatore, Milano 1991.

[30] R. Dworkin, *Life's Dominion*, Harper Collins, London 1993.

[31] A.D. Naik, C.B. Dyer, M.E. Kunik., L.B. McCullough, *Patient Autonomy for the Management of Chronic Conditions: A Two-Component Re-Conceptualization*, in «The American Journal of Bioethics», (2009), 9, 2, pp. 23-30.

[32] G. Belinguer, *Bioetica quotidiana e bioetica di frontiera*, in S. Di Meo, C. Mancina, *Bioetica*, Laterza, Roma-Bari 1989, pp. 5-18.

[33] A.D. Naik, *et al.*, *Patient Autonomy for the Management of Chronic Conditions*, cit., p. 26.

[34] M. Schremer, *Telecare and self-management: opportunity to change a paradigm?*, in «Journal of Medical Ethics», (2009), 35, pp. 688-691.

Tornando al contesto preso da noi in esame, ossia quello della telemedicina, sembra che questo senso maggiormente esecutivo o gestionale dell'autonomia venga favorita e valorizzata dalle TIC. Una delle maggiori applicazioni, infatti, che la telemedicina ha incontrato, è proprio quella alla cura e al monitoraggio delle malattie croniche[35]: dispositivi sempre più sofisticati permettono al paziente, fornendogli istruzioni e misurazioni, di prendere in carico la propria patologia senza bisogno dell'intervento e della supervisione dei professionisti e del rapporto faccia-a-faccia. Il paziente quindi impara ad avere a che fare con la propria patologia e a gestirla, comprende e interpreta il senso dei dati ottenuti dalle misurazioni, si comporta di conseguenza (ad esempio adattando i dosaggi o cambiando alcuni aspetti del suo stile di vita). Nella maggior parte dei casi, le tecnologie di telemonitoraggio favoriscono un aumento dell'auto-sufficienza del paziente in quanto questo diviene una sorta di proto-professionista[36] che segue su un piano esecutivo delle indicazioni pratiche ma il medico rimane la figura di riferimento per quanto riguarda gli aspetti decisonali.

Non solo per il monitoraggio e i dispositivi di misurazione, ma anche nel caso delle televisite, infatti, la gestione indipendente del rapporto sembra incrementato, rendendo possibile una relazione più immediata da un punto di vista temporale e integrata all'ambiente del paziente[37]. Tuttavia, a tale indipendenza esecutiva non corrisponde ad un'altrettanta crescita nel ruolo decisionale del paziente, il quale spesso si trova come uditore passivo ad accogliere un confronto tra varie figure professionali:

> Although telemedicine may be a shared care environment that is characterized by collaborative interaction between all participants, there is a risk that interprofessional interaction might exclude the patient from important aspects of the consultation[38].

Sembra quindi che nell'ambito della telemedicina il livello esecutivo dell'autonomia, intesa come indipendenza nella gestione del proprio pia-

[35] F. Aberer, D.A. Hochfellner, J.K. Mader, *Application of Telemedicine in Diabetes Care: The Time is Now*, in «Diabetes Therapy», (2021), 12, pp. 629-639. P. Alvarez, A. Sianis, J. Brown, A. Ali, A. Briasoulis. *Chronic disease management in heart failure: focus on telemedicine and remote monitoring*, in «Reviews in Cardiovascular Medicine», (2021), 22, pp. 403-413. H. Bitar, S. Alismail, *The role of eHealth, telehealth, and telemedicine for chronic disease patients during COVID-19 pandemic: A rapid systematic review*, in «Digital Health», (2021), 7, pp. 1-19.

[36] M. Schremer, *Telecare and self-management*, cit. p. 689.

[37] *Ibidem*.

[38] Y. Pappas, J. Vseteckova, N. Mastellos, G. Greenfield, G. Randhawa, *Diagnosis and Decision-Making in Telemedicine,* in «Journal of Patient Experience», (2019), 6, 4, p. 302.

no terapeutico e della misurazione e monitoraggio dei propri parametri, sia quello più sviluppato e fondamentale, lasciando invece in secondo piano la componente maggiormente decisionale. Va inoltre considerato che, anche dal punto di vista esecutivo, tale accrescimento non è omogeneamente distribuito e nemmeno esente da complicazioni.

L'*e-patient*, deve possedere un bagaglio di specifiche competenze per l'accesso e la gestione delle TIC coinvolte nella sua cura: per essere pazienti è necessario essere in grado di relazionarsi con gli strumenti tecnologici, avere una adeguata formazione e un accesso adeguato ad essi. Si tratta di un cambiamento nella identità del paziente di portata considerevole. Anche prima, nella medicina tradizionale, indubbiamente, una forma di famigliarità con le tecniche era richiesta, ma in misura meno pervasiva e specializzata, e soprattutto mediata dalla competenza medica e professionista. Oggi, invece, così come la gestione della terapia è affidata al paziente, anche il funzionamento dei dispositivi e delle tecnologie utilizzate vi dipendono. I pazienti che più frequentemente usufruiscono di questo servizio, o coloro che ne avrebbero maggior necessità, sono tuttavia spesso coloro che con maggior difficoltà riescono ad accedervi. Come dice giustamente Dorn, «mHealth use currently is skewed towards who need the least help: the young, the fit, and the educated»[39]. La popolazione anziana, in particolare, è quella che ha meno familiarità con il mezzo tecnologico, e per cui l'apprendimento del funzionamento risulta maggiormente problematico. Anche laddove possieda una conoscenza di internet, o una basilare capacità di utilizzo dei mezzi tecnologici, questa fascia di popolazione incontra in molti casi difficoltà nella relazione con l'interfaccia dei dispositivi[40]. Non solo: anche nel caso delle televisite e della realtà virtuale, i pazienti anziani mostrano maggiori difficoltà a riconoscere il volto del medico come una figura professionale e a prenderla sul serio[41]. Anche

[39] S.D. Dorn, *Digital helath: Hope, hype, and Amara's law,* in «Gastroenterology», (2015), 149, 3, p. 516. Simili considerazioni sono espresse anche in A. Ho, O. Quick, *Leaving patients to their own devices? Smart Technology, safety and therapeutic relationships*, in «BMC Medical Ethics», (2018), 19, 18.

[40] «[…] one interesting thing found was that the older adults tendend to blame themselves rather than the interface features when they could not coplete the tasks more often than the younger adults did» (Y.J. Chun, P.E. Patterson, *A usability gap between older adults and younger adults on interface design of an Internet-based telemedicine system*, in «Work», (2012), 41 Suppl. 1, p. 352.

[41] N.M. Hjelm, *Benefits and drawbacks of telemedicine*, in «Journal of Telemedicine and Telecare», (2005), 11, p. 67. Sul tema si veda anche il più recente E. Hargittai, A.M. Piper, M.R.

per i pazienti con difficoltà cognitive o percettive[42] o coloro che sono meno alfabetizzati dal punto di vista digitale si trovano a dover affrontare l'ulteriore criticità legata all'acquisizione delle competenze necessarie all'utilizzo del mezzo. Si presenta quindi con forza, anche in questo contesto, il problema del *digital divide*[43], questione che soprattutto nell'attuale epoca di pandemia ha rivelato la sua importanza, non soltanto riguardo all'autonomia, ma anche rispetto alla giustizia e all'equità della telemedicina[44].

Proprio a commento della proposta di Naik e colleghi, Camila Scaclan e Ian Kerridge sottolineano questo punto problematico: «One of the problems with autonomy is that it fails to account for the moral significance of vulnerability in the setting of serious illness and dependency on healthcare and assumes that decisions are, and even should be rational»[45].

Sembra quindi che l'acquisizione di maggior autonomia, anche nel senso esecutivo, sia favorita in molti casi in telemedicina, ma anche ostacolata. Il dispositivo digitale diviene così organo-ostacolo per il paziente, che si trova quindi ad oscillare tra un accrescimento della propria autosufficienza e l'incremento delle proprie vulnerabilità. Alle già presenti vulnerabilità legate al suo stato di salute e quindi alla malattia, troviamo sommarsi, nel contesto della telemedicina, un insieme di vulnerabilità che possiamo definire *digitali*[46], legate alla sua relazione con i dispositivi e alle capacità necessarie per utilizzarli adeguatamente.

Morris, *From internet access to internet skills digital inequality among older adults*, in «Univers Access Inf Soc», (2019), 18, pp. 881-890.

[42] J. Van Cleave, C. Stille, D.E. Hall, *Child Health, Vulnerability, and Complexity: Use of Telehealth to Enhace Care for Children and Youth with Special Health Care Needs*, in «Academic Pediatrics», (2022), 22, 2, pp. 34-40.

[43] I. Litchfield, D. Shukla, S. Greenfield, *Impact of COVID-19 on the digital divide: a rapid review*, in «BMJ Open», (2021), 11, e053440.

[44] J. Roy, D.R. Levy, Y. Senathirajah, *Defining Telehealth for Research, Implementation, and Equity*, in «J Med Internet Res», (2022), 24(4), e35037.

[45] C. Scanlan, I.H. Kerridge, *Autonomy and Chronic Illness: Not Two Components But Many*, in «The American Journal of Bioethics», (2009), 9, 2, p. 41.

[46] Un senso simile può ritrovarsi nell'utilizzo del concetto di «tecno-vulnerabili» proposto da Antonio Carnevale, il quale individua tre livelli di vulnerabilità: una, naturale, dovuta alla nostra finitezza, una dovuta alla dipendenza della tecnica come correttivo a questa prima vulnerabilità, e infine la tecno-vulnerabilità, dovuta alla multiforme azione delle nuove tecnologie su di noi (A. Carnevale, *Tecno-vulnerabili. Per un'etica della sostenibilità tecnologica*, Orthotes, Napoli-Salerno 2017).

L'autonomia esecutiva, quindi, rivela le difficoltà di conciliazione con il concetto di vulnerabilità, in quanto rischia in alcuni casi di accrescerne la portata piuttosto che alleviarne il peso. Proprio a causa di ciò, essa si rivela non solo potenziata dalla telemedicina, ma anche presupposta[47].

4. *Conclusioni. Per un'autonomia relazionale in telemedicina*

Per tornare quindi alla domanda di partenza, ossia se possiamo parlare di un potenziamento dell'autonomia nel contesto della telemedicina, la risposta non può che essere composita.

Da un lato, infatti, il senso più gestionale ed esecutivo sembra effettivamente accresciuto dalle possibilità offerte dalle TIC, soprattutto nel caso delle malattie croniche, in cui il paziente può integrare nella sua quotidianità la cura senza necessità di un intervento esterno e senza il ricorso a un numero maggiore di ospedalizzazioni. Il ruolo dei professionisti, in quest'ottica, rimane quello di controllori esterni a cui i dati raccolti dal paziente sono inviati e continuano a intervenire in modo specifico per quanto riguarda l'ambito della decisione terapeutica. Ma, anche in questo caso, come nei precedenti, la prospettiva è quella di un'autonomia intesa come *indipendenza*, e la relazione è interpretata come qualcosa che deve assumere il carattere minimale di scambio di informazioni necessarie al mantenimento dello stato di salute. I limiti di un'idea di autonomia concepita in questo modo si fanno però più evidenti dal momento in cui si riconosce la difficoltà di estendere in modo uniforme questo obiettivo: le vulnerabilità digitali che comporta questa prospettiva fanno quindi sì che l'autonomia, ancor prima di essere un obiettivo, sia data per presupposta e laddove non sia già presente rischi di non essere mai raggiunta.

La possibilità di un possesso diretto dei propri dati e del loro monitoraggio, così come il diretto controllo sulla gestione quotidiana del proprio stato di salute, di certo sono entrambi aspetti che contribuiscono in modo fondamentale alla crescita dell'autonomia del paziente e alla sua capacità di prendere decisioni. Non si può inoltre negare che tale visione della relazione clinica in termini di rapporti di forza sia una possibilità concreta e

[47] Ricoeur, sulla scia di Kant, in una sua riflessione sul soggetto di diritto descrive in termini paradossali l'autonomia a causa di questa doppia natura di condizione di possibilità e di compito. Egli parla, a questo proposito, di autonomia come «idea-progetto» (P. Ricoeur, *Autonomia e vulnerabilità*, in *Il Giusto. Vol. 2*, Effatà Editrice, Torino 2007, p. 95).

che, malgrado il superamento di un paradigma paternalistico permangano in molti contesti atteggiamenti di gestione conflittuale della relazione e di sottomissione del paziente all'autorità del medico. Pur riconoscendo questi aspetti, tuttavia, un'idea di autonomia intesa esclusivamente in questo modo non conduce il paziente all'indipendenza *dalla* relazione e non, piuttosto, *all'interno di essa.*

Affinché ciò avvenga, e questa è la proposta con cui concludiamo questo saggio, è necessario leggere il concetto di autonomia in un senso maggiormente *relazionale.* Un concetto, questo di «autonomia relazionale», che ha visto negli ultimi anni un grande sviluppo, su un terreno filosofico e politico[48] ma soprattutto bioetico, tanto da parlare di un vero e proprio «relational turn of bioethics»[49]. Tale riflessione ha trovato nell'indagine femminista[50] e dell'etica della cura[51] il suo maggior spazio, proponendo una messa in discussione dell'idea individualistica di autonomia classica[52], che permettesse di pensare insieme autonomia e vulnerabilità, autodeterminazione e dipendenza. Seguendo questa proposta l'autonomia è quindi inserita e complicata alla luce delle relazioni sociali in cui siamo quotidianamente immersi: è impossibile prescindere da questa rete nella ricostruzione del processo decisionale del singolo. Oltre ad essere irrealistico, ciò non è nemmeno auspicabile, in quanto priverebbe il paziente stesso di elementi fondamentali al raggiungimento dell'autodeterminazione. Tra le relazioni affettive e sociali in cui è inserito nella sua vita, anche il rapporto col medico è uno spazio di creazione della propria autonomia, in quanto garantisce non solo un supporto tecnico per la comprensione dei dati (aspetto che non può avvenire tramite l'esclusivo rapporto paziente-dato) ma anche di un vero rapporto di cura. L'idea di un'autonomia relazionale consente quindi di pensare il paziente come individuo inserito in una rete di relazioni di cura che contribuiscono attivamente come supporto del processo decisionale. Ciò significa, insieme, riconoscere la dipendenza

[48] C. Mackenzie, N. Stoljar, *Relational Autonomy. Feminist Perspectives on Autonomy, Agency and the Social Self*, Oxford University Press, New York 2000.

[49] R. Jennings, *Reconceptualizing Autonomy: A Relational Turn in Bioethics*, in «Hastings Centre Report», (2016), May, 46(3), pp. 11-16.

[50] Oltre al già citato C. Mackenzie, N. Stoljar, *Relational Autonomy*, cit., si veda anche S. Sherwin, *Whither Bioethics? How Feminism Can Help Reorient Bioethics*, in «IJFAB: International Journal of Feminist Approaches to Bioethics», (2008), 1 (1), pp. 7-27.

[51] M. Verkek, *The care perspective and autonomy*, in «Medicine, Health Care and Philosophy», (2001), 4, pp. 289-294.

[52] Sulla critica all'autonomia come "mito" si veda M. Fineman, *The autonomy myth. A theory of dependency*, New Press, New York 2004.

(sia dal medico che dai fattori sociali), ma senza ridurla al dominio e alla subordinazione del paziente. A differenza delle visioni precedenti, non tenta di risolvere la criticità legata all'abuso di potere nella relazione con l'eliminazione della relazione stessa, bensì ne riconosce l'importanza nei suoi aspetti positivi come strumento per valorizzare la stessa centralità del paziente e ne combatte gli elementi nocivi legati al rischio di prevaricazione e all'asimmetria tra membri del rapporto clinico. Come si legge nella raccolta sul tema curata da Downie e Llewelyn, questa prospettiva «affirms the fact of relationships and the importance of attending to their nature and to what is required of them to ensure well-being and flourishing» e allo stesso tempo, secondo l'influenza femminista da cui sorge, essa permette di «recognizing oppression (particularly, but not exclusively, of women) and seeking its end»[53].

Si assiste quindi a un doppio movimento che da un lato permette di valorizzare la decisione del paziente, ma senza per questo sradicarla dalla sua realtà socialmente situata (e quindi nemmeno dal rapporto col medico); dall'altro lato nel rintracciare contesti di abuso dell'autonomia e di diminuzione dell'indipendenza, si riconosce come soluzione il coinvolgimento positivo nel processo decisionale di figure di riferimento quali famigliari e medici.

Un altro aspetto particolarmente significativo di questa idea relazionale risiede inoltre nell'essere il *risultato* del processo decisionale condiviso, non il presupposto, non generando, per questo, ulteriori vulnerabilità. L'autonomia, quindi, si amplia andando a esercitarsi anche in casi di dipendenza, dando un maggior rilievo a figure di cura, non solo quella del medico, ma anche infermieri, *caregivers* e famigliari. Il fine ultimo, quindi, non è più l'indipendenza del paziente, ma il suo coinvolgimento nelle scelte e nelle cure che lo riguardano. In questo modo il paziente non è privato della propria capacità decisionale, ma essa si sviluppa all'interno delle relazioni in cui si trova a vivere, facendo sì che possa mantenere un ruolo da protagonista ma possa riconoscersi nella sua completezza solo se inserito nella rete di relazioni significative.

Se tale idea ha avuto sin da subito, su un piano teorico, una funzione critica rispetto alla concezione dell'autonomia come indipendenza, essa ha assunto, nel corso del tempo, anche un valore maggiormente costruttivo

[53] J. Downie, J.J. Llewellyn, *Being Relational: Reflections on Relational Theory and Health Law*, UBC Press, Vancouver 2012, p. 6.

e applicativo in vari settori medici[54]. I terreni in cui maggiormente sono state riscontrate applicazioni proficue sono quelle del fine vita e delle cure palliative, in cui il tema del coinvolgimento di figure terze è di grande centralità e impone una riflessione sulla relazionalità dell'autonomia del paziente[55]. Vista, poi, la radice femminista della teoria, non stupisce che essa abbia poi trovato spazio di approfondimento nell'etica riproduttiva e su temi quali l'aborto o la maternità surrogata[56].

Ci sembra che questo possa trovare un ulteriore terreno proficuo nell'ambito della medicina digitale e della telemedicina, che a sua volta può accrescere le sue potenzialità e risorse. Solo in un approccio relazionale di questo genere, infatti, gli aspetti precedentemente posti in luce nella prospettiva di telemedicina, ossia quelli del controllo sui dati e quelli della gestione quotidiana della terapia, contribuiscono in modo proficuo e significativo al potenziamento del coinvolgimento del paziente nelle decisioni e non rimangono attestati di semplice indipendenza. Essi, inseriti in questo contesto relazionale, forniscono un quadro più verosimile rispetto alla situazione del paziente. In questa direzione vanno, ad esempio, le riflessioni di Dove e colleghi, i quali prendono in considerazioni vari esempi di autonomia relazionale in ambito medico, tra cui uno legato al nostro. In particolare, essi considerano il caso del ministero della salute britannico, il quale dal 2015 ha favorito la digitalizzazione delle cartelle e dei dati del paziente, divenute accessibili anche dai propri dispositivi personali. Questa strategia, volta all'accrescimento dell'autonomia dei pazienti come diretti possessori dei propri dati e delle proprie informazioni, parte da un'idea di autonomia individualistica, collegata all'indipendenza, come abbiamo precedentemente mostrato. Tuttavia, in molti casi, alcuni individui desiderano o necessitano condividere i propri dati, ma il sistema, progettato al fine di una fruizione individuale (per tutelare la *privacy* del paziente),

[54] A quasi un decennio di distanza dall'articolo di Sherwin che introduceva la proposta di una bioetica relazionale su basi femministe, l'autrice si sofferma sugli effettivi risultati raggiunti in questa direzione: S. Sherwin, K. Stokdale, *Whither Bioethics Now?: The Promise of Relational Theory*, in «IJFAB: International Journal of Feminist Approaches to Bioethics», (2017), 10, 1, pp. 7-29.

[55] C. Gómez-Vírseda, Y. de Maeseneer, C. Gastmans, *Relational autonomy: what does it mean and how is it used in end-of-life care? A systematic review of argument-based ethics literature*, in «BMC Medical Ethics», (2019), 20, 76.

[56] A. Superson, *The Right to Bodily Autonomy and the Abortion Controversy*, in A. Veltman, M. Piper (eds.), *Autonomy, Oppression, and Gender*, Oxford University Press, Oxford 2014, pp. 301-325; A. Ballantyne, *Exploitation in Cross-Border Reproductive Care*, in «IJFAB: International Journal of Feminist Approaches to Bioethics», (2014), 7 (2), pp. 75-99.

sembra ostacolare coloro che necessitano di questo approccio condiviso, senza comprendere che questo coinvolgimento relazionale non è semplicemente un ostacolo all'autonomia del paziente, ma in molti casi una sua completa realizzazione: «For it is often in these messy grey zones of in-betweenness and hybridity where the impact of our decisions on others is not only contemplated but valued, that the truest expression of self-rule is manifest»[57].

Questo caso ci mostra la possibilità di una riunificazione equilibrata dell'idea stessa di autonomia anche nell'ambito della telemedicina. Le indubbie possibilità su questi campi, se sfruttate all'interno di una prospettiva relazionale, permettono di parlare di un accrescimento non solo delle possibilità del paziente, ma delle possibilità della relazione, attraverso la costruzione di valori comuni, fiducia reciproca favorita da un possesso simmetrico delle informazioni, responsabilizzazione del paziente. Le potenzialità aperte dalle TIC nella cura potranno così diventare vettore non solo di trasmissione di dati, ma anche il luogo del coinvolgimento nella creazione di valori[58].

English title: Autonomy in telemedicine. The e-patient between independence and involvement

Abstract

In this article we will analyse the impact of ICT in the field of medicine. In particular, we will focus on patient autonomy (e-patient), which seems to be increasing thanks to telemedicine. In the first part, we will reconstruct the debate on the subject between those who think that telemedicine is an incentive for patient autonomy and those who see it as a tool for control and domination over patients' bodies. In the second part we will explore the concept of autonomy in an executive sense, which seems to be enhanced by telemedicine. In the second part we will explore the concept of autonomy in an executive sense, which seems to be enhanced by telemedicine. After showing the limits of this idea, we will conclude by proposing a more relational per-

[57] E.S. Dove, S.E. Kelly, F. Lucivero, M. Machirori, S. Dheensa, B. Prainsack, *Beyond individualism: Is there a place for relational autonomy in clinical practice and research?*, in «Clinical Ethics», (2017), 12 (3), p. 162.

[58] R. Palumbo, *The Bright Side and the Dark Side of Patient Empowerment. Co-Creation and Co-Destruction of Value in the Healthcare Environment*, Springer, Cham 2017.

spective, which may favour the involvement between doctor and patient and the protection of dependencies and vunerabilities.

Keywords: Telemedicine; relational autonomy; indipendence; vulnerability.

Silvia Dadà
Università di Pisa
silvia.dada@cfs.unipi.it

T

Francesca Marin

Decisioni di fine vita, dipendenza e vulnerabilità

1. *Premessa*

Negli ultimi anni si è sviluppato in Italia un acceso dibattito sul tema del fine vita, con approfondimenti di carattere bioetico e giuridico che hanno riguardato in particolare la pratica del suicidio medicalmente assistito[1]. La discussione sul tema ha preso avvio dalla richiesta di alcuni pazienti di assistenza medica al suicidio a cui hanno fatto seguito delle risposte giurisprudenziali che, oltre ad avere un inevitabile impatto sull'opinione pubblica, hanno sollecitato il Parlamento a intervenire in materia. Malgrado tali sollecitazioni, nel nostro Paese vi è ancora un vuoto legislativo in merito a questa determinata pratica di fine vita. Va altresì segnalato che l'iter del Disegno di Legge "Disposizioni in materia di morte volontaria medicalmente assistita" (DDL n. 2553, approvato dalla Camera dei Deputati il 10 marzo 2022) non si è completato a causa della conclusione della XVIII legislatura della Repubblica Italiana.

Il presente contributo intende evidenziare come le argomentazioni sino-

[1] A dire il vero, di recente nel nostro Paese si è sviluppata una discussione anche in merito all'eutanasia: nel 2021 sono state raccolte ben oltre 1,2 milioni di firme a sostegno della proposta referendaria dell'Associazione "Luca Coscioni", finalizzata – a detta dei promotori – a legalizzare l'eutanasia attraverso l'abrogazione parziale dell'art. 579 del codice penale (omicidio del consenziente). Con sentenza n. 50 del 2022, la Corte Costituzionale ha dichiarato inammissibile il referendum perché, come riporta il Comunicato della Corte stessa emesso il 15 febbraio 2022, «a seguito dell'abrogazione, ancorché parziale, della norma sull'omicidio del consenziente, cui il quesito mira, non sarebbe preservata la tutela minima costituzionalmente necessaria della vita umana, in generale, e con particolare riferimento alle persone deboli e vulnerabili» (https://www.cortecostituzionale.it/documenti/comunicatistampa/CC_CS_20220215193553.pdf).

 DOI: 10.4454/teoria.v42i2.156

ra proposte nel dibattito pubblico incoraggino ad assumere posizioni sempre più permissive. Alla base di tali argomentazioni sembra infatti esservi una visione omogenea delle diverse pratiche di cura di fine vita che condiziona inevitabilmente il processo decisionale. Detto altrimenti, si registra un graduale misconoscimento delle differenze mediche e bioetiche tra i vari trattamenti e procedure, con il conseguente rischio di adottare un approccio riduttivo rispetto alle cosiddette *End-of-life-decisions* (decisioni di fine vita).

Per cogliere questo aspetto, si volgerà innanzitutto lo sguardo all'odierno lessico del fine vita. Nello specifico, la prima parte dello scritto intende evidenziare come la pratica del suicidio medicalmente assistito venga sempre più denominata con una terminologia che ne offusca i tratti peculiari, giungendo così ad equiparare erroneamente tale pratica ad altre procedure assistenziali. Espressioni quali "morire medicalmente assistito" e "aiuto medico a morire" potrebbero infatti alludere non solo all'assistenza medica al suicidio, ma anche ad altre pratiche assistenziali, come ad esempio la sedazione palliativa profonda, che non sono finalizzate a procurare la morte del paziente. Da qui la necessità di fare emergere le differenze mediche e bioetiche tra le varie *End-of-life-decisions*, affinché si riconosca come la decisione di accedere al suicidio medicalmente assistito avvii un percorso diverso rispetto a quello previsto per la sospensione dei trattamenti e l'avvio della sedazione palliativa profonda.

Nella seconda parte di questo scritto, si mostrerà come le suddette differenze non siano state debitamente riconosciute nelle varie fasi del dibattito italiano sul suicidio medicalmente assistito: pur delineando un ambito assai limitato di legittimazione dell'aiuto medico al suicidio, l'ordinanza n. 207/2018 e la sentenza n. 242/2019 della Corte Costituzionale adottano una linea argomentativa che giunge erroneamente ad equiparare pratiche che sono in realtà ben diverse in termini di procedure e obiettivi. Per di più, focalizzando l'attenzione su una delle condizioni dettate dalla stessa Corte, ovverosia la dipendenza del richiedente da trattamenti di sostegno vitale, alcune sentenze giurisprudenziali hanno incluso tra i suddetti trattamenti non solo i supporti tecnologici, ma anche i farmaci e l'assistenza da parte di terzi. Nelle battute finali di questo contributo si intende mostrare le problematicità di una tale prospettiva: essa favorisce una visione omogenea delle diverse pratiche assistenziali, ha un impatto sulla nostra concezione di vulnerabilità e rischia di declinare ogni bisogno di cura nei termini di una dipendenza che comporta necessariamente un vissuto negativo.

2. *Un significativo slittamento terminologico*

Nel DDL n. 2553 il termine "suicidio" compariva solo una volta, precisamente all'interno dell'art. 6, c. 3 dove si affermava che l'esercente la professione sanitaria può sollevare obiezione di coscienza solo rispetto al «compimento delle procedure e delle attività specificamente dirette al suicidio» e non all'«assistenza antecedente l'intervento»[2]. In questo passaggio del testo normativo, il ricorso alla parola "suicidio" appariva doveroso perché specificava la prestazione medica che poteva essere oggetto di obiezione da parte del professionista sanitario. Discutibile era invece l'assenza del suddetto termine in ogni altra parte del DDL, e questo già a partire dal titolo della proposta di legge dove compariva l'espressione "morte volontaria medicalmente assistita".

A dire il vero, la scelta del Legislatore era prova di uno slittamento terminologico che è già in atto da diversi anni nella riflessione bioetica. Nel dibattito bioetico internazionale, la pratica del suicidio medicalmente assistito (*Physician-Assisted-Suicide*) viene sempre più indicata ricorrendo alle locuzioni "*Physician-Assisted-Dying*" e "*Physician Aid-in-Dying*", che si possono rispettivamente tradurre in "morire medicalmente assistito" e "aiuto medico a morire". La tendenza a modificare il lessico per così dire tradizionale non può essere sottovalutata perché giunge ad offrire un'immagine in un certo qual modo sfuocata della pratica del suicidio medicalmente assistito. Le espressioni "morire medicalmente assistito" e "aiuto medico a morire" sembrano infatti svalutare l'elemento dell'intenzionalità e offuscare il contributo offerto dal medico: mentre la dicitura "*Physician-*

[2] Vi è qui una ripresa dell'approccio adottato nella Legge 194/78 sull'interruzione volontaria della gravidanza laddove una parte dell'art. 9 recita: «L'obiezione di coscienza esonera il personale sanitario ed esercente le attività ausiliarie dal compimento delle procedure e delle attività specificamente e necessariamente dirette a determinare l'interruzione della gravidanza, e non dall'assistenza antecedente e conseguente all'intervento». Esula dall'obiettivo del presente contributo esaminare la tematica dell'obiezione di coscienza; per di più, non sarebbe qui possibile proporre un adeguato approfondimento rispetto a questo tema alquanto complesso e delicato. Ci si limita ad evidenziare come la sentenza n. 242/2019 della Corte Costituzionale non contenesse la previsione dell'obiezione di coscienza da parte dei professionisti sanitari, giustificando la scelta in questi termini: «la presente declaratoria di illegittimità costituzionale si limita ad escludere la punibilità dell'aiuto al suicidio nei casi considerati, senza alcun obbligo di procedere a tale aiuto in capo ai medici» e quindi «Resta affidato alla coscienza del singolo medico scegliere se prestarsi, o no, a esaudire la richiesta del malato». Per un'analisi dell'art. 6 del DDL n. 2553, cfr. P. Benciolini, *L'aiuto medico a morire. Un contributo nell'ottica della "medicina legale clinica"*, in «Responsabilità medica», 6 (2022), n. 1, pp. 19-24 (in particolare §4 dal titolo *"Obiezione": per chi?*, pp. 22-24).

Assisted-Suicide" esplicita già da sé l'atto del professionista sanitario che assiste il paziente a commettere il suicidio fornendogli i farmaci letali per l'autosomministrazione, le locuzioni "morire medicalmente assistito" e "aiuto medico a morire" oscurano l'intenzione del medico di aiutare qualcuno a suicidarsi.

Malgrado questi nodi problematici, appare in una certa misura condivisibile la ragione che porta a evitare il termine "suicidio": stando ad esempio ai vissuti esistenziali sinora presi in esame dalla giurisprudenza italiana, colui che richiede un aiuto medico a morire si trova in una situazione estrema e drammatica perché affetto da una patologia irreversibile che cagiona sofferenze fisiche e psicologiche ritenute assolutamente intollerabili nonché capace di prendere decisioni libere e consapevoli ma non in grado di porre fine da solo alla propria esistenza. Dinnanzi a queste storie di vita, il termine "suicidio" risulterebbe inappropriato perché la richiesta di un aiuto medico a morire non sembra dettata da un proposito suicidario, bensì dalla volontà di liberarsi da un corpo che viene oramai vissuto come una prigione[3].

Le considerazioni appena proposte potrebbero allora giustificare la tendenza a sostituire "*suicide*" con "*dying*". Così facendo, non si pone più l'accento su una determinata azione, nella fattispecie il suicidio, bensì sul processo del morire. Questo diverso focus ha l'indebito vantaggio di non circoscrivere il fine vita all'atto che traduce la volontà del paziente conducendolo alla morte, riconoscendo così l'intero processo decisionale nonché i vari fattori contestuali che lo caratterizzano. Ciononostante, se prese alla lettera, le espressioni "morire medicalmente assistito" e "aiuto medico a morire" potrebbero rinviare non solo agli atti finalizzati a procurare la morte del paziente, ma anche a tutte quelle pratiche assistenziali che caratterizzano la cura del fine vita. Detto altrimenti, sostituendo "*suicide*" con "*dying*", non si può escludere che vengano erroneamente declinate nei termini del *Physician-Assisted-Dying* la sospensione dei trattamenti e l'avvio della sedazione palliativa profonda: in effetti, queste procedure possono caratterizzare il processo del morire, ma, come emergerà in seguito, non sono finalizzate a provocare la morte del paziente come invece avviene effettuando il suicidio medicalmente assistito.

[3] È quanto sostenuto anche in alcuni passaggi del parere del Comitato Nazionale per la Bioetica dal titolo *Riflessioni bioetiche sul suicidio medicalmente assistito* (18 luglio 2019, https://bioetica.governo.it/media/4310/vr__p135_2019_parere-suicidio-medicalmente-assistito.pdf, pp. 13 e 24).

3. *Per rimarcare le differenze*

Le considerazioni di carattere terminologico appena proposte evidenziano il rischio di una prospettiva omologante rispetto alle varie pratiche di fine vita. Tale prospettiva ha caratterizzato il recente dibattito italiano sul suicidio medicalmente assistito: Fabio Ridolfi, 46enne di Fermignano (Pesaro-Urbino) immobilizzato da 18 anni a letto a causa di una tetraparesi da rottura dell'arteria basilare, aveva ottenuto l'accesso al suicidio assistito ma, essendo ancora in attesa dall'Asur (Azienda Sanitaria Unica Regionale) Marche di conoscere modalità e farmaco per l'autosomministrazione e ritenendo non più tollerabile il prolungamento delle sue sofferenze, è deceduto il 13 giugno 2022, dopo aver chiesto che venissero sospese idratazione e alimentazione artificiale e avviata la sedazione palliativa profonda. Il diretto interessato ha parlato a riguardo di una scelta obbligata dovuta alle lungaggini burocratiche e i mass media hanno per lo più diffuso la notizia equiparando la decisione iniziale di Fabio di accedere al suicidio medicalmente assistito a quella compiuta a fronte dei ritardi burocratici. Ne danno prova i titoli di alcune testate giornalistiche: *Suicidio assistito, Fabio inizia la sedazione profonda*[4], *Suicidio assistito, è morto Fabio Ridolfi*[5].

In realtà, il suicidio medicalmente assistito e la sedazione palliativa profonda non sono pratiche interscambiabili perché avviano percorsi nettamente differenti e attuano decisioni di fine vita molto diverse fra loro. Tra le due procedure vi sono infatti delle differenze assai rilevanti dovute a ragioni mediche e bioetiche: da un lato, le sostanze letali impiegate nell'assistenza medica al suicidio (come il tiopental sodico) differiscono dai farmaci utilizzati per la sedazione profonda (quali midazolam, diazepam e propofol); dall'altro, la somministrazione di sedativi che conduce al-

[4] https://www.ilrestodelcarlino.it/pesaro/cronaca/non-siate-tristi-fabio-avra-quel-che-voleva-1.7778034. Si veda anche https://www.ansa.it/sito/notizie/topnews/2022/06/13/suicidio-assistito-per-fabio-inizia-sedazione-profonda_a2b5ec6a-e96a-4102-b33e-ebef93185936.html; https://www.rainews.it/articoli/2022/06/suicidio-assistito-inizia-oggi-la-sedazione-profonda-per-fabio-ridolfi-f7707d49-4c1a-43ed-bdda-a2169a23db19.html.

[5] https://www.repubblica.it/cronaca/2022/06/13/news/suicidio_assistito_per_fabio_ridolfi_inizia _la_sedazione_profonda-353698151/. A dire il vero, qualcosa di simile si era già verificato nel nostro Paese qualche anno fa. A inizio gennaio 2018, la stilista Marina Ripa di Meana, che aveva in precedenza espresso la volontà di recarsi in Svizzera per la pratica del suicidio medicalmente assistito, diffondeva in un video messaggio il seguente appello: non è necessario andare oltralpe per porre termine alle proprie sofferenze perché si può percorrere l'equivalente «via italiana delle cure palliative con la sedazione profonda» (cfr. https://www.youtube.com/watch?v=7INNQm_ZTV4).

la perdita dello stato di coscienza è finalizzata ad alleviare le sofferenze in imminenza di morte, e non a procurare il decesso come invece avviene fornendo aiuto all'aspirante suicida. Detto altrimenti, le due pratiche in questione sono ben diverse in termini di procedure e di obiettivi, e quindi non si può parlare, come è avvenuto invece a livello mediatico, di suicidio medicalmente assistito e di sedazione palliativa profonda in maniera contigua o interscambiabile.

Il mancato riconoscimento delle suddette differenze incoraggia una visione omogenea delle varie pratiche di fine vita e incrementa quelle diffidenze che purtroppo ancora oggi persistono nei confronti delle cure palliative in generale e della sedazione palliativa profonda in particolare. Tali diffidenze sono spesso dovute a una scarsa conoscenza della materia (sia a livello di opinione pubblica sia in ambito sanitario), ma anche a una lettura distorta degli obiettivi di questi percorsi di cura, perché si ravvisa erroneamente in essi una modalità per accelerare o procurare la morte del paziente[6]. In realtà, come già detto, la sedazione palliativa profonda è finalizzata ad alleviare le sofferenze in imminenza di morte. Nello specifico, l'obiettivo è quello di ridurre o eliminare i cosiddetti sintomi refrattari, cioè quei sintomi quali irrequietezza psicomotoria, dispnea e angoscia non più controllabili attraverso gli altri trattamenti palliativi. Ora, non vi è dubbio che tale obiettivo possa subire uno stravolgimento nel momento in cui si accosta questa particolare procedura palliativa al suicidio medicalmente assistito.

Purtroppo la risonanza mediatica della storia di Fabio ha dato conferma di quanto appena detto perché la decisione di sospendere i trattamenti di sostegno vitale e di avviare la sedazione palliativa profonda è stata descritta come una «soluzione "di ripiego"» per «anticipare la propria morte in modo medicalmente assistito»[7]. Oltre a dar luogo a un totale travisamento delle finalità della sedazione palliativa profonda, il suddetto messaggio mediatico scredita questa particolare procedura perché la considera una via sostitutiva al suicidio medicalmente assistito nei casi in cui quest'ultimo non si possa attuare per ritardi burocratici o non sia stato concesso a chi ne ha fatto richiesta.

[6] Cfr. L. Orsi, *Differenze e rapporti fra Cure Palliative e Morte Medicalmente Assistita. Per favore, non facciamo confusione!*, in «Responsabilità medica», 6 (2022), n. 1, pp. 49-52.

[7] https://www.fondazioneveronesi.it/magazine/i-blog-della-fondazione/le-ragioni-delletica/suicidio-assistito-e-sedazione-profonda-la-storia-di-mario-e-fabio.

4. *Le argomentazioni della Corte Costituzionale*

Una certa visione omogenea delle pratiche di fine vita fa da sfondo anche ai pronunciamenti della Corte Costituzionale in tema di suicidio medicalmente assistito. Per cogliere questo aspetto, occorre evidenziare le ragioni che hanno condotto la medesima Corte ad intervenire in materia.

Con ordinanza 14 febbraio 2018, la Corte d'Assise di Milano aveva sollevato la questione di legittimità costituzionale dell'art. 580 del codice penale nella parte in cui sanziona in modo indiscriminato le condotte di istigazione e di aiuto al suicidio. In quell'occasione, si analizzava il comportamento di Marco Cappato, che aveva accompagnato Fabiano Antoniani (noto come DJ Fabo) presso l'associazione svizzera *Dignitas* per effettuare il suicidio medicalmente assistito. Per la suddetta Corte, la condotta di Cappato non risultava sanzionabile perché, pur agevolando l'esecuzione del suicidio di Antoniani, non aveva inciso sul processo deliberativo dell'aspirante suicida, cioè non aveva comportato un rafforzamento della decisione suicidiaria. Da qui la sospetta illegittimità dell'art. 580 del codice penale, precisamente nella parte in cui non distingue le condotte di agevolazione da quelle di istigazione al suicidio.

Con l'ordinanza n. 207/2018 e la sentenza n. 242/2019, la Corte Costituzionale ha escluso che «l'incriminazione dell'aiuto al suicidio, ancorché non rafforzativo del proposito della vittima, possa ritenersi di per sé in contrasto con la Costituzione»[8]: tale incriminazione costituisce una forma di tutela del diritto alla vita, soprattutto delle persone più deboli e vulnerabili. Tuttavia, sempre a parere della Corte, l'art. 580 del codice penale non si applica nei casi in cui l'aiuto al suicidio venga fornito a una persona «(a) affetta da una patologia irreversibile e b) fonte di sofferenze fisiche o psicologiche, che trova assolutamente intollerabili, la quale sia (c) tenuta in vita a mezzo di trattamenti di sostegno vitale, ma resti (d) capace di prendere decisioni libere e consapevoli»[9].

Per legittimare l'aiuto al suicidio nelle circostanze che rispondono a queste quattro condizioni, la Corte evidenzia come la libertà di autodeterminazione del malato nella scelta delle terapie debba comprendere anche quelle finalizzate a liberarlo dalle sofferenze. In particolare, tra le procedure che consentono al paziente di porre fine alla propria vita non possono

[8] Sentenza della Corte Costituzionale n. 242/2019, par. 2.2, https://www.cortecostituzionale.it/actionSchedaPronuncia.do?anno=2019&numero=242.

[9] *Ivi*, par. 2.3.

rientrare solo il rifiuto o la rinuncia ai trattamenti di sostegno vitale e il conseguente avvio della sedazione palliativa profonda (come previsto dalla Legge 219/2017, art. 1, c. 5 e art. 2), ma anche l'assistenza medica al suicidio.

Questa linea argomentativa presenta diversi nodi problematici: essa palesa innanzitutto una lettura distorta della Legge 219/2017[10] perché quest'ultima, sulla base del principio di autodeterminazione individuale, riconosce sì il diritto del paziente di rifiutare o di rinunciare ai trattamenti sanitari, ma non afferma la libertà di porre termine alla propria esistenza[11]. Per rivendicare una tale libertà si dovrebbe riconoscere il diritto di morire, che viene però esplicitamente negato dalla Corte Costituzionale: quest'ultima, infatti, ricorda come dall'art. 2 della Costituzione discenda «il dovere dello Stato di tutelare la vita di ogni individuo» e «non quello – diametralmente opposto – di riconoscere all'individuo la possibilità di ottenere dallo Stato o da terzi un aiuto a morire».

Vi è da dire però che, giustificando il suicidio medicalmente assistito per evitare una limitazione della libertà individuale, si è dinnanzi a un riconoscimento, seppure implicito, del diritto di morire. Per di più, in base a questo approccio, si potrebbe legittimare non solo l'assistenza medica al suicidio, ma anche la richiesta di eutanasia da parte di un paziente libero e consapevole. In effetti, per garantire la libera scelta sui modi per porre termine alla propria esistenza, si dovrebbe attuare perfino la decisione di morire espressa da colui che è fisicamente impossibilitato ad autosomministrarsi il farmaco letale[12].

Oltre a incoraggiare posizioni sempre più permissive, le argomentazioni della Corte Costituzionale promuovono una visione omogenea delle diverse

[10] Purtroppo il livello di conoscenza della suddetta Legge è ancora scarso sia in ambito sanitario sia nella discussione pubblica. Diverse lacune si registrano poi rispetto all'applicazione del testo normativo, come dimostra ad esempio il persistente squilibrio tra le regioni italiane riguardo all'erogazione delle cure palliative. A riguardo, si vedano i dati pubblicati a inizio 2018 dall'Huffington Post e relativi all'inchiesta *L'ultima cura non è per tutti* (https://projects.huffingtonpost.it/cure-palliative/) e le conclusioni di un recente studio empirico pubblicate in D. Pietersz, *I risultati della ricerca: Italia*, in L. Gaudino (eds.), *La relazione di cura tra legge e prassi. Un'indagine comparativa tra Italia, Francia, Spagna e Inghilterra*, Pacini Giuridica, Pisa 2021, pp. 135-149.

[11] Recita infatti l'art. 1, c. 6 della L. 219/2017: «Il medico è tenuto a rispettare la volontà espressa dal paziente di rifiutare il trattamento sanitario o di rinunciare al medesimo e, in conseguenza di ciò, è esente da responsabilità civile o penale. Il paziente non può esigere trattamenti sanitari contrari a norme di legge, alla deontologia professionale o alle buone pratiche clinico-assistenziali; a fronte di tali richieste, il medico non ha obblighi professionali».

[12] Cfr. M. Reichlin, *Fondamenti di bioetica*, il Mulino, Bologna 2021, p. 126.

pratiche di fine vita; nello specifico, non si riconosce come il rifiuto o la rinuncia dei trattamenti sanitari e la sedazione palliativa profonda da un lato e l'assistenza medica al suicidio dall'altro avviino percorsi assai diversi in termini di procedure e obiettivi[13]. Attuando queste pratiche, il medico crea una condizione che contribuisce al verificarsi dell'evento fatale, ma nel caso del suicidio medicalmente assistito egli agevola l'esecuzione del proposito di suicidio del paziente tant'è vero che gli fornisce un farmaco letale. L'intervento del professionista sanitario è quindi finalizzato a procurare il decesso dell'assistito, mentre la sedazione palliativa profonda ha l'obiettivo di alleviare le sofferenze in imminenza di morte. Il mancato riconoscimento di queste differenze ha delle ripercussioni sull'ascrizione di responsabilità morale perché porta a ritenere il medico parimenti responsabile della morte del paziente sia qualora fornisca all'assistito una sostanza letale, sia nel caso in cui non avvii o sospenda i trattamenti oppure effettui la sedazione palliativa profonda[14]. Un tale approccio rende altresì accostabili o interscambiabili *End-of-life-decisions* che sono in realtà ben diverse fra loro. Si verificano così delle equiparazioni indebite che distorcono il dibattito bioetico sul fine vita e giungono persino a gettare ombra sulle pratiche palliative che, come già detto, avviano un percorso totalmente diverso da quello dell'assistenza medica al suicidio.

5. *Sull'espressione "trattamenti di sostegno vitale"*

Malgrado le problematicità appena evidenziate, i pronunciamenti della Corte Costituzionale delineano un ambito assai circoscritto di legittimazione dell'aiuto medico al suicidio. Ciononostante, è già in atto una messa in discussione delle condizioni limitative enunciate dalla Corte. Infatti, nella sentenza del 27 luglio 2020, la Corte d'Assise di Massa ha assolto Marco Cappato e Mina Welby dai reati di rafforzamento e agevolazione del suicidio: nell'aprile 2017 gli imputati avevano accompagnato Davide Trentini, paziente affetto da sclerosi multipla, presso la Fondazione svizzera *Lifecircle* per effettuare il suicidio medicalmente assistito. L'assoluzione degli imputati è avvenuta ricorrendo alle argomentazioni della Corte Costituziona-

[13] Cfr. A. Da Re, *La falsa analogia tra rifiuto-rinuncia alle cure e suicidio medicalmente assistito. Riflessioni bioetiche sull'ordinanza della Corte Costituzionale n. 207/2018*, in «Medicina e Morale», 68 (2019), n. 3, pp. 281-295.

[14] Per un approfondimento sul tema, rinvio a F. Marin, *Bioetica di fine vita. La distinzione tra uccidere e lasciar morire*, Orthotes Editrice, Napoli-Salerno 2017, pp. 163-170.

le, ma, diversamente da Antoniani, Trentini non dipendeva da trattamenti di sostegno vitale. Egli assumeva farmaci antispastici e antidolorifici ed era dipendente dalla funzione meccanica manuale evacuativa delle feci. La Corte d'Assise di Massa giustificava però la sua decisione in questi termini: «dipendenza da "trattamenti di sostegno vitale" non significa esclusivamente "dipendenza da una macchina" ma comprende anche la dipendenza da qualsiasi trattamento sanitario senza il quale si verificherebbe la morte del malato […], anche se in tempi non rapidi»[15].

Con la sentenza del 28 aprile 2021, la Corte d'Assise di Genova ha confermato la suddetta conclusione giurisprudenziale nei seguenti termini:

> La malattia gravissima da cui era affetto Trentini Davide non richiedeva il ricorso a macchinari; il trattamento farmacologico era tuttavia per lui essenziale per la sopravvivenza, poiché se non lo avesse assunto si sarebbe fatalmente alterato il delicato equilibrio che gli permetteva di sopravvivere. Anche Trentini dunque viveva una vita artificiale, fonte di insopportabile dolore fine a se stesso, perché la guarigione non sarebbe stata possibile, mentre la malattia sarebbe progredita fino a provocargli la morte in un giorno non definibile, ma certo[16].

Stando a questa tesi, essere sottoposti a trattamenti di sostegno vitale significa essere dipendenti da tutto ciò che è direttamente funzionale alla propria sopravvivenza. Nella categoria 'trattamenti di sostegno vitale' rientrano così dispositivi medici (quali ventilazione, alimentazione e idratazione artificiale), farmaci e persino l'assistenza da parte di terzi.

È questa una lettura estensiva dell'espressione 'dipendenza da trattamenti di sostegno vitale' che è stata recentemente proposta anche dal Comitato Etico Regionale delle Marche (CERM) nel parere relativo al caso del paziente tetraplegico Mario (nome di fantasia), la cui vera identità (Federico Carboni) è stata rivelata dopo la sua morte avvenuta il 16 giugno 2022[17]. Chiamato ad accertare la sussistenza delle quattro condizioni stabilite dalla Corte Costituzionale, il CERM ha affermato che i trattamenti a cui Mario era sottoposto (pacemaker, catetere vescicale permanente ed evacuazioni manuali) potevano essere considerati di sostegno vitale perché la loro sospensione avrebbe provocato complicanze tali da condurlo alla morte.

[15] https://www.biodiritto.org/ocmultibinary/download/3939/46390/4/7ac8ce6da583d7be605b7ffc6bd7772f/file/Sentenza-Massa.pdf, pp. 31-32.

[16] https://www.giurisprudenzapenale.com/wp-content/uploads/2021/07/trentini-corte-assise-appello-genova.pdf, p. 6.

[17] https://www.biodiritto.org/ocmultibinary/download/4162/48892/2/1c70a9bb3b641206a3505c6a356ae66e/file/Estratto-parere.pdf.

In tal modo, a renderci particolarmente dipendenti e quindi assai vulnerabili non sono più solo i dispositivi offerti dalla medicina altamente tecnologizzata, ma anche quelle forme di assistenza che consentono il soddisfacimento dei bisogni vitali. Si guarda così all'intero piano clinico-assistenziale, inserendo nella categoria "trattamenti di sostegno vitale" ogni singolo elemento che lo caratterizza. In altre parole, tale categoria si estende in modo così ampio al punto da includere presidi medici, apparecchi tecnologici, farmaci e interventi assistenziali di competenza infermieristica quali ad esempio tracheoaspirazione, svuotamento di vescica e intestino nonché il trattamento delle piaghe da decubito. Per indicare tutti questi interventi, il linguaggio per così dire tradizionale sembrerebbe lacunoso, tant'è vero che è stata recentemente proposta la dicitura "trattamenti sanitari medico-infermieristici di sostegno vitale"[18].

Ora, la lettura estensiva dell'espressione "trattamenti di sostegno vitale" può accentuare le fonti della vulnerabilità umana dovute ai vissuti di malattia, al deterioramento del corpo e all'avanzamento dell'età. Nello specifico, non si può escludere che tale lettura porti i soggetti particolarmente deboli e vulnerabili a considerarsi ancor più dipendenti dal contesto assistenziale e dalle figure in esso coinvolte. Tali soggetti potrebbero infatti declinare la loro esistenza come una vita artificiale che grava in termini assistenziali ed economici sulla famiglia e sulla società. In una tale cornice, chi avanza la richiesta di suicidio medicalmente assistito potrebbe trovarsi dinnanzi a una scelta obbligata sostenuta dalla seguente motivazione: non voglio più essere di peso a coloro che si prendono cura di me.

Non si può escludere che possano giungere alla medesima conclusione persino coloro che necessitano di un'assistenza medico-specialistica a lungo termine perché affetti ad esempio da malattie croniche, quali diabete, patologie cardiovascolari e malattie respiratorie croniche. In genere, per fronteggiare tali patologie, è necessario un uso integrato di farmaci, presidi medici e dispositivi digitali di automonitoraggio per effettuare in maniera autonoma e pressoché istantanea rilevazioni fisiologiche come l'indice glicemico nel sangue, il battito cardiaco e la pressione arteriosa. Poiché ogni mezzo utilizzato nel piano di cura sembra essere direttamente funzionale

[18] D. Mazzon, *Sul concetto di "trattamenti di sostegno vitale"*, in «Responsabilità medica», 6 (2022), n. 1, pp. 47-48. Per un ridimensionamento dell'espressione "trattamenti di sostegno vitale", cfr. F.M. Zambotto, *Sulla questione della nozione dei* life sustaining treatments, in «Responsabilità medica», 6 (2022), n. 1, pp. 59-60.

alla sopravvivenza del singolo, si potrebbe paradossalmente affermare che tutti i malati cronici sono sottoposti a trattamenti di sostegno vitale. Portando agli estremi quanto sinora detto, risulterebbero un sostegno vitale, e quindi segno del carattere artificiale della vita, anche i trattamenti farmacologici che si effettuano durante un breve decorso di malattia e persino i farmaci utilizzati per curare raffreddore e influenza.

In sintesi, la lettura estensiva dell'espressione "dipendenza da trattamenti di sostegno vitale" presenta dei tratti paradossali e ha delle inevitabili ripercussioni a livello sociale e culturale, perché rischia di declinare ogni bisogno di cura nei termini di una dipendenza che comporta necessariamente un vissuto negativo. In altre parole, il rischio è quello di guardare alla vulnerabilità come a un ostacolo da superare sempre e comunque o addirittura da annullare, vivendo in maniera negativa tutte le possibili ferite dovute al nostro essere fragili e vulnerabili. In realtà, il nostro bisogno di cura è espressione della vulnerabilità umana, cioè di quel tratto peculiare dell'essere umano che lo apre all'ambiente circostante e gli consente l'incontro con l'altro. Letta in questi termini, la vulnerabilità è una dimensione costitutiva del nostro essere, che ci accompagna in ogni momento ed età della nostra vita; come tale essa non si traduce necessariamente in un'esperienza negativa. Essa semmai, pur potendo essere causa di ferite, costituisce un'esposizione radicale verso la realtà tutta e verso gli altri, che richiede cura e attenzione.

English title: End-of-life-decisions, dependence and vulnerability

Abstract

The paper critically analyzes the Italian debate on Physician-Assisted-Suicide (PAS), showing that it is characterized by a misrecognition of the medical and bioethical differences between the various treatments and procedures, with the consequent risk of adopting a reductive approach to the so-called end-of-life decisions. As will be addressed in the first part of the article, the current end-of-life lexicon proves what has just been said: the practice of PAS is increasingly indicated with expressions like "Physician-Assisted-Dying" and "Physician Aid-in-Dying", which could allude not only to assisted suicide but also to other end-of-life practices that are not aimed at the patient's death, such as deep palliative sedation. In this way, different practices in terms of procedures and objectives are mistakenly equated. The

second part of the paper argues that improper undue equations can also be noticed by analyzing both the arguments of the Constitutional Court (provided in Order n. 207/2018 and Judgment n. 242/2019) and the recent attempt to extend the expression "life-sustaining treatments".

Keywords: physician-assisted-suicide; end-of-life-decisions; withholding/withdrawing treatments; deep palliative sedation; life-sustaining treatments.

Francesca Marin
Università di Padova
francesca.marin@unipd.it

T

Federico Zilio

Tough Decisions in Unclear Situations. Dealing with Epistemic and Ethical Uncertainty in Disorders of Consciousness

1. The limbo of disorders of consciousness

Disorders of consciousness (DoCs) such as coma, vegetative state/unresponsive wakefulness syndrome, and minimally conscious state are characterized by the impairment or even complete loss of self-awareness and awareness of the environment. The entire spectrum of disorders of consciousness can be represented as a "grey zone"[1] or a "limbo" characterized by a qualitative continuum (different phenomenal states and dissociations between wakefulness and consciousness) and a quantitative continuum (different state transitions and fluctuations from one degree of consciousness to another). The best known DoC to laypeople is probably coma, represented by the absence of both dimensions (closed eyes, absence of sleep-wake cycle, response to painful stimuli and/or light, and absence of voluntary actions). After a period of generally ten to thirty days, patients who survive the coma can go through different stages of disorder of consciousness. The vegetative state (VS), recently referred to also as unresponsive wakefulness syndrome (UWS), denotes a state of wakefulness (spontaneous opening of the eyes and recovery of the sleep-wake cycle) in the absence of symptoms of awareness (the patient may produce some non-intentional movements and reflexes). The minimally conscious state involves the recovery of some degree of awareness, represented by signs of intention-

[1] A. Owen, *Into the Gray Zone: A Neuroscientist Explores the Border between Life and Death*, Simon&Schuster, New York 2017.

 DOI: 10.4454/teoria.v42i2.157

al behaviour, execution of simple commands, and non-complex gestural or verbal responses (MCS, MCS+, MCS-, depending on the level of recovery). Finally, patients who recover functional communication and/or functional use of objects are defined as emerging from MCS (EMCS), although in some individuals who remain in a confusional state (CS) there may persist impairment of attention, memory, orientation, perception, etc.

2. *Epistemic and ethical uncertainty*

In recent years, bioengineers and neuroscientists have developed several tools to investigate these pathological conditions; in particular, the recent application of machine learning methods, classification technology, and brain-computer interfaces as complementary approaches to the diagnosis of states of consciousness is indicative of new opportunities and challenges[2]. However, these undefined states of consciousness are raising an increasing number of ethical issues. Although neurotechnologies have certainly improved the differentiation of DoC diagnoses, at the same time, they reveal the complexity of such situations, especially regarding clinical decision-making. In this regard, uncertainty about the concrete state of consciousness of unresponsive patients results in a lack of sufficient information to identify the patient's best interest and make appropriate clinical decisions. Moreover, even in cases where communication can be established with people with disorders of consciousness through simple human-to-human interaction (e.g., blinking) or human-to-machine interaction (e.g., brain-computer interface, BCI), it is not clear whether and to what degree a response (perhaps through a BCI) given by a DoC patient with a degraded or fluctuating level of consciousness can be taken into account for making high-stake decisions.

Taken together, these issues highlight a state of "scientific uncertainty" (or, more generally, epistemic uncertainty), defined as «uncertainty about the diagnosis, prognosis, causal explanation of disease, or treatment rec-

[2] D. Sinitsyn et al., *Machine Learning in the Diagnosis of Disorders of Consciousness: Opportunities and Challenges*, in B. Velichkovsky, P.M. Balaban, V.L. Ushakov (eds.), «Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics. Intercognsci 2020. Advances in Intelligent Systems and Computing, vol 1358», Springer, Cham 2021, pp. 729-735. D. Larrivee, *Improving Objective Assessment in Disorders of Consciousness: An Option for Classification Technology?*, in «Clinical Sciences Research and Reports», 1 (2017), pp. 1-4.

ommendations»[3], which consequently produces a state of ethical uncertainty[4]. I will now present some of the main issues that characterise the intrinsic uncertainty in neuroscience of disorders of consciousness.

2.1. The problem of misdiagnosis

The assessment of consciousness is a challenging topic that involves ethical and legal implications, as misdiagnosis can have devastating consequences for the lives of people with DoCs. In fact, misdiagnosis can lead to attitudes of underestimation or overestimation of the patient's level of consciousness arise; while overestimation of consciousness (false positives) can lead to ethical problems related to resource allocation and false hope, underestimation (false negatives) can lead to nefarious ethical consequences, such as suspension of treatments for patients whose lives no longer seem worth living without the consent of the conscious patient him/herself.

The usual assessments based on the clinical consensus of the medical team are not always sufficient to discern different levels of consciousness; indeed, there is still a high rate of diagnostic error based on clinical consensus (~40%, i.e., several patients are considered to be in a vegetative state when instead they preserve some degree of consciousness); consequently, a purely behavioural approach seems insufficient to characterize the conscious state of DoC patients[5]. In this regard, to better interpret the wide spectrum of consciousness in patients with DoCs, recent guidelines suggest advanced neurological investigations (use of mirrors, familiar voices, naturalistic paradigms, etc.) in addition to the standard behavioural scales, supplemented also by repeated neuroimaging examination.

Some neurodiagnostic tools are, for example, positron emission tomography (PET), functional magnetic resonance imaging (fMRI), and electroencephalography (EEG); the latter can also be accompanied by a brain-computer interface (BCI) system to attempt communication with patients. These neurotechnologies are progressively entering neurological diagnostic routines; however, they are not yet homogeneously diffused and / or they are used only in exceptional cases. Neurological data are not easy

[3] L.S.M Johnson, C. Lazaridis, *The Sources of Uncertainty in Disorders of Consciousness*, in «AJOB Neuroscience», 9 (2018), n. 2, pp. 76-82, p. 79.

[4] L.S.M Johnson, *The Ethics of Uncertainty: Entangled Ethical and Epistemic Risks in Disorders of Consciousness*, Oxford University Press, New York 2022.

[5] M.J. Young et al., *The Neuroethics of Disorders of Consciousness: A Brief History of Evolving Ideas*, in «Brain», 144 (2021), n. 11, pp. 3291-3310.

to analyse and interpret (e.g., a single case of neural-cognitive correlation does not necessarily imply consciousness), and an accurate diagnosis requires several tests applied at different times to avoid false positives (i.e., the subject seems conscious when s/he is not actually conscious) or false negatives (i.e., the subject seems unconscious when s/he actually is)[6]. Furthermore, even the most sophisticated devices should not be considered infallible; current neurodiagnostic tools appear to be accompanied by an intrinsic risk of ambiguity and uncertainty with respect to diagnosis and prognosis[7]. This is probably due to the incomplete development of the technology, which is still in the research and development phase; nevertheless, the neurodiagnostic tool is essentially based on the formulation of indirect and inductive inferences from the neuronal to the conscious activity (when mental state x is engaged, then the neural state y is active; neural state y is active; therefore the mental state x is engaged).

This "reverse inference" is widely used as a good probabilistic tool in cognitive sciences due to its predictive power[8], however it implies a logical fallacy, that is, the so-called "affirming the consequent" (if p then q, q is true; therefore, p is true); therefore, it must be carefully considered, as the presence of a neuronal-cognitive correlation does not necessarily imply a neuronal-phenomenal inference[9]. For this reason, the reverse inference process implied in neurological evaluation is often supported by other tests and should be considered as an ancillary and complementary tool rather than a substitutive one for any clinical and bioethical decision[10].

2.2. The reception of neurodiagnostic information among relatives and caregivers

So far, we have considered some of the problems related to the assessment of disorders of consciousness, not only on the technical and diagnos-

[6] D. Cruse et al., *Detecting Awareness in the Vegetative State: Electroencephalographic Evidence for Attempted Movements to Command*, in «PLoS ONE», 7 (2021), n. 11.

[7] L.S.M Johnson, C. Lazaridis, *art. cit.*

[8] M. Nathan, G. Del Pinal, *The Future of Cognitive Neuroscience? Reverse Inference in Focus*, in «Philosophy Compass», 12 (2017), n. 7.

[9] G. Northoff, *Does Task-Evoked Activity Entail Consciousness in Vegetative State? 'Neuronal-Phenomenal Inference' versus 'Neuronal-Phenomenal Dissociation'*, in M. Farisco, K. Evers (eds.), *Neurotechnology and Direct Brain Communication*, Routledge, New York 2016, pp. 104-116.

[10] A. Peterson et al., *Risk, Diagnostic Error, and the Clinical Science of Consciousness*, in «NeuroImage: Clinical», 7 (2015), pp. 588-597.

tic level, but also on the epistemic and ethical level. Moving now outside the clinical sphere, an additional problem emerges when information from neurological analysis is communicated to laypeople, such as family members and caregivers. Physician-patient-family communication is intrinsically characterized by an epistemic asymmetry, that is, a state of disparity in skills and knowledge; indeed, on the one hand, the physician has much more technical expertise than both the patient and family members, on the other hand, sometimes patients possess privileged subjective knowledge about their illness experience to which the physician cannot access. In this sense, good medical communication must try to prevent this asymmetry from increasing – although it is impossible to remove it – and from producing instances of epistemic injustice, for example, when subjective patient reports or family testimony are underestimated or not considered in formulating diagnosis and prognosis.

In the context of disorders of consciousness, this situation becomes even more complex, as the asymmetry of expertise between clinicians and laypeople increases for two reasons. First, the uncertainty and ambiguity of some neurodiagnostic findings do not help effective communication; second, if the patient seems completely unresponsive (e.g., UWS), no other information about the state of consciousness can be relied upon, making communication with family members even more difficult. Some recent studies have pioneeringly investigated the reception of neurodiagnostic data among family members and caregivers of patients with DoCs. This kind of study can be extremely important to understand the role of neurodiagnostic tools in caregiver attitudes (but also in healthcare professionals) and how they influence end-of-life decisions regarding patients with DoCs.

Schembs and colleagues recently investigated by semi-structured interviews the interpretations, attitudes, and opinions of a group of patients' next of kin (seven) regarding the EEG examination[11]. They found that caregivers tend to adapt neurodiagnostic findings to their belief system, as the preservation of hope is essential to maintaining their ability to care. Therefore, an unfavorable evaluation implied questioning the validity of this type of results, while a positive evaluation allowed us to confirm optimism towards the recovery of their loved one. In particular, they specifically report on some parts of the interviews. Peterson and colleagues also interviewed

[11] L. Schembs et al., *How Does Functional Neurodiagnostics Inform Surrogate Decision-Making for Patients with Disorders of Consciousness? A Qualitative Interview Study with Patients' Next of Kin*, in «Neuroethics», 14 (2021), n. 3, pp. 327-346.

some caregivers (twenty) of patients with DoCs regarding their reactions to and understanding of the EEG evaluation[12]. The results show an overall understanding of the meaning of these data, but with various reactions (acceptance, rejection, emotional exhaustion, disagreement, etc.). At the same time, the authors highlight the risk of misinterpreting the data and the degree of certainty along with strong expectations about diagnosis or prognostic value, either by overestimation (false positive) or underestimation (false negative).

Overall, these and other studies[13] have shown instances of cognitive dissonance and lack of realization in caregivers and next of kin with respect to patient clinical situations. Less attention (it was not the main point of the studies) has been devoted to investiganting how neurodiagnostic information was provided by healthcare professionals and what epistemic status these data have. For example, it is not explained how neurological information has been provided to family members or whether doctors have followed a specific communication protocol that is identical for all families (adapting the content according to the case). Although it may be true that the next of kin of patients with DoCs does not «share the assumption that an 'objective assessment' of consciousness is possible or valid»[14], but it is also true that the neurodiagnostic assessment of consciousness currently has several problems that prevent it from being defined as "objective". Therefore, it would be a mistake to interpret this dissonance between neurodiagnostic data and the reception of relatives only as a problem of one side of the communication (family and caregivers). In this regard, it would be important to analyze not only the ability of laypeople to understand neuroinformation, but also if and how the diagnostic and prognostic uncertainty of disorders of consciousness is communicated.

[12] A. Peterson et al., *Caregiver Reactions to Neuroimaging Evidence of Covert Consciousness in Patients with Severe Brain Injury: A Qualitative Interview Study*, in «BMC Medical Ethics», 22 (2021), n. 105.

[13] L.M. Andersen, H.B. Boelsbjerg, M.T. Høybye, *Disorders of Consciousness: An Embedded Ethnographic Approach to Uncovering the Specific Influence of Functional Neurodiagnostics of Consciousness in Surrogate Decision Making*, in «Neuroethics», 14 (2021), n. 3, pp. 351-356. A. Peterson, *How Will Families React to Evidence of Covert Consciousness in Brain-Injured Patients?*, in «Neuroethics», 14 (2021), n. 3, pp. 347-350.

[14] L. Schembs et al., *art. cit.*, p. 339.

2.3. The performativity of clinical language and the concept of vegetative state

The methodological and epistemic uncertainty of the neurodiagnostic data is not the only factor that makes it difficult to understand (before) and communicate (after) the clinical situation and make decisions. Even the use of a certain kind of clinical language can have unintended consequences in medical communication, as well as in the consideration of disorders of consciousness. To address this point, it is necessary to consider the difference between constative utterances and performative utterances[15]. A constative utterance is truth evaluable, for example, a statement that intends to describe the state of some portion of the world (e.g., "this chair is blue"); instead, a performative utterance does not describe or report anything, nor are they true or false; rather, it performs a certain kind of action (e.g., "I promise I will pay for this chair"). Medical language is highly performative as it can alter the reality of the patient[16]. Even the formulation of the diagnosis, that is, the act of identifying a disease or syndrome from signs and symptoms, is not a purely descriptive act, as in specific cases it can change the socio-ontological status of the patient, influencing possible ethical choices towards the patient. Perhaps the clearest example of this performative effect is the diagnosis of death, in which the state of death is defined through a series of clinical criteria (e.g., flat encephalogram), but it is the act of declaration by the physician that makes it real and establishes a concrete spatio-temporal dimension. Or, as argued by Havi Carel, diagnoses of certain degenerative diseases or cancers involve a global transformation of the subject's existence (e.g., loss of opportunities, possibilities, openness to the future, agency and subjectivity, wholeness, certainty and control)[17], even if the disease was already present before the act of communicating the diagnosis.

In the case of disorder of consciousness, the power of performative medical language is particularly relevant. The concept of a vegetative state is a crucial example of this problem, as it highlights a strong indirect performative significance. Initially, Bryan Jennett and Fred Plum proposed the diagnosis of "persistent vegetative state"[18] to indicate a persistent state of

[15] J.L. Austin, *How to Do Things with Words*. Clarendon Press, Oxford 1962.

[16] E. Lalumera, *Etica della comunicazione sanitaria*, Il Mulino, Bologna 2022.

[17] H. Carel, *Phenomenology of Illness*, Oxford University Press, Oxford 2016.

[18] B. Jennett, F. Plum, *Persistent Vegetative State after Brain Damage. A Syndrome in Search of a Name*, in «Lancet», 1 (1972), n. 7753, pp. 734-737.

wakeful unresponsiveness where the vegetative nervous system (the sleep-wake cycle and autonomic functions) remains intact (something similar to the Aristotelian idea of a vegetative soul[19]). In 1994, the Multi-Society Task Force on PVS declared that a vegetative state can be judged "permanent" (or irreversible) twelve months after a traumatic injury and three months after in patients with non-traumatic aetiology[20].

Now, the truth or falsity of this description of permanence (currently disproved by the number of recoveries even after twelve months) is not the real problem, but the performative consequence of that diagnosis is. First, even though it is not in the original idea of the term, "vegetative state" has acquired a pejorative connotation over time with dehumanizing connotations (referring to a patient as if he or she were a "vegetable")[21]. In 2010, a more neutral terminology, "unresponsive wakefulness syndrome", was proposed[22] as it better recognizes diagnostic uncertainty[23]. Second, the terms "permanent" and "irreversible", unlike the original and more prudent term "persistent", are not just descriptive, as they establish the absence of any recovery capacity. For this reason, recent guidelines suggest that the term "permanent" should be replaced by the term "chronic" to indicate the stability of the condition[24]. In other words, in order to achieve adequate communication and to prevent clinical language

[19] Z.M. Adams, J.J. Fins, *The Historical Origins of the Vegetative State: Received Wisdom and the Utility of the Text*, in «Journal of the History of the Neurosciences», 26 (2017), n. 2, pp. 140-153.

[20] Multi-Society Task Force on PVS, *Medical Aspects of the Persistent Vegetative State*, in «New England Journal of Medicine», 330 (1994), n. 22, pp. 1572-1579.

[21] C. Lazaridis, *Withdrawal of Life-Sustaining Treatments in Perceived Devastating Brain Injury: The Key Role of Uncertainty*, in «Neurocritical Care», 30 (2019), n. 1, pp. 33-41. There is a long series of other terminologies that indicate – more or less voluntary – processes of dehumanisation of the patient with a disorder of consciousness: "s/he is a corpse with a beating heart", "s/he is a piece of meat or an empty shell", "pulling the plug", "it is a fate worse than death", etc. All this highlights that the treatment of patients with DoCs does not depend exclusively on technical-methodological issues concerning diagnosis, prognosis, and rehabilitation, but also depends on the ethical-ontological background that underpins the clinical attitude towards patients. Cfr. F. Zilio, *Personhood and Care in Disorders of Consciousness. An Ontological, Patient-Centred Perspective*, in «Medicina e Morale», 69 (2020), n. 3, pp. 327-346.

[22] S. Laureys et al., *Unresponsive Wakefulness Syndrome: A New Name for the Vegetative State or Apallic Syndrome*, in «BMC Medicine», 8 (2010), n. 68.

[23] However, being limited only to behavioural description, this terminology remains "agnostic" regarding consciousness and, therefore, although it avoids diagnostic error about consciousness, it is not concretely informative alone for clinical decision-making. Cfr. L.S.M Johnson, *op. cit.*, pp. 21-23.

[24] J.T. Giacino et al., *Practice Guideline Update Recommendations Summary: Disorders of Consciousness*, in «Neurology», 91 (2018), n. 10, pp. 450-460.

itself from negatively influencing attitudes towards treatment, greater nosological humility would be required, i.e., recognising the current lack of knowledge about disorders of consciousness[25], the fluidity and variability of such clinical conditions, and consequently the need for conceptual and temporal prudence regarding diagnosis and prognosis (e.g., "persistent" instead of "permanent").

Additionally, the way clinical information is presented can influence decision-making processes toward a specific choice (e.g., withdrawing, withholding, or continuing life-support treatments) without actually denying any option. In this sense, the above-mentioned epistemic asymmetry could turn into 'epistemic manipulation' where information is presented (or intentionally left out) for socioeconomic factors or to promote something useful to the hospital (e.g., allocation of healthcare resources), rather than to the person with DoCs *per se*[26]. For example, some may focus the communication on the severity of brain injury and the low probability of recovery during the acute phase, when it is still very difficult to make an accurate diagnosis and prognosis, in order to (more or less consciously) suggest hasty end-of-life decisions, such as withdrawing life-supporting treatments, which could particularly influence the decisions of families in economic poverty and lack of health insurance.

Together, the misconception of the vegetative state, along with the clinical language of performative use and the risk of epistemic manipulation, has major implications for family counseling, decision-making, and ethics of the field. Although they may appear as a mere exposition of descriptive contents, diagnoses and prognoses regarding disorders of consciousness can lead to biased clinical attitudes and decisions that limit the possibility of patient recovery. Indeed, what Joseph Fins has called "therapeutic nihilism" and "prognostic pessimism"[27] can also depend on conceptual ambiguities and the underlying prescriptive values of certain clinical categorizations. In this sense, patients with a bad diagnosis may have a lower chance of recovery, not just due to the bad outcome of the clinical analysis *per se*, but because the prediction of mortality and the lack of neurorehabilitation programs could prompt premature practices of withdrawing or

[25] J.J. Fins, *Syndromes in Search of a Name: Disorders of Consciousness, Neuroethics, and Nosological Humility*, in M.D. Lockshin, M.K. Crow, M. Barbhaiya (eds.), *Diagnoses Without Names: Challenges for Medical Care, Research, and Policy*, Springer, Cham 2022, pp. 163-175.

[26] L.S.M Johnson, *op. cit*.

[27] J.J. Fins, *Rights Come to Mind: Brain Injury, Ethics, and the Struggle for Consciousness*, Cambridge University Press, New York 2015.

withholding treatments. Therefore, we could speak of "self-fulfilling negative prognoses" (from the self-fulfilling prophecy bias[28]), that is, a vicious circle in which these negative expectations influence clinical choices and outcomes[29].

3. *Clinical decision-making for people with DoCs*

As presented above, disorders of consciousness present several epistemic and methodological issues, which, in turn, generate a number of issues on the ethical and clinical levels. I have discussed some of the problems related to the intrinsic uncertainty of disorders of consciousness: diagnostic error, prognostic uncertainty, communication with family and caregivers, and the performative value of clinical language. All this particularly affects clinical decision-making processes, as it prevents the formulation of a proper balance between scientific evidence, known best practices, knowledge of the clinical situation of the individual case, and physician-patient-family communication. In fact, scientific knowledge on disorders of consciousness and neurodiagnostic technologies, despite important recent steps, is still not as developed and spread as in other clinical areas, and clinical guidelines often suggest a cautious attitude (in both clinical practice and communication) due to the above-mentioned state of uncertainty[30].

Additionally, uncertainty about the patient's state of consciousness also complicates the surrogate decision-making process. The surrogate decision-making process is initiated when a person is unable to make decisions about personal health care (i.e., incompetence); in that case, other legal instruments or persons provide in decision making: first of all, any written advance healthcare directives or any trustees/surrogates/attorneys who should interpret the patient's current wishes according to his past actions and decisions are taken into account. In the absence of these (advanced directives, surrogates, knowledge about past wishes), it becomes necessary to determine the best interest for the patient.

[28] M. Mertens et al., *Can We Learn from Hidden Mistakes? Self-Fulfilling Prophecy and Responsible Neuroprognostic Innovation*, in «Journal of Medical Ethics», (2021), pp. 1-7.

[29] F. Zilio, *art. cit.*

[30] D. Kondziella et al., *European Academy of Neurology Guideline on the Diagnosis of Coma and Other Disorders of Consciousness*, in «European Journal of Neurology», 27 (2020), n. 5, pp. 741-756. J.T. Giacino et al., *art. cit*.

Two problems can be highlighted here with respect to DoCs. First, there is often no certainty about the degree of consciousness of the patient, and this compromises the effectiveness of advanced directives or surrogates because, if there is a possibility that the patient is conscious, it is important to respect her/his autonomy first. Second, even if one recognises that the patient has a certain level of consciousness (e.g., MCS or CMD), the criteria are not clear to include such a patient in supported decision making for important medical decisions[31]. Furthermore, even if one could communicate with a patient with covert consciousness through the use of BCI, it is first to understand whether and what ethical and legal value a response or message displayed through a computer that decodes and classifies brain states has, particularly when limited to closed questions with "yes / no" answers[32].

In general, several issues hinder the formulation of a classic clinical decision-making process in the field of disorders of consciousness, not because the person is unconscious or incompetent (this is already the case in several other pathologies), but because of the epistemic uncertainty about consciousness that consequently implies ethical uncertainty. How do we overcome this impasse? L. Syd M Johnson proposes an inductive balance (instead of deductive, given the intrinsic uncertainty) between two types of risk: epistemic risk, i.e., the risk of being wrong in accepting an incorrect hypothesis, and ethical risk, i.e., the ethical consequences of being wrong. These two types of risk should mutually constrain each other through two principles of inductive risk. The first principle of inductive risk (taken from Richard Rudner) says that «the tolerable level of epistemic risk […] should be limited by the ethical risk of being wrong»[33]. Applying this principle to DoCs, given the high level of ethical risk (e.g., risk of undertreatment, self-fulfilling prognosis, unwanted death), a low level of epistemic risk should be allowed, which unfortunately cannot yet be guaranteed.

[31] A. Peterson, K. Mintz, A.M. Owen, *Unlocking the Voices of Patients with Severe Brain Injury*, in «Neuroethics», 15 (2022), n. 9.

[32] M.N. Abbott, S.L. Peck, *Emerging Ethical Issues Related to the Use of Brain-Computer Interfaces for Patients with Total Locked-in Syndrome*, in «Neuroethics», 10 (2017), n. 2, pp. 235-242. W. Glannon, *Communication with Brain–Computer Interfaces in Medical Decision-Making*, in I. Opris, M.A. Lebedev, M.F. Casanova (eds.), *Modern Approaches to Augmentation of Brain Function*, Springer, Cham 2021, pp. 141-161.

[33] L.S.M Johnson, *op. cit*., p. 87.

Johnson thus proposes to couple this principle with a "second principle of inductive risk" that determines the level of ethical risk by the level of epistemic risk[34]. There are different levels of ethical risks in the medical field that are related to the potential consequences of continuing, limiting, withdrawing or withholding therapeutic treatments, such as minor discomfort, considerable side effects, or even death. Applying this principle to DoCs, given the current high diagnostic error and prognostic uncertainty, the ethical risk should be limited; in other words, in situations where there is diagnostic uncertainty about the state of consciousness and prognostic uncertainty about the chances of survival and recovery, high-stakes decisions (i.e., decisions where the risk-to-benefit ratio is substantially worse than alternatives)[35], should be avoided. To give an example, prompt decisions to withhold or withdraw treatments, donate organs, and recommend deep palliation or "do not resuscitate" orders in acute brain injury should be considered premature and extremely risky from an ethical point of view, given the high epistemic uncertainty regarding acute coma (see also the therapeutic nihilism and prognostic pessimism mentioned above)[36].

Another example can be used with respect to the evaluation of DoC patients (e.g., MCS) based on their ability to use a brain-computer interface. Given the experimental stage of many BCIs, while a high BCI performance indicates a good level of consciousness, a low or absent BCI performance does not necessarily imply low or absent consciousness (denying the antecedent fallacy/inverse error); indeed, the subject may not want to answer, questions may be misunderstood, or the BCI device might have low precision. Therefore, BCI *per se* does not provide a reliable marker for assessing consciousness and, consequently, influencing clinical decision-making.

[34] L.S.M Johnson, *op. cit.*, p. 91.

[35] A. Peterson, K. Mintz, A.M. Owen, *art. cit.*, p. 9.

[36] J.J. Fins, *op. cit*. In this respect, many authors criticise the high number of hospitality deaths (~70%) due to hasty withdrawals (within very few days of brain injury) of treatments in neurointensive care (acute coma). Cfr. B. Edlow, J.J. Fins, *Assessment of Covert Consciousness in the Intensive Care Unit: Clinical and Ethical Considerations*, in «The Journal of Head Trauma Rehabilitation», 33 (2018), n. 6, pp. 424-434. L.S.M. Johnson, *op. cit.*, pp. 98-99.

4. *Conclusions*

The limbo of disorders of consciousness is characterised by an inherent uncertainty involving both technological-methodological factors (neuroimaging), conceptual and linguistic factors (clinical communication and terminology), and ethical factors (nihilistic and pessimistic attitudes on diagnosis and prognosis). This epistemic and ethical uncertainty significantly affects clinical decisions for patients with DoCs. Consequently, greater epistemic humility and recognition of such uncertainty could improve clinical and ethical attitudes, avoiding hasty end-of-life decisions and cases of misinterpretation and manipulation in physician-family communication[37].

Abstract

Disorders of consciousness (DoC) are characterized by impaired or complete loss of self-awareness and awareness of the environment. It is not easy to assess the level of consciousness of people with DoCs; indeed, there may be cases of covert awareness, that is, people who manifest complete behavioural unresponsiveness but preserve some degree of consciousness. This makes the search for neuronal markers of consciousness in subjects with DoC quite urgent, and the improvement and dissemination of innovative neuroimaging technologies a moral imperative. Neuroethics, considered here as a special branch of clinical ethics, should deal with the ethical implications of these neurotechnologies and the intrinsic uncertainty of diagnosis and prognosis about disorders of consciousness, with a focus on how these issues affect clinical decision-making. First, I will present some epistemic and methodological issues that characterise the disorders of consciousness: diagnostic error, prognostic uncertainty, communication with family and caregivers, and the performative value of clinical language. The epistemic uncertainty emerging from these problems is deeply intertwined with ethical uncertainty, especially when dealing with clinical decisions that may lead to the death of persons whose states of consciousness (and wishes) are not entirely clear. I will suggest the need for epistemic and ethical prudence, through the formulation of a balance between the two principles of inductive risk as proposed by L. Syd M Johnson. Consequently, recognition of intrinsic uncertainty in the field

[37] I would like to thank the Centro Universitario Cattolico (CUC) for supporting my project on pluralist epistemology in neuroscience. I particularly thank the CUC director, Prof. Ernesto Diaco, and my fellow researchers.

of disorders of consciousness could improve clinical and ethical attitudes, avoiding hasty end-of-life decisions and cases of misinterpretation and manipulation in physician-family communication.

Keywords: epistemic uncertainty; ethical uncertainty; disorders of consciousness; consciousness; clinical decision-making.

Federico Zilio
Università di Padova
federico.zilio@unipd.it

T

Mario De Caro, Massimo Marraffa

Consciousness and responsibility

There undoubtedly is a strong tension between cognitive science and folk psychology. On the one hand, some cognitive scientists drastically downplay introspection, and with that they cast radical doubt on the ordinary conception of ourselves as conscious agents: except for perceptual data, they claim, conscious mental states are illusionary. On the other hand, naive ethics – as reconstructed by experimental philosophy – looks to consciousness as the fundamental basis for attributing responsibility: agents are responsible for an action if it reflects a conscious deliberation on their part.

After exposing this disagreement, we will advocate adopting an intermediate position between traditional philosophers, who continues to ascribe primacy to consciousness in action in spite of the data emerging from the mind-brain sciences, and scientists (or empirically oriented philosophers) who, overgeneralizing from specific cases, claim that all conscious mental states are epiphenomenal. An example of this intermediate position can be gleaned from some authors (Levy, 2014; Carruthers, 2015a; Carruthers and King, 2022), who convincingly argue that cognitive neuroscience, rather than proving the epiphenomenalism of consciousness, allows for a finer-grained articulation of the dialectic between unconscious processing and conscious reflection.

1. *Introspection as theorizing*

In the last decades, a psychological tradition of research has developed experimentally the Freudian hypothesis of our propensity for self-decep-

 DOI: 10.4454/teoria.v42i2.158

tion, i.e. a tendency to fabricate "convenient" explanations of our conduct. This has happened especially in social and group psychology, where experimental designs have been devised with participants that have no direct introspective access to their real motivations (i.e., the true causes) of their conduct in the experiment; unaware of these motivations, they nevertheless fabricate a posteriori – on the basis of socially shared explanatory theories or idiosyncratic theorizing – reasonable but imaginary explanations of their own conduct (a form of nonclinical "confabulation"). Here, unconscious everyday mechanisms of self-deception have been shown to be more pervasive, articulate, varied, and profound than Freud thought (cf. Wegner, 2002; Wilson, 2002; Johansson *et al*. 2013).

Consider a classic case of confabulation of intentions. In a study by Wegner and Wheatley (1999), a participant P and an experimenter's accomplice rested their fingers on a tablet mounted on a computer mouse, moving a cursor on a screen where about fifty small objects appeared. Subjects heard words in headphones and had to keep moving the mouse until the stop signal came (about every 30 sec). P was induced to mistakenly believe that she was the one who made the decision to stop the cursor movement; this was achieved by having her listen to the name of one of the objects that appeared on the screen just before the accomplice locked the cursor next to the image of the named object. In addition to the confabulation of decisions, there were fluctuations in the perception of intentionality depending on when P heard the word.

These kinds of experimental data (which could be multiplied at will) are the source of theories in which "introspection" is judged to be a misnomer for an interpretive process, that is, a process that makes use of information concerning states of affairs external to the mind (the agent's manifest behavior and/or the situation in which that behavior takes place) in order to *theorize* about the causal etiology of one's own and others' behavior. This is the theory of self-knowledge that establishes a *Self/Other Parity* (cf. Schwitzgebel, 2019, §2.1), whose historical referent is Ryle (1948)[1].

In this view, introspective consciousness is redefined as the ability to ex post facto remotivate one's actions, that is, the ability to continuously "approve" what one is doing. The agent is no longer – as a stereotype

[1] "The sort of things I can find out about myself are the same as the sorts of things I can find out about other people, and the methods of finding them out are much the same […I]n principle, as distinct from practice, John Doe's ways of finding out about John Doe are the same as John Doe's ways of finding out about Richard Roe" (Ryle, 2009, p. 139).

implicit in the naive way of examining animal-type living systems would have it – a primarily quiescent organism, which 'then' moves, each time for a given purpose; it is rather a primarily self-propelled structure. So, one can never really tell when an action begins nor when an identifiable plan of behavior directed toward an end arises. It is more accurate to say that we have always been immersed in a system of behavioral patterns (or, more precisely, cognitive-motor patterns) that we have begun to articulate since we exist as individuals, and that we relentlessly modify and repurpose according to circumstances and the stimuli that modulate them. And immersed in this flow of actions, we sometimes say and tell ourselves "This is just the thing I want to do", or "What I did is the thing I really wanted to do", or again "This thought is just what I feel like thinking". In this view, what characterizes "voluntary" human action is not so much the presence of anticipatory mental events, but (i) the fact that we are not surprised that we have performed that action[2]; and (ii) that we then explain it. As Anscombe (1957) noted, it is incorrect to assume that we know what our intentions are; what is to correct to say, rather, is that we can tell what our intentions are.

2. *Do conscious thoughts exist?*

Having reached this point, it is important to note that no serious scholar has endorsed a *purely* self/other parity view. Nisbett and Wilson (1977), for example, distinguished between "cognitive processes" (i.e., the causal processes underlying judgments, decisions, emotions, and feelings) and mental "content" (the judgments, decisions, emotions, and feelings themselves). This private content can be accessed directly, resulting in knowledge endowed with "almost complete certainty". And Ryle (1949) himself, when he stresses the importance of outward behavior in our mentalistic self-attribution practices, acknowledges the presence of "twinges", "thrills", "tickles", and even "silent soliloquies", which we know of in our own case and that do not appear to be detectable by observing outward behavior. However, since none of these scholars has offered any hypothesis about the mechanisms of this apparently more direct self-knowledge, their

[2] "[D]ie willkürliche Bewegung sei durch die Abwesenheit des Staunens charakterisiert" ("Voluntary movement is marked by the absence of surprise") (Wittgenstein, 1953, Engl. transl. 1986, §628).

theory is *incomplete* (Schwitzgebel, 2019, §2.1). With this in mind, it is of the utmost importance to turn attention to Peter Carruthers's (2011, 2015a, 2019) enhanced version of the self/other parity view.

Carruthers's theory of introspective self-knowledge rests on the validity of a global workspace account of the conscious accessibility of our perceptual experiences, first postulated by Baars (1988) and widely confirmed since (Dehaene, 2014). In particular, analyses of functional connectivity patterns in the human brain have shown which sort of neural architecture is necessary to realize the main elements of a global broadcasting account. Specifically, these studies show the existence of two main neurocomputational spaces within the brain, each characterized by a distinct pattern of connectivity.

The first space is a processing network, composed of a set of parallel, distributed, and functionally specialized processors or modular subsystems subsumed by topologically distinct cortical domains with highly specific local or medium-range connections that encapsulate information relevant to its function. These subsystems compete with each other to access the Global Neuronal Workspace (GNW), which is implemented by long-range cortico-cortical connections, mostly originating from the pyramidal cells of layers 2 and 3 that are particularly dense in prefrontal, parieto-temporal and cingulate associative cortices, together with their thalamo-cortical loops.

The global broadcasting architecture provides Carruthers with a framework within which it can be argued that *occurrent thoughts* are always unconscious and direct the stream of consciousness and reflection from behind the scenes. The expression "occurrent thoughts" refers to propositional attitude events (such as "judging something to be the case", "deciding to do something", or "actively intending to do something") that are *episodic* rather than persisting, and have a *non-sensory format* (they are "amodal"). Carruthers claims that only sensory or sensory-involving states can participate in consciousness (and, a fortiori, reflection), while amodal propositional attitudes operate unconsciously in the background. This thesis is argued in two steps.

First, according to Carruthers occurrent thoughts cannot be *first-order* access-conscious. The global broadcasting architecture affords to explain the conscious accessibility of our sensory or sensory-involving states. When one of the functionally specialized processors accesses the global workspace, its outputs (i.e., sensory information including perceptions of the world, the deliverances of somatosensory systems, imagery, and inner

speech) are broadcast to an array of executive, conceptual, and affective "consumer" systems. These systems process ("consume") sensory information according to their various specialisms – e.g., drawing inferences, forming memories, producing emotional responses, forming judgments, planning and making decisions, and verbally reporting. By contrast, thoughts – that is, the outputs of the consumer systems – are not capable of being globally broadcast. The reason is that the mechanism by which a state is broadcast is *top-down attention*; and in reviewing the literature on attention in cognitive neuroscience, Carruthers finds that "attention itself has an exclusively sensory focus", primarily targeting "midlevel sensory areas" (2015a, pp. 91-2). (More precisely, a top-down attentional network links the dorsolateral prefrontal cortex, the frontal eye-fields, and the intraparietal sulcus. The "business end" of the system is the latter, which projects both boosting and suppressing signals to targeted areas of mid-level sensory cortices.) Hence the anticipated conclusion: only states with a sensory-based format are capable of becoming first-order access-conscious.

Let us come to Carruthers's second argumentative step. According to him, occurrent thoughts cannot be *higher-order* access-conscious either. It seems obvious that thoughts are available in a way that enables us to know of their occurrence without requiring self-interpretation, of the sort that makes us aware of the thoughts of other people. The global broadcasting architecture, however, allows Carruthers (2011) to develop a robust version of the self/other parity account of self-knowledge. According to Carruthers' version of the self-other parity theory of the nature and sources of self-knowledge (the so-called "Interpretive Sensory-Access", ISA), we can have non-interpretive access only to our sensory or sensory-involving states; all knowledge of our own occurrent thoughts is instead a matter of *interpretation*.

Among the consumer systems that form judgments (i.e. events of belief-formation), a "mindreading system" exists that is a multi-componential faculty that exploits a corpus of folk-psychological theoretical knowledge in order to generate metarepresentational beliefs about the mental states of others and of oneself. This faculty, Carruthers argues, was originally designed for "reading" other minds; only at a later stage the ancestral mindreaders started to apply this skill to themselves, forming beliefs about their own mental states as they did about other people's. Since the mindreading system evolved for understanding other people, it is *outward looking*: it has access to all sensory information broadcast by our perceptu-

al systems, and hence it also has non-interpretive access to one's own sensory states. However, it does not give us direct access to our own thoughts; so we must infer them from observations of our circumstances and behavior, interpreting ourselves just as we interpret others. In this light, the only difference between self- and other- knowledge of thoughts is that in one's own case, the mindreading system has more available information upon which to base its interpretation. As a matter of fact, in addition to using overt behavior, in one's own case it can also draw on a subject's affective, sensory, and quasi-sensory states such as visual imagery or inner speech tokens that are globally broadcast in the mind. In brief, Carruthers's ISA theory restricts self/other parity to a particular subclass of mental states, i.e. propositional attitude events as opposed to mental events with a sensory-based format, which are introspectable (cf. Schwitzgebel, 2019, §§ 2.1.3 and 4.2.2).

Here, then, is how the ISA theory is able to explain what earlier versions of the self/other parity failed to explain, namely, why mentalistic self-attribution can occur even in the absence of behavioral and contextual data, and why one is able to "read" one's own mind better than that of others. Even when I am sitting in my room, motionless and with my eyes closed, I have no difficulty in attributing mental states to myself because I can still rely on a great deal of information regarding the situation I am in, in the form of sensory, imaginative and somatosensory data.

The moral to be drawn is an eliminativist in relation to conscious thought. Since the distinctive feature of the global-broadcasting mechanism is that it is sensory-based, amodal propositional attitudes cannot broadcast themselves, though they might cause sensory-like events (e.g., a sentence in inner speech) which are so broadcast. Outside of the broadly sensory domain (sensation, perception and affect) none of our mental states is ever conscious.

The disappearance of conscious thought still leaves room for a distinction between unconscious *intuitive* processes and conscious *reflective* processes. The latter are forms of mental activity that are directed, for example, toward solving a problem, arriving at a judgment, or reaching a decision. These reflective processes rest on *working memory*, the executive system for directing attention and sustaining and manipulating imagery in the global workspace; and working memory is a *sensory-based* system. First, working memory is a process that emerges and constitutively depends on sensory systems (Postle, 2006); second, top-down attention directed at mid-level perceptual regions of the brain is necessary not only for

conscious perception but also for that contents to enter working memory. The latter uses top-down attention to activate and sustain imagistic representations in conscious form; there is no place within it for amodal propositional attitudes. Since working memory is the system that underlies conscious reflective processes, the latter must be sensorily laden. Supposed conscious thoughts are sensory images in working memory, typically imaged utterances[3].

It is of utmost importance to note that within this framework consciousness is by no means an epiphenomenon, since it performs an essential coordinating function in the mental lives of humans and many other animal species. Perceptual information becomes available to consumer systems only by virtue of global diffusion, and this allows them (and thereby the entire organism) to coordinate around a "common focus"[4].

Even so, the essential feature of the global broadcasting mechanism is its sensory character: an amodal propositional attitude event cannot be globally broadcast, although it can cause a sensory event that can be (e.g., a sentence in internal language). So, except for the sensory domain (sensations, perceptions, and emotions), none of our mental states is available to access consciousness. In particular, there are no entities such as (nonperceptual) judgments, intentions or conscious decisions.

3. *The nexus of moral responsibility and conscious thought in naive ethics*

If ISA theory is well grounded, it puts a strong constraint on the construction of a theory of (moral and legal) responsibility congruent with the findings of neurocognitive sciences: the existence of conscious amodal

[3] As Gomez-Lavin (2017) noted, Carruthers' philosophical treatment of the constructs of attention and working memory leads us to Aristotle's *De Anima*, where the capacity of *phantasia*, like working memory, enables us to entertain a perceptual image in the absence of any stimulus; more crucially, *phantasia* is deemed necessary for all thought, as "the soul never thinks without an image" (431a16).

[4] "Consciousness does make a difference. Indeed, it is vital to the overall functioning of the human mind. […] I certainly don't think consciousness is epiphenomenal. On the contrary, it plays a crucial coordinating function in the minds of humans and most other animals. It is only when information becomes globally broadcast (= becomes access-conscious) that it is made available to a wide range of down-stream systems for drawing inferences, forming memories, evaluating, and so on. This enables all those systems (and thereby the organism as a whole) to become coordinated around a common focus." (Carruthers, 2015b, 1 e 7 agosto).

propositional attitude events cannot be among the theory's commitments (King and Carruthers, 2012, 2022).

That naive ethics establishes a link between moral responsibility and conscious intentional mental states seems to be attested by some research conducted in the field of experimental philosophy applied to the concepts of freedom and responsibility. In the free will debate, philosophers often resort to ordinary intuitions – in particular, it is often claimed that naive ethics is *incompatibilist*. However, Nahmias, Morris, Nadelhoffer and Turner (2006) have argued that their experimental results attest to precisely the opposite: common sense is – as Strawson (1962) had already argued – *compatibilist*. However, Nichols and Knobe (2007), reviewing the findings of Nahmias' group, wondered why so many philosophers who are interested in the question of free will today have become convinced of the incompatibilist nature of ordinary intuitions. Their hypothesis is this: there may be a tendency in people to provide compatibilist answers to concrete questions about particular cases, but incompatibilist answers to abstract questions about general moral principles. If so, the divergence between the data of psychological studies and the conclusions of philosophers would be attributable to a difference between two different ways of *framing* the relevant question.

To test this hypothesis, Nichols and Knobe presented participants with descriptions of two universes, A and B. Universe A is a universe in which everything takes place in accordance with deterministic laws. In universe B, on the other hand, everything occurs in accordance with deterministic laws except for human decisions. Participants were first asked the question "Which universe is most similar to ours?" to which 90% responded by opting for the indeterministic universe B. Then participants were randomly assigned to one of two conditions, abstract and concrete.

Participants placed in the *abstract* condition were asked the following low-emotion question: "In universe A is it possible for a person to have full moral responsibility for his or her actions?" In this condition 86% of the participants gave the incompatibilist answer that in universe A full moral responsibility is not possible. In contrast, participants in the *concrete* condition were presented with a deterministic universe in which a specific agent, Bill, committed a morally reprehensible act (killing his wife and children). The question was: "In your opinion, does Bill bear full moral responsibility for the death of his wife and children?" (Nichols and Knobe, 2007, p. 670). In this concrete and emotionally charged condition, 72% of the subjects gave the compatibilist response that Bill bears full moral responsibility for the murder of his wife and children.

Thus, these data seem to confirm the hypothesis that intuitions about the determinism/responsibility relationship vary depending on the emotional framing of the imagined case. When participants are confronted with macroscopic violations of moral norms, they experience a strong affective reaction (a *reactive* attitude such as moral anger or indignation) that renders them unable to properly apply the underlying naive theory of moral responsibility, which – Nichols and Knobe argue – is incompatibilist. Compatibilist intuitions are then the result of a *performance error* caused by the disruptive influence of emotion on moral judgment. In other words, the bias triggered by strong affect prevents subjects from making the inference that is instead made at the abstract level, leading to the conclusion that determinism excludes responsibility. From this perspective, the conclusion is that the compatibilist intuitions of the ordinary individual are only apparent; and must be set aside as they are subject to the distorting influence of emotional responses.

According to Eddie Nahmias and collaborators (Nahmias, Coates and Kvaran, 2007; Nahmias and Murray, 2010; Nahmias, 2011), however, the scenarios constructed by Nichols and Knobe do not allow the results of their study to be interpreted as evidence of the incompatibilist character of the naive theory of moral responsibility. In fact, Nahmias *et al.* argue, what led the participants in the experiment to deny free will and moral responsibility is the interpretation of determinism as a thesis that implies the idea that the causes of behavior *bypass* the conscious and rational control of the agent. In other words, the description of determinism used by Nichols and Knobe ("everything must occur the way it does in fact occur") may have suggested to the participants that conscious deliberations and ends play no causal role in determining the agent's conduct – i.e. they are *epiphenomenal*[5]. And, indeed, if determinism is interpreted in terms of "bypassing" consciousness, compatibilism is actually doomed (since this view implies that conscious mental states play a relevant role in the generation of action); and the same happens if determinism is interpreted as a form of fatalism – that is, as the belief that certain events will take place regardless of what we decide or try to do. However, Nahmias et al. maintain, determinism does *not* entail bypassing or epiphenomenalism about mental

[5] The key difference is that in universe A each decision is completely caused by what happened before the decision – given the past, each decision must be made the way it is in fact made; in universe B, on the other hand, decisions are not completely caused by the past, and each decision does not have to be made the way it is in fact made.

states or fatalism. In a deterministic universe, natural events remain contingent. Also, determinism does not exclude that conscious mental states play a causal role in human conduct. Quite the contrary: to the extent that our mental states are part of a deterministic sequence of events, they play an essential role in determining what will happen. On this view, then, it is not so much that freedom and responsibility are threatened by determinism as such, but only when it is conceived of as a *reductionist mechanism* – that is, when it is claimed that the higher-level properties of a system (and its changes over time) are reduced to and can be exhaustively explained by its lower-level mechanisms – as when human conduct is reduced to the causal mechanisms of the nervous system in which conscious mental states play no role. In brief, reductionist mechanism asserts that human actions are caused by lower-level mechanisms rather than by his conscious mental states and rational capacities.

In short, while Nichols and Knobe argue that judgments made in high-emotional-impact cases are the outcome of a performance error attributable to the disruptive influence of our emotions and from this conclude that the naive concept of responsibility is incompatible with the truth of determinism, Nahmias *et al*. advance the opposite thesis, namely that performance errors take place when participants mistakenly assume that determinism excludes the possibility of conscious, rational control.

4. *Reconceptualizing the consciousness thesis*

As said, there is some evidence that naive ethics looks to consciousness as the fundamental basis for attributing responsibility – an agent is responsible for an action if it reflects a conscious deliberation on his part. This was translated into normative terms by Levy (2014), who argued for the *consciousness thesis*, which maintains that "consciousness of some of the facts that give our actions their moral significance is a necessary condition for moral responsibility" (p. 1). He contends that since consciousness plays the role of integrating representations, behaviour driven by non-conscious representations are inflexible and stereotyped, and only when a representation is conscious "can it interact with the full range of the agent's personal-level propositional attitudes" (ibid., p. vii). This fact entails that consciousness of key features of our actions is a necessary (though not sufficient) condition for moral responsibility since consciousness of the morally significant facts to which we respond is required for these facts to be

assessed by, and expressive of, the agent. Furthermore, he argues that the two leading accounts of moral responsibility – *real self* account (Frankfurt, 1971, 1988) and *control-based* account – are committed to the truth of the consciousness thesis despite what proponents of these accounts maintain. According to Levy, (a) only the actions that are performed consciously can express our evaluative agency, and the expression of moral attitudes requires consciousness of that attitude; and (b) we possess responsibility-level control only over actions that we perform consciously, and control over their moral significance requires consciousness.

However, the consciousness thesis seems to contradict the constraint that ISA theory imposes on the construction of a theory of responsibility. In fact, to be congruent with data from the neurocognitive sciences, such a theory must not presuppose the existence of conscious amodal propositional attitude events. In the case of the real self, it is claimed that an agent can be held responsible exclusively for those actions that have been caused by psychological states reflecting its identity as practical agent. But if the propositional attitudes that define the agent's real self are the conscious ones, the elimination of conscious thought implies the non-existence of the real self (King and Carruthers, 2012, pp. 217ff).

Now let us ask: would a theory of responsibility that satisfies this constraint allow us to preserve at least part of the considerations that motivate the idea that the actions for which we are responsible are the actions that originate from conscious attitudes and decisions? For example, would such a theory allow us to distinguish between actions that originate from so-called "implicit attitudes" and actions that arise from conscious reflection? Suppose, for example, that an individual is totally unaware that they have an *implicit bias* against people of colour. Consequently, as they review some job applications, they prefer a less qualified white applicant to a black applicant. Should this person be blamed for doing so? Certainly we should be in a position to say that this individual is far less culpable than someone who, while reading the resume, thinks "I would never hire a person of colour" and for that very reason trashes the application.

According to Carruthers (2015a, §§3.3 and 3.5) this distinction can still be drawn in his ISA theory. Indeed, although the latter does not allow a distinction to be drawn between conscious and unconscious amodal attitudes (amodal attitudes being all unconscious), a kindred distinction can still be drawn – that is, one can still distinguish between attitudes that are formed by virtue of one's conscious reflections and those that are caused by unconscious processes. Attitudes that originate from conscious reflec-

tion are still unconscious attitudes (they are those whose existence is often known to the subject as the result of the mentalistic interpretation of the sensory contents of reflection). Nevertheless, they are attitudes to the formation of which the whole person has contributed (and note that here Carruthers is following Levy, 2014):

> Asking oneself in inner speech, "What should I decide?," for example, issues in a globally broadcast request for information, thereby allowing all the different consumer subsystems that receive such broadcasts a chance to contribute an answer. There is a good sense, then, in which attitudes that are formed as a result of conscious reflection are owned by the whole person, in a way that a decision to redirect attention to the sound of one's own name is not (Carruthers, 2015a, p. 237).

Within this framework, the distinction between personal and subpersonal attitudes can be reformulated. They are attitudes of the same type but differ with regard to their etiologies: personal attitudes, but not subpersonal attitudes, are unconscious attitudes that are caused by conscious reflection.

Applying this distinction to the case of the implicit bias, we obtain the following. A decision that arises from conscious reflection on the alleged demerits of people of color is one to which the whole person contributes. It therefore reflects, in a sense, *the self as a whole*. In contrast, where the decision is caused by an unconscious bias, it reflects that bias and nothing more. All of the person's other purposes and values might tend in the opposite direction, so that if his attention had been focused on the difference in competence between the two candidates as well as the implicit bias, they would have immediately chosen the black candidate.

From this perspective, cognitive neuroscience by no means leads to the epiphenomenalism of consciousness; rather, it allows for a finer-grained articulation of the dialectic between unconscious processing and conscious reflection. And this undeniably is an important piece in a theory of responsibility that aspires to hinge the normative plane on the descriptive one.

References

Anscombe, E. (1957). *Intention*. Oxford: Blackwell.

Carruthers, P. (2011). *The Opacity of Mind*. Oxford: Oxford University Press.

Carruthers, P. (2015a). *The Centered Mind*. Oxford: Oxford University Press.

Carruthers, P. (2015b). Who's in charge anyway? Pubblicato sul blog della Oxford University Press il 1 agosto 2015: <http://blog.oup.com/2015/08/whos-in-charge-conscious-mind/>.

Carruthers, P. (2019). *Human and Animal Minds*. Oxford: Oxford University Press.

Dehaene, S. (2014). *Consciousness and the Brain*. New York: Viking.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy*, 68(1), pp. 5-20.

Frankfurt, H. (1988). *The Importance of What We Care About*. Cambridge: Cambridge University Press.

Johannson, P., Hall, L., Tärning, B., Sikström, S., Chater, N. (2013). Choice Blindness and Preference Change: You Will Like This Paper Better If You (Believe You) Chose to Read It!. *Journal of Behavioral Decision Making*, https://doi.org/10.1002/bdm.1807Legal Information Institute of Cornell University.

King, M., Carruthers, P. (2012). Moral responsibility and consciousness. *Journal of Moral Philosophy*, 9, pp. 200-28.

King, M., Carruthers, P. (2022). Responsibility and consciousness. In D. Nelkin and D. Pereboom (eds.), *Handbook of Moral Responsibility*. Oxford: Oxford University Press, pp. 448-67.

Levy N. (2014). *Consciousness and Moral Responsibility*. Oxford: Oxford University Press.

Nahmias, E. (2011). Intuitions about free will, determinism, and bypassing. In *The Oxford Handbook on Free Will*. Oxford: Oxford University Press, 2nd edn., pp. 555-75.

Nahmias, E., Coates, D., Kvaran, T. (2007). Free will, moral responsibility, and mechanism: Experiments on folk intuitions. *Midwest Studies in Philosophy*, 31, pp. 214-42.

Nahmias, E., Morris, S., Nadelhoffer, T., Turner, J. (2006). Is incompatibilism intuitive? *Philosophy and Phenomenological Research*, 73, pp. 28-53.

Nahmias E, Murray D. (2010). Experimental philosophy on free will: an error theory for incompatibilist intuitions. In *New Waves in Philosophy of Action*. New York: Palgrave-Macmillan, pp. 189-215.

Nichols, S., Knobe, J. (2007). Moral responsibility and determinism. *Nous*, 41, pp. 663-85.

Nisbett, R., Wilson, T.D. (1977). *Telling more than we can know: Verbal reports on mental processes*. *Psychological Review*, 84, pp. 231-59.

Ryle, G. (1949). *The Concept of Mind*. London: Routledge, 2009.

Schwitzgebel, E. (2019). Introspection. In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/archives/win2019/entries/introspection/>.

Strawson, P.F. (1962). Freedom and resentment. *Proceedings of the British Academy*, 48, pp. 1-25.

Wegner, D.M. (2002). *The Illusion of Conscious Will*. MIT Press, Cambridge (MA).

Wegner, D.M., Wheatley, T.P. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist*, 54, pp. 480-92.

Wilson, T.D. (2002). *Strangers to Ourselves*. Cambridge (MA): Harvard University Press.

Wittgenstein, L. (1953). *Philosophische Untersuchungen*. Oxford: Blackwell (Engl. transl. 1986).

Abstract

Nowadays there is a strong tension between cognitive neuroscience and many ethical views based on the ordinary view of the world. On the one hand, many cognitive neuroscientists and empirically oriented philosophers raise a radical doubt about the ordinary conception of ourselves as conscious thinking agents who causally control their actions – where conscious thinking includes our beliefs, goals, decisions, and intentions. On the other hand, many ethicists still accept the ordinary conception of ourselves and, consequently, look at consciousness as one of the two fundamental bases for attributing responsibility: agents are responsible for their actions as long as such actions reflect their conscious deliberations (the other basis for the attribution of responsibility is that conscious deliberations do contribute causally to the generation of actions).

After exposing this disagreement, we will advocate the adoption of an intermediate position between that advocated by traditional ethicists (who, in spite of the data emerging from mind and brain sciences, keep attributing an absolute primacy to conscious thought in moral agency) and that held by cognitive neuroscientists and philosophers (who venture to claim that the conscious mind is indeed epiphenomenal). We will argue that an alternative and more promising model may be built by referring to some suggestions by Neil Levy, Peter Carruthers, and Matt King. In this light, we will claim that

cognitive neuroscience's findings – rather than showing that the conscious mind is epiphenomenal – require that we offer a finer-grained and unbiased articulation of the dialectic between unconscious processing and conscious reflection.

Keywords: conscious thought; experimental moral philosophy; moral responsibility; personal and subpersonal attitudes.

Mario De Caro
Università di Roma 3
mario.decaro@tlc.uniroma3.it

Massimo Marraffa
Università di Roma 3
massimo.marraffa@uniroma3.it

T

Marco Menon

L'esternalizzazione dei processi di decisione nella società postindustriale. Vilém Flusser e il funzionario nell'apparato*

1. Introduzione

Sostenere che le tecnologie hanno radicalmente mutato il mondo umano e sociale, plasmandolo sotto quasi ogni aspetto della vita quotidiana, individuale e collettiva, è pressoché riconosciuto come un luogo comune, un'ovvietà. Tuttavia, già il fatto di non provare quasi più meraviglia, e dare quindi per scontata una condizione che è frutto di un lungo processo di mutazioni e interazioni tra essere umano e tecnologie racchiude un rischio enorme, cioè quello di non tematizzare, e peggio ancora di naturalizzare, tutta una serie di dinamiche che invece richiedono una discussione pubblica partecipata e un'attenta analisi filosofica (sotto diversi punti di vista: non da ultimo quello etico-politico), pena l'adeguarsi a quella "normatività nascosta" che tutte le innovazioni tecnologiche, in modi più o meno efficaci, sempre veicolano[1]. Come scrive Paolo Benanti, «tutti gli ambiti della nostra realtà sono ora soggetti non più a una mediazione diretta tra questi stessi ambiti e l'uomo; sono invece relazioni mediate da artefatti»[2]. Ciò che allo stato attuale più ci deve interessare sono due aspetti: in primo luo-

* Parti di questo articolo sono confluite in M. Menon, *Vilém Flusser e la «rivoluzione dell'informazione». Comunicazione, etica, politica*, Edizioni ETS, Pisa 2022.

[1] Si vedano, al proposito, le considerazioni di R. Casati, *Contro il colonialismo digitale. Istruzioni per continuare a leggere*, Laterza, Roma-Bari 2014.

[2] P. Benanti, *Le macchine sapienti. Intelligenze artificiali e decisioni umane*, Marietti 1820, Bologna 2018, p. 51.

 DOI: 10.4454/teoria.v42i2.159

go, che questi artefatti sono dotati di intelligenze artificiali. Sono in grado, cioè, con una autonomia sempre maggiore, di interagire in maniera indipendente dall'agente umano con l'ambiente che li circonda, di modificare i loro processi operativi apprendendo per prove ed errori. In secondo luogo, anche per la ragione appena detta, è necessario parlare di una condizione in cui non abbiamo di fronte a noi delle tecnologie specifiche, che possiamo affrontare, per così dire, una alla volta, come se potessimo isolarle, asservirle ai nostri scopi e *scegliere come e quando usarle o riporle*. La tecnologia, come già aveva inteso negli anni '50 Günther Anders, non è più fatta di dispositivi o attrezzi singoli, ma costituisce un *mondo*, all'interno del quale ci troviamo già catturati e anticipati, e di cui non possiamo disporre in senso meramente strumentale, per perseguire i nostri fini. Scrive ancora Benanti, quella con cui oggi abbiamo a che fare è una «tecnologia generale, cioè […] un tipo di tecnologia che non serve a svolgere un singolo, determinato, compito, ma che cambia il modo in cui facciamo tutte le cose»[3].

Il riferimento al "mondo" sembra limitare la presenza di questa mediazione tecnologica alla "dimensione esteriore", ma in realtà le cose sono ben più complesse. Come, tra gli altri, osservava il pensatore francese Paul Virilio, queste tecnologie sempre più sofisticate non solo si sono progressivamente innervate nel «corpo territoriale», ma hanno anche conquistato «il corpo dell'uomo, il corpo proprio di un individuo rapidamente sottomesso al regno della bio-tecnologia, di queste nano-macchine capaci di *colonizzare* non più solo l'estensione del mondo, ma lo spessore stesso del nostro organismo»[4]. C'è da chiedersi se, e in che misura, in virtù della potenza modellante e manipolatrice di tali tecnologie la colonizzazione di cui parla Virilio possa aver avuto un'influenza non solo sul modo di "fare le cose", ma anche sul modo di pensare. Il riferimento è a quanto nei comportamenti umani, sotto l'influenza e il condizionamento dell'automazione, può essere descritto come una sorta di mero "funzionamento" che progressivamente contrae l'orizzonte della nostra autonomia decisionale. Infatti, grazie ai più recenti sviluppi degli algoritmi, deleghiamo sempre più alcuni dei nostri stessi processi decisionali alle macchine e ai dispositivi, esternalizzando per così dire tutta una serie di operazioni che possono essere svolte con maggiore velocità ed efficacia da agenti non umani[5]. Ciò solleva una serie

[3] *Ibidem*. Dello stesso autore si veda anche *La condizione tecno-umana. Domande di senso nell'era della tecnologia*, EDB, Bologna 2016.

[4] P. Virilio, *La velocità di liberazione, Strategia della lumaca edizioni*, Roma 1997, p. 113.

[5] Un inquadramento delle problematiche specifiche viene offerto in particolare da due articoli, ormai punti di riferimento consolidati nella letteratura: B.D. Mittelstadt et al., *The ethics*

di interrogativi, spesso dal tono apocalittico, sull'autentica natura di tali fenomeni. Si tratta di una liberazione da compiti gravosi, che ci permette di risparmiare tempo e risorse, oppure di una sottile forma di schiavitù mascherata da emancipazione[6]? Un pessimista come il già citato Virilio argomentava in quest'ultima direzione, laddove osservava, in ambito bellico, che la vertiginosa contrazione degli spazi e dei tempi innescata dall'accelerazione tecnologica e informazionale avrebbe portato alla «fatale messa in opera della *automazione della decisione*» in cui «si cancella la responsabilità diretta della decisione umana»[7]. La crescente presenza e penetrazione dei dispositivi "intelligenti" nel nostro mondo della vita non lascerebbe quindi incontaminato alcuno spazio, e ciò vale non soltanto per la dimensione corporea, sia essa "esterna" o "interna", ma anche per le attività superiori dell'intelletto umano.

È forte quindi la percezione del rischio reale non tanto di una servitù (più o meno) volontaria[8], quanto di una inconsapevole abdicazione allo statuto di esseri agenti consapevoli e responsabili, di una sospensione delle decisioni esistenziali per lasciarsi guidare da una procedura prestabilita, da un protocollo, o da agenti artificiali, come nel caso emblematico, raccontato da Anders, del generale McArthur, che si affidò a una "macchina-oracolo" per valutare l'opportunità di iniziare quella che sarebbe stata a tutti gli effetti la terza guerra mondiale[9]. Il pericolo è una narcosi della co-

of algorithms: Mapping the debate, in «Big Data & Society», III (2016), n. 2, pp. 1-21; L. Royakkers et al., *Societal and ethical issues of digitization*, in «Ethics and Information Technology», XX (2018), pp. 127-142; per un confronto tra le questioni poste dall'intelligenza artificiale e la dimensione della decisione, considerando la storia della filosofia moderna, si veda C. Canullo, *Chi decide? Intelligenza artificiale e trasformazioni del soggetto nella riflessione filosofica*, in E. Calzolaio (a cura di), *La decisione nel prisma dell'intelligenza artificiale*, Wolters Kluwer-CEDAM, Milano 2020, pp. 25-35.

[6] Si veda ad esempio il cosiddetto fenomeno dell'algoritmocrazia (*algocracy*), introdotto da A. Aneesh, *Virtual Migration. The Programming of Globalization*, Duke University Press, Durham-London, 2006, p. 5: «definisco *algoritmocrazia* il governo dell'algoritmo, o il governo del codice, che forse rappresenta la differenza chiave tra la fase attuale e quella precedente nell'integrazione globale»; tale concetto è ampliato e discusso in profondità da J. Danaher, *The Threat of Algocracy: Reality, Resistance and Accommodation*, in «Philosophy and Technology», XXIX (2016), n. 3, pp. 245-268.

[7] P. Virilio, *Lo spazio critico*, Edizioni Dedalo, Bari 1988, pp. 138-139.

[8] Sul tema si veda R. Cubeddu, *Nuove tirannidi. Conseguenze inintenzionali della dipendenza della politica dalla scienza*, IBL Occasional Papers 105, online, 2016, https://www.brunoleoni.it/nuove-tirannidi-conseguenze-inintenzionali-della-dipendenza-della-politica-dalla-scienza [ultimo accesso: 13/06/2022].

[9] G. Anders, *L'uomo è antiquato I. Considerazioni sull'anima nell'epoca della seconda rivoluzione industriale*, Bollati Boringhieri, Torino 2003, pp. 65-67.

scienza critica. Tra i pensatori che si sono interessati a tale problematica va contato il filosofo ceco Vilém Flusser (1920-1991), noto anzitutto per i suoi scritti sulla comunicazione e sui media. Ma Flusser è stato anche il pensatore che si è posto il problema del "carattere non-umano" della decisione nelle società postindustriali e che nell'esternalizzazione dei processi di decisione alle intelligenze artificiali, che lui chiamava "apparati", ha visto tanto il rischio di un nuovo totalitarismo quanto l'occasione di una radicale emancipazione dell'essere umano.

Il presente contributo discuterà tale ambiguità nella prospettiva flusseriana, e sarà diviso in quattro parti. La prima illustrerà cosa Flusser intendesse per società postindustriale. La seconda parte fornirà una spiegazione della sua nozione di "apparato". Nella terza verrà articolato il problema della decisione nella società "governata" dagli apparati, mentre nella quarta e ultima parte verrà discussa la natura dell'ambiguità delle macchine che decidono, al fine di metterne in luce il potenziale utopico.

2. *Società postindustriale e tecnocrazia*

Flusser è un pensatore che si concentra sulle grandi trasformazioni culturali, su cambiamenti strutturali che comportano quelle che Friedrich Nietzsche definiva "trasvalutazioni dei valori". Nelle sue opere è possibile identificare almeno tre trasformazioni del genere, qui elencate in ordine di radicalità crescente:

– il passaggio dalla società industriale a quella postindustriale, ovvero il passaggio dalle macchine con le quali lavorano gli operai agli apparati nei quali operano i funzionari. L'epoca industriale si estende, indicativamente, dal XVIII al XX secolo;
– il passaggio dalla storia alla post-storia, ovvero un mutamento che riguarda i codici di comunicazione e il loro influsso sulla costruzione dell'esperienza. L'epoca cosiddetta "storica", cioè quella informata dalla scrittura lineare, si estende dal II millennio a.C. al XX secolo;
– il passaggio dalla sedentarietà al nomadismo, cioè la trasformazione delle strutture comunicative con l'emergere di un nuovo coordinamento tra una struttura discorsiva dominante, il *broadcasting* dei mezzi di comunicazione di massa, e una struttura dialogica resiliente, quella reticolare di media come telefono, posta, reti telematiche. L'epoca della sedentarietà, da intendere in senso comunicologico, si estende dall'8.000 a.C. al XX secolo.

Nel presente intervento verrà preso in esame un fenomeno che caratterizza il passaggio alla società postindustriale. Per questa ragione, in primo luogo è necessario chiarire che cosa Flusser intenda con questo termine[10]. A questo scopo, verrà preso in considerazione il capitolo *Nosso trabalho* dal libro *Pós-História*, composto nel 1979 in portoghese e pubblicato nel 1983[11]. Qui il filosofo praghese presenta la trasformazione che porta da un tipo di società all'altra come un cambiamento di portata ontologica. Sostiene infatti che il «passaggio dalla società agricola a quella industriale ha degli effetti *ontologici*. Il contadino esperisce la realtà in maniera differente dall'operaio. Il passaggio attuale dalla società industriale a quella postindustriale avrà un effetto paragonabile. L'operaio esperisce la realtà in maniera differente dal funzionario»[12]. Il passaggio sembra allora consistere in un cambiamento della visione del mondo, che a sua volta implica una specifica ontologia. Flusser procede quindi a illustrare dapprima il cambiamento che si è verificato nel passaggio dalla società agricola del contadino a quella industriale dell'operaio, analizzando la visione del mondo delle due forme di vita che determinano il carattere delle società in cui sono predominanti. L'agricoltura viene definita come «manipolazione paziente della natura vivente», mentre l'industria è la «manipolazione violenta della natura inanimata: la costringe a riformularsi in modo conforme a dei modelli concepiti in precedenza»[13]. Il contadino fa un'esperienza di attesa: egli infatti aspetta, con cura e attenzione, che il bestiame e le colture si sviluppino e raggiungano il grado di "utilità" desiderato. L'operaio (o l'ingegnere) al contrario fa un'esperienza di costrizione: obbliga la materia grezza a conformarsi al suo modello o al suo progetto. Le due prassi sottendono due visioni differenti della realtà: «per i contadini la realtà è un ente animato posto sotto la sua tutela. Per l'ingegnere la realtà è un materiale che deve essere preso a martellate, bruciato, gassificato»[14]. Queste visioni informano rispettivamente la concezione degli altri esseri umani e l'imma-

[10] Cfr. D. Bell, *The Coming of Post-Industrial Society: A Venture in Social Forecasting*, Basic Books, New York 1973.

[11] In lingua tedesca esce prima il testo ricostruito e redatto da Volker Rapsch (con approvazione dell'autore) in V. Flusser, *Nachgeschichten. Essays, Vorträge, Glossen*, Bollmann, Düsseldorf 1990; poi, in versione integrale e non editata, in V. Flusser, *Nachgeschichte. Eine korrigierte Geschichtsschreibung*, hrsg. von S. Bollmann und E. Flusser, Fischer Verlag, Frankfurt a.M. 1997.

[12] V. Flusser, *Pós-História. Vinte instantâneos e um modo de usar*, Annablume, São Paulo 2011, p. 47.

[13] *Ivi*, p. 47.

[14] *Ibidem*.

gine del mondo più generale. Estendendo le visioni emergenti dalle prassi alla sfera del sociale, per il contadino gli esseri umani saranno una specie di gregge che necessita un pastore, mentre per l'operaio o l'ingegnere gli esseri umani saranno una specie di materia da forgiare; estendendole invece alla cosmologia, avremo da una parte una visione "aristotelica" del mondo e dell'ordine naturale, e dall'altra quella della scienza moderna dove è l'intelletto a dare le leggi alla natura. Di conseguenza anche la concezione della teoria cambia: contemplazione di forme immutabili nel primo caso, invenzione di forme mutevoli nel secondo. A chi obbietta che nella società agricola erano presenti artigiani, e che esistono ancora contadini nella società industriale, Flusser risponde che gli artigiani nella società agricola pensavano, esperivano e valutavano in piena coerenza con la visione "agricola" del mondo, quindi al servizio "dell'attesa" dei contadini, e viceversa. Per il praghese il passaggio da una forma di società all'altra non è solamente una questione sociologica: comporta, come detto, una vera e propria trasvalutazione di valori, che coinvolge pensiero e azione.

Una trasformazione dello stesso tipo è quella a cui stiamo assistendo noi, secondo Flusser, con il passaggio dalla società industriale a quella postindustriale. Non si tratta meramente del fatto che oggi la maggior parte degli esseri umani, almeno nelle società occidentali, è occupata nel settore dei servizi, nel terziario, e solo una minoranza lavora ancora nei settori secondario e primario. Sulla base di quanto detto, questa sarebbe solo una concezione quantitativa della società postindustriale. Per comprendere la visione che la informa, Flusser analizza la prassi della forma di vita che la caratterizza: il funzionario.

Il tipico funzionario sta seduto alla scrivania, dove riceve da altri funzionari fogli di carta ricoperti da simboli. Questi documenti vengono o archiviati oppure elaborati (ricoperti di altri simboli) e trasmessi ad altri funzionari. «Il funzionario riceve simboli, immagazzina simboli, produce simboli ed emette simboli. Lo fa in parte ancora manualmente, e in parte già grazie ad apparati cibernetici [*aparelhos cibernéticos*] del tipo "word processors". La sua prassi ha luogo in un contesto chiamato "mondo codificato"»[15]. I simboli con cui opera il funzionario sono dei fenomeni che stanno per altri fenomeni, ovvero sono segni che significano qualcosa sulla base di una convenzione, sia essa pianificata o emergente. Questi simboli possono essere di due tipi: vengono detti "osservazionali" quando significano un fenomeno "concreto" del mondo "reale", oppure sono

[15] *Ivi*, pp. 49-50.

"teorici" quando significano altri simboli. Si potrebbe pensare che, facendo risalire tutti i simboli teorici a simboli osservazionali e quindi a fatti concreti, il funzionario operi in maniera indiretta per cambiare il mondo reale. Ma per Flusser le cose non stanno così. Normalmente pensiamo che, ad esempio, un documento ricoperto di simboli come un passaporto "significhi" una persona reale, in carne e ossa. La dinamica immanente agli «apparati entro i quali i funzionari funzionano» inverte però "i vettori di significato", nel senso che è «la persona concreta, che riceve il passaporto, a significare il passaporto. È essa il simbolo, e il passaporto è il significato»[16]. La realtà del funzionario è una realtà fatta di simboli convenzionali, e gli esseri umani hanno un qualche valore nella misura in cui rimandano in maniera conforme a questi ultimi. Ciò non significa però che la società del funzionario, la società postindustriale, «sarà burocratica». Al contrario, per Flusser

> *dove c'è burocrazia, la società postindustriale non è ancora ben programmata*. Tutto indica che la società postindustriale sarà dominata dai programmi di funzionamento, nei quali i funzionari funzioneranno come ingranaggi sempre più invisibili all'interno di scatole nere [*caixas pretas*]. Ovvero, che [la società postindustriale] sarà una *tecnocrazia*.

Con ciò non intende dire che la vera classe dominante sarà composta dai programmatori degli apparati, da coloro che padroneggiano i codici con cui vengono messi a punto i software che governano i nostri dispositivi. Anche costoro non sono altro che funzionari specializzati. Quando Flusser parla di tecnocrazia, intende dire che «la vera classe dominante sarà quella degli apparati»[17], cioè entità artificiali in cui gli esseri umani funzionano come componenti sostituibili. Nella tecnocrazia gli esseri umani sono cifre da inserire in "giochi formali", memorie da programmare per un determinato comportamento o per svolgere mansioni specifiche.

L'analisi del pensatore praghese va naturalmente vista alla luce delle grandi innovazioni tecnologiche verificatesi nei successivi 40 anni, le quali però hanno effettivamente portato, per usare le sue parole, «all'invenzione di computer e dispositivi intelligenti, e alla trasformazione della società in un sistema cibernetico composto da funzionari e apparati»[18]. Le tendenze

[16] *Ivi*, p. 50.
[17] *Ivi*, p. 52.
[18] *Ivi*, p. 53. Per un quadro generale della situazione attuale, si veda L. Floridi, *La quarta rivoluzione: Come l'infosfera sta trasformando il mondo*, Raffaello Cortina Editore, Milano 2017.

che più interessano a Flusser dal punto di vista filosofico-esistenziale, infatti, non hanno fatto altro che accentuarsi e consolidarsi. Il quadro della società postindustriale qui presentato sembra quindi essere valido per la situazione odierna.

3. *Gli apparati come simulazione di processi mentali*

È necessario ora capire che cosa egli intendesse per apparato[19]. Uno dei luoghi classici in cui Flusser offre un'analisi distesa e relativamente chiara del concetto di apparato si trova in *Per una filosofia della fotografia*. La versione portoghese del saggio ha un titolo diverso da quello tedesco (*Für eine Philosophie der Fotografie*), in realtà più efficace e diretto: *Filosofia da caixa preta. Ensaios para uma futura filosofia da fotografia* (*Filosofia della scatola nera. Saggi per una filosofia futura della fotografia*). Lo spostamento di accento è rivelatore: abbiamo a che fare con una filosofia della scatola nera, o *black box*, che è il termine usato da Flusser per designare gli apparati, i quali, superando le competenze dei singoli funzionari che operano al loro interno o interagiscono con dispositivi intelligenti, restano in larga parte imperscrutabili[20]. Se ne può ricavare una chiave di lettura, ovvero trattare la filosofia della fotografia di Flusser anzitutto come una filosofia degli apparati. Cosa che viene suggerita anche dalla prefazione alla versione portoghese, dove il filosofo praghese ammette che «l'intenzione che muove questo saggio è quella di contribuire a un dialogo filosofico sull'*apparato* in funzione del quale vive la nostra epoca, prendendo come pretesto il tema della *fotografia*»[21]. Lo sviluppo dell'argomentazione

[19] Il background del concetto di apparato in Flusser, che fra le altre cose comprende figure mitiche e letterarie come il Golem di Praga e il dramma teatrale fantascientifico *R.U.R.* (ovvero *Rossumovi univerzální roboti*) di Karel Čapek, è stato ricostruito in modo esauriente da R. Guldin, *Golem, Roboter und andere Gebilde. Zu Vilém Flussers Apparatbegriff*, in «Flusser Studies», IX (2009), pp. 1-17.

[20] Sul concetto di *black box* in Flusser, e le problematiche a esso legate, cfr. D. Irrgang, *Vilém Flussers Black Box*, in E. Geitz, C. Vater, S. Zimmer-Merkle (a cura di), *Black Boxes – Versiegelungskontexte und Öffnungsversuche. Interdisziplinäre Perspektiven*, De Gruyter, Berlin-Boston 2020, pp. 53-69.

[21] V. Flusser, *Filosofia da caixa preta. Ensaios para uma futura filosofia da fotografia*, Editora Hucitec, São Paulo 1985, p. 8. Questo è un aspetto che in realtà Flusser esplicita a più riprese nel corso del saggio, ma che evidentemente ha ritenuto opportuno chiarire nella prefazione al pubblico brasiliano. Si veda ad esempio V. Flusser, *Per una filosofia della fotografia*, cit., p. 96: «l'apparato fotografico si rivelerà l'avo di tutti quegli apparati che si apprestano a robotizzare tutti gli aspetti della nostra vita, dal gesto più esteriore fino all'aspetto più intimo del pensare,

del saggio nella sua versione tedesca è coerente con questa affermazione. Flusser infatti scrive:

> si può supporre che l'apparato fotografico racchiuda – in forma elementare, embrionale – le proprietà caratteristiche degli apparati in generale e che tali caratteristiche possano essere elaborate a partire da esso. […] In quanto prototipo degli apparati che sono divenuti così determinanti per il presente e per l'immediato futuro, l'apparato fotografico offre un approccio adeguato per un'analisi generale degli apparati – di quegli apparati che da una parte crescono a dismisura e rischiano di scomparire dal campo visivo (come gli apparati amministrativi), e dall'altra si ritraggono fino a raggiungere dimensioni microscopiche, per sottrarsi completamente al nostro intervento (come i chip degli apparati elettronici)[22].

Si pone immediatamente la questione dell'ambiguità legata al termine tedesco *Apparat*, che può essere tradotto sia come "apparato" sia come "apparecchio". Per Flusser sistemi amministrativi, come ministeri e dipartimenti, e dispositivi elettronici, come la fotocamera o il minitel, sono tutti *Apparate*, nonostante si tratti di fenomeni a prima vista molto diversi. È necessario perciò analizzare più da vicino questo concetto e vedere in che modo possa funzionare nel tenere assieme cose apparentemente estranee tra loro.

Flusser "gioca" con l'etimologia del termine latino *apparatus*, e osserva che deriva da *apparare*, cioè *vorbereiten* in tedesco. Ma in latino esiste anche il termine *praeparare*, che in tedesco viene reso sempre con *vorbereiten*. Cercando di restituire la differenza tra i prefissi *prae-* e *ad-* allo scopo di illustrare due sfumature dell'*apparatus*, Flusser propone di distinguere tra *vor-bereiten* e *für-bereiten*. L'apparato è dunque qualcosa di pronto per qualcos'altro, nella doppia disposizione di essere in *attesa* e in *agguato*. La rapacità dell'apparato ne mette in evidenza lo statuto ambiguo rispetto all'essere umano. Se si tratta di uno strumento (*Werkzeug*), di che tipo di strumento si tratta? Argomentando in maniera non distante da Marshall McLuhan[23], Flusser sostiene che «gli utensili [*Werkzeuge*] in senso comune

del sentire e del volere»; p. 102: «una filosofia della fotografia può essere il punto di partenza per ogni filosofia che si occupi dell'*esistenza attuale e futura dell'uomo*»; p. 111: «la filosofia della fotografia è necessaria per portare alla coscienza la pratica fotografica; e ciò è a sua volta necessario perché in questa pratica appare un *modello di libertà nel contesto generale postindustriale*», corsivo aggiunto.

[22] V. Flusser, *Per una filosofia della fotografia*, cit., p. 21, trad. it. modificata.

[23] V. Flusser, *La leva passa al contrattacco*, in Id., *Filosofia del design*, Bruno Mondadori, Milano 2003, p. 43: «Le macchine sono simulazioni degli organi del corpo umano. La leva, per esempio, è un prolungamento del braccio. In essa è potenziata la capacità di sollevare, mentre tutte le

sono prolungamenti degli organi del corpo umano […] Essi simulano l'organo che prolungano»[24]. Nel caso degli strumenti semplici, preindustriali, abbiamo a che fare con simulazioni empiriche: il martello, ad esempio, prolunga e simula il pugno sulla base dell'esperienza. Con la rivoluzione industriale, invece, il prolungamento e le simulazioni avvengono sulla base di teorie scientifiche. In questo caso non si tratta più di meri *Werkzeuge* ma di macchine. Ma questa non è ancora una definizione sufficiente per caratterizzare l'apparato. Flusser ci mette in guardia dall'interpretare il fenomeno dell'apparato con categorie appartenenti alla società industriale. Gli apparati non eseguono un lavoro come fanno invece le macchine: non producono degli oggetti prendendo "pezzi" di natura per dare loro una forma (per *in-formarli*). Come si è visto in precedenza nel trattare la prassi del funzionario, gli apparati *producono simboli* e operano nella dimensione chiamata "mondo codificato".

Se l'apparato non è uno strumento (*Werkzeug*), né una macchina (*Maschine*), è allora possibile accostarlo o identificarlo con il *dispositif* foucaultiano, come accade in letteratura[25]? Non si tratta in realtà della stessa cosa. Se si prende come riferimento la definizione che ne ha dato Giorgio Agamben, un dispositivo è

> qualunque cosa abbia in qualche modo la capacità di catturare, orientare, determinare, intercettare, modellare, controllare e assicurare i gesti, le condotte, le opinioni e i discorsi degli esseri viventi. Non soltanto, quindi, le prigioni, i manicomi, il Panopticon, le scuole, […] ma anche […] il linguaggio stesso, che è forse il più antico dei dispositivi, in cui migliaia e migliaia di anni fa un primate […] ebbe l'incoscienza di farsi catturare[26].

Nonostante vi sia in comune tra il dispositivo e l'apparato il gesto predatore della cattura ai danni dell'umano, la differenza emerge chiaramente quando Flusser parla del linguaggio come se fosse un apparato, correggendosi subito dicendo che il suo esempio in realtà non è appropriato:

altre funzioni che il braccio possiede vengono trascurate. La leva è più "stupida" del braccio, ma arriva più lontano e solleva pesi maggiori». Sulla vicinanza a McLuhan, basti pensare a passi come questo: «Ogni invenzione o tecnologia è un'estensione o un'autoamputazione del nostro corpo, che impone nuovi rapporti o nuovi equilibri tra gli altri organi e le altre estensioni del corpo», M. McLuhan, *Gli strumenti del comunicare. Mass media e società moderna*, Net, Milano 2002, p. 55.

[24] V. Flusser, *Per una filosofia della fotografia*, cit., p. 25.

[25] Si veda ad esempio B. Stricklin, *Apparatus*, in A. Jaffe, M.F. Miller, R. Martini (a cura di), *Understanding Flusser, Understanding Modernism*, Bloomsbury Publishing, London 2021, pp. 272-274.

[26] G. Agamben, *Che cos'è un dispositivo?*, Nottetempo, Roma 2006, pp. 21-22.

i programmi degli apparati consistono in simboli. Funzionare significa dunque giocare con i simboli e combinarli. Un esempio anacronistico illustrerà la cosa: possiamo considerare lo scrittore un funzionario dell'apparato "linguaggio" che gioca con i simboli contenuti nel programma linguistico – le parole – combinandoli. […] L'esempio è anacronistico in quanto *il linguaggio non è un apparato*; non è stato creato per *simulare un organo del corpo*, e la sua creazione non *poggia su alcuna teoria scientifica*[27].

Dal questo breve confronto tra Agamben e Flusser è possibile trarre due informazioni. In primo luogo, un apparato può anche essere inteso come una specie di dispositivo, ma non tutti i dispositivi sono anche degli apparati. In secondo luogo, (i) essere simulazione di un organo del corpo e (ii) poggiare su teorie scientifiche sono due condizioni necessarie perché si possa parlare di apparato. Ma come si vedrà subito non sono anche condizioni sufficienti.

Finora alla domanda sulla natura dell'apparato sono state fornite risposte sostanzialmente negative. Sappiamo che cosa non è, e ora bisogna darne una definizione positiva. Flusser sembra farlo in modo soddisfacente quando scrive che è un «giocattolo complesso [*komplexes Spielzeug*]», sottolineando così l'assenza del *Werk* che viene sostituito dallo *Spiel*[28], un gioco però che trascende la comprensione di chi ci gioca, ovvero che supera la capacità, da parte del funzionario, di esaurire le possibilità programmate nell'apparato con cui opera[29]. Per chiarire questo aspetto però Flusser fa un passo ulteriore, che esplicita davvero la natura dell'apparato:

> gli apparati furono inventati per simulare determinati processi mentali. Soltanto ora (dopo l'invenzione dei computer), e per così dire a posteriori, iniziamo a capire che genere di processi mentali simulino tutti gli apparati. […] Tutti gli apparati (e non soltanto i computer) sono calcolatori e, in questo senso, "intelligenze artificiali"[30].

[27] V. Flusser, *Per una filosofia della fotografia*, cit., pp. 31-32, trad. it. modificata.

[28] Si veda anche più avanti, *ivi*, p. 90: «gli apparati sono infatti simulazioni del pensiero, giocattoli che giocano al "pensiero"; e simulano processi mentali non in conformità a quella comprensione del pensiero che corrisponde all'introspezione o alle conoscenze della psicologia e della fisiologia, ma in conformità a una concezione del pensiero così come abbozzata nel modello cartesiano».

[29] Questo punto era già stato esposto con chiarezza da Flusser negli scritti degli anni '60 dedicati alla figura del funzionario, prendendo le mosse da Kafka: cfr. V. Flusser, *Da religiosidade. A literatura e o senso de a realidade*, Escrituras, São Paulo 2002.

[30] *Ivi*, p. 36; cfr. p. 100: «gli apparati sono effettivamente titani antropomorfi, poiché sono stati prodotti con quest'unica intenzione […] essi non sono sovrumani, ma subumani – simulazioni esangui e semplificatorie di processi mentali umani che, proprio in virtù della loro testardaggine, rendono superflue e non funzionali le decisioni umane».

Il passo è molto rivelatore. È ormai evidente che Flusser quando sostiene che un apparato dev'essere una simulazione di un organo del corpo umano sta pensando a un organo in particolare: il cervello[31]. Ed è importante quindi notare che la simulazione per lui paradigmatica del cervello è quella offerta dal computer, quel *black box* che simula il pensiero calcolante e computa bit informazionali generando delle realtà "immateriali". Ne consegue che la sua nozione di apparato è pensata su questo modello, proiettato poi a ritroso su tutti quei dispositivi che, a partire dalla macchina fotografica, riescono a codificare simboli.

Questa simulazione dei processi mentali permette agli esseri umani di trasferire alcuni compiti agli apparati, i quali possono portarli a termine in maniera più veloce ed efficiente, e sempre più spesso riescono ad eseguire operazioni altrimenti impossibili per delle intelligenze umane. Nel caso dei processi di decisione, in cui si tratta di «scegliere una sola conseguenza tra le numerose conseguenze possibili e prevedibili, e scartare tutte le altre»[32], l'operazione è resa possibile dal fatto che tali processi possono essere scomposti in "decidemi" (gli elementi più piccoli in cui può essere ridotto il processo decisionale). Questo tipo di esternalizzazione delle decisioni è illustrato in maniera indimenticabile nel già menzionato episodio del generale McArthur e della sua "macchina-oracolo" che calcola le varie conseguenze e "decide" quali siano il comportamento e la condotta ottimali.

[31] Vale la pena, ancora, di confrontare questo passo con M. McLuhan, *Gli strumenti del comunicare*, cit., pp. 53, 68: «Con l'avvento della tecnologia elettrica l'uomo estese, creò cioè al di fuori di se stesso, un modello vivente del sistema nervoso centrale. […] Inserendo con i media elettrici i nostri corpi fisici nei nostri sistemi nervosi estesi, istituiamo una dinamica mediante la quale tutte le tecnologie precedenti, che sono soltanto estensioni delle mani, dei piedi, dei denti e dei controlli termici del corpo – tutte queste estensioni, comprese le città – saranno tradotte in sistemi d'informazione».

[32] V. Flusser, *Von linearen Entscheidungen zu synthetischen Projektionen*, in «gdi impuls», VII (1989), n. 4, pp. 17-27, qui citato secondo il dattiloscritto conservato nell'archivio, V. Flusser, *Von linearen Entscheidungen zu synthetischen Projektionen*, Vilém Flusser Archive 597, p. 2. Vale la pena di notare che il passo prosegue mettendo in discussione il concetto stesso di decisione all'interno della nuova "immagine digitale del mondo" delineata da Flusser: «tutti i processi sono diventati calcolabili: quelli fisici in particelle, quelli biologici in geni, quelli linguistici in fonemi, quelli culturali in culturemi, le azioni in actomi e (cosa decisiva per questo saggio) le decisioni in decidemi […] finché pensiamo in maniera lineare (finché siamo imprigionati nel codice alfabetico), restiamo prigionieri del contraddittorio, drammatico e addirittura disperato "albero della decisione". Ci dobbiamo perciò decidere di nuovo in ogni istante, senza poter sapere se prima ci siamo decisi in maniera "giusta", e ciononostante non possiamo liberarci dai condizionamenti che ci vincolano. Ma non appena saltiamo dal pensiero lineare a quello calcolante (ad esempio, invece di pensare in testi pensiamo in codici informatici), tutto ciò svanisce. Non ci vediamo più quindi in quelle catene causali che si diramano e si annodano, nelle quali dobbiamo deciderci, ma ci vediamo circondati e immersi in uno sciame di possibilità, delle quali possiamo concretizzarne solo alcune».

4. *Il calcolo della decisione: fine della libertà?*

Nel saggio sulla fotografia il tema dell'esternalizzazione della decisione è rintracciabile nel contesto della critica della cultura, e da lì va estratto e ricostruito[33]. La critica flusseriana ai mezzi di comunicazione di massa e alla tendenza degli individui a trasformarsi in pubblico inerte, soggetto alla manipolazione da parte delle strutture comunicative che producono e distribuiscono informazioni sempre più ridondanti[34], è molto vicina al "pessimismo" francofortese. Flusser non condivide però l'impostazione marxista di quegli autori, che critica dacché essi cercano di smascherare gli interessi dei capitalisti che agirebbero nell'ombra all'interno degli apparati. Con un tono che rasenta lo scherno, Flusser squalifica l'approccio tradizionale alla *Kulturkritik* come una forma di "paganesimo di secondo grado", un tentativo di invocare fantasmi già esorcizzati[35]. Cercando le intenzioni di individui o élite sfruttatrici dietro l'operare degli apparati si contribuisce, al contrario, all'incomprensione della vera natura degli apparati, e quindi a sancirne il dominio e il potere manipolatore. Nell'illustrare il tipo corretto di critica a cui devono essere sottoposti gli apparati, Flusser compie un passo di grande interesse ai fini della presente discussione: riconosce che esiste un'intenzione originaria alle spalle degli apparati, poiché essi «sono stati inventati per funzionare automaticamente, ovvero in modo autonomo rispetto ai futuri interventi umani. Questa è l'intenzione che li ha creati: disinserire l'essere umano da essi. E questo intento ha avuto indubbiamente successo». Ma proprio perché l'intenzione originaria

[33] Flusser affronta la questione della decisione nel mondo degli apparati anche in altri testi molto brevi, come ad esempio V. Flusser, *Dos centros de decisão na década dos 70*, in «O Estado de São Paulo, Suplemento Literário», 31.1.1970, n. 658, p. 1; Id., *Politische Entscheidungen. Essay über das Verschwinden des Großen Staatsmannes*, in «Freitag», 28.6.1991, n. 27, p. 19; Id., *Von linearen Entscheidungen zu synthetischen Projektionen*, cit.; infine un testo probabilmente risalente al 1980 dal titolo *O poder de decidir*, Vilem Flusser Archive 2900. In questa sede si è preferito dedicare attenzione ai testi in cui la questione dell'apparato è articolata con maggiore respiro, fornendo il contesto più adeguato per ricostruire il problema dei processi decisionali esternalizzati.

[34] Si tratta di un circolo vizioso di condizionamento e feedback in cui la libertà umana, che per Flusser è libertà di creare nuove informazioni (differenze che fanno la differenza), finisce per essere soffocata e addirittura spegnersi. Cfr. N.A. Roth, *Out of Language: Photographing as Translating*, in M. Durden, J. Tormey (a cura di), *The Routledge Companion to Photography Theory*, Routledge, London-New York 2020, pp. 398-409; S. Köppl, *Wir müssen die Frage nach der Freiheit neu formulieren. Von der unwürdigen Entscheidungsfreiheit zur Freiheit als Projektion – eine Spurensuche bei Vilém Flusser*, in «Flusser Studies», XVI (2013), pp. 1-14.

[35] Cfr. V. Flusser, *Per una filosofia della fotografia*, cit., pp. 85 e 97-99.

si è realizzata con successo non ha senso cercare in questo stadio del loro sviluppo un'intenzione umana dietro al loro funzionamento. Per la medesima ragione nemmeno si può parlare di qualcuno che li governi. Infatti

> gli apparati funzionano in modo automatico e *non obbediscono ad alcuna decisione umana*, nessuno può possederli. *Ogni decisione umana è presa sulla base di decisioni dell'apparato; essa si è ridotta a decisione puramente "funzionale", ovvero: l'intenzione umana si è volatilizzata* […] gli apparati funzionano ormai come fine a se stessi, "automaticamente" appunto, all'unico scopo di conservare e migliorare se stessi[36].

Questo passo mette in luce un'altra caratteristica degli apparati. In esso vengono contrapposte la decisione umana e la decisione dell'apparato. La prima si riduce a decisione puramente "funzionale", cioè inscritta nei programmi che predeterminano le operazioni nell'apparato. Finisce per essere ricondotta ai processi mentali simulati da queste intelligenze artificiali, laddove non venga addirittura resa superflua, ridondante, o addirittura un disturbo. Gli apparati sono stati infatti progettati per funzionare in maniera automatica[37], cioè il più indipendentemente possibile dall'intervento umano, che viene limitato al feedback che serve all'apparato a migliorare se stesso. Flusser è molto drastico a questo proposito: nella misura in cui opera come funzionario, l'individuo non è un essere umano in senso autentico. Non *ek-siste*, ma funziona solamente. Descritta nei termini di apparati che prendono il sopravvento e relegano il fattore umano a mero serbatoio di feedback, la società postindustriale altro non è che una visione distopica di un futuro che si sta avverando[38]. Un paio di anni più tardi, Flusser avrebbe bilanciato il suo pessimismo, mettendo in luce l'ambiguità degli apparati.

[36] *Ivi*, pp. 99-100, corsivo aggiunto.

[37] In questa prospettiva, la distinzione tra uno strumento e un apparato ricorda quella tra tecnica e tecnologia proposta da A. Fabris, *La filosofia e lo specchio delle macchine*, in «InCircolo», VI (2018), pp. 28-38, p. 33: «gli strumenti tecnici dipendono per il loro uso dall'essere umano e senza l'azione dell'essere umano non sono in grado di esercitare la loro funzione – ragion per cui essi sono sotto il controllo dell'essere umano e da questi dipendono – le nuove tecnologie sono invece capaci di autoalimentarsi, di autoregolarsi, di interagire autonomamente con il rispettivo ambiente. […] nella misura in cui si configurano come veri e propri sistemi, esse sono in grado d'inglobare in sé e di raccordare, per il raggiungimento di uno specifico obbiettivo, non solo vari strumenti tecnici, ma lo stesso agire degli esseri umani».

[38] Come evidenzia F. Krückel, *Bildung als Projekt. Eine Studie im Anschluss an Vilém Flusser*, Springer, Wiesbaden 2015, p. 155: «Diese Programmierung vollzieht sich bis in das Konsumverhalten des Einzelnen. Freie Entscheidungen sind in einer nachmodernen programmierten Welt nicht mehr möglich. Der Mensch wird im Extremfall zum Stereotyp der Apparate, die in der radikalen Form die dialogischen Kreise automatisiert haben».

Nell'opera del 1985 *Ins Universum der technischen Bilder*, forse il libro più ambizioso della fase matura del suo pensiero, Flusser ridisegna l'affresco della società postindustriale in chiave utopica a partire dal concetto di telematica. Con questo termine intende la tecnologia comunicativa che avvicina automaticamente ciò che è lontano, creando una potenziale prossimità globale degli esseri umani che potrebbero così entrare in relazioni dialogiche con chiunque, al di là delle distanze spaziali. Essa presuppone uno sviluppo molto avanzato degli apparati e delle intelligenze artificiali, una condizione che Flusser stava osservando nelle sue fasi nascenti e che oggi noi diamo pressoché per scontata nelle più banali interazioni quotidiane: si tratti di fare una video-chiamata con un amico in un altro continente, di ordinare la cena a domicilio, oppure di guardare una serie tv sul nostro smartphone durante un viaggio. L'esercizio di futurizzazione proposto da Flusser, che spesso utilizza il registro della caricatura e dell'esagerazione ironica per mettere in risalto alcune tendenze del presente che altrimenti resterebbero inavvertite, si concentra con particolare attenzione sul potenziale delle tecnologie che implementano le reti telematiche. Quelle che lui qui chiama "memorie artificiali" sono gli apparati in grado di simulare processi mentali grazie alle loro capacità di calcolo e che possono sostituire in molti ambiti l'essere umano. Flusser osserva che

> la telematica si mostra come una tecnica che permette di sostituire l'essere umano con macchine automatiche […] anche nel processo decisionale. […] già oggi, la maggior parte delle decisioni vengono prese automaticamente, molto prima che la tecnica dell'informatica e il metodo del calcolo delle proposizioni abbiano raggiunto la loro maturazione e molto tempo prima che la telematica sia divenuta effettivamente funzionale. Al punto che, da questa prospettiva, la telematica non sembra tanto una rivoluzione nella produzione dell'informazione, e nemmeno nella preparazione di questa produzione, quanto piuttosto una rivoluzione nella decisione; uno scaricare la coscienza critica dall'essere umano alle macchine automatiche. Fine della libertà[39].

L'amara constatazione con cui si conclude il paragrafo andrebbe letta aggiungendo il punto interrogativo. Davvero non resterebbe alcun margine di libera decisione umana nella società telematica? Elementi per una risposta sono ricavabili da quei passi in cui Flusser discute la possibilità che si affermino globalmente dei "diavoletti di Maxwell", ovvero degli algoritmi concepiti in modo tale da rendere automatica la selezione di in-

[39] V. Flusser, *Immagini*, cit., pp. 165-166.

formazioni non ridondanti all'interno delle reti telematiche. Oggi diremmo che si tratta di attrezzare le intelligenze artificiali in modo tale da affidare a loro il compito di selezionare, all'interno degli ambienti digitali, le informazioni reali, filtrando la "spazzatura" informazionale. In questo scenario utopico, in cui gli apparati hanno sostituito l'essere umano nella fase della "critica", si pone la questione della libertà decisionale. Flusser si chiede se non sia possibile

> automatizzare questa critica presente dappertutto, in modo tale [da poter] risparmiare agli esseri umani di dover verificare secondo il contenuto informativo ogni singola informazione che scorre nella rete […] gli esseri umani sarebbero così liberi di prendere solo le "decisioni decisive", cioè quelle metadecisioni che fanno riferimento alla programmazione dei critici automatici. Secondo la mia opinione, questi sono i […] passi che condurranno ai diavoletti di Maxwell, che si stanno già installando in ogni luogo. Sono passi che muovono in direzione di una libertà sempre maggiore[40].

Flusser, come si può apprendere da questo passo, rovescia in maniera simmetrica la questione discussa sinora. Il timore che affidandosi a degli algoritmi e a delle intelligenze artificiali nei processi di decisione, qui esemplificati dall'attività di filtro e selezione delle informazioni rilevanti e significative, cioè non ridondanti, si arrivi a una nuova forma di schiavitù, sottomissione o rinuncia alla coscienza critica, si tramuta in una speranza. L'esternalizzazione dei processi di decisione alle intelligenze artificiali diventa una forma di liberazione. Infatti non sarebbe più necessario dover decidere. A ciò penserebbero gli apparati.

Flusser precisa però che agli esseri umani resterebbero ad ogni modo le "metadecisioni", le decisioni che contano davvero, riguardanti la "programmazione" delle intelligenze artificiali. Una possibile interpretazione di questa affermazione è la seguente: Flusser intende riferirsi al fatto che le intelligenze umane, a differenza delle intelligenze artificiali, sono capaci di autoriflessione. Sono capaci, cioè, di relazionarsi al proprio agire e stabilire i criteri secondo cui giudicare l'azione. I processi decisionali ordinari, quelli che vengono affidati in maniera liberatoria agli apparati, sono da intendersi come dei "giochi" nei quali le intelligenze artificiali ci superano in velocità, precisione ed efficienza[41]. Ma sono dei giochi, appunto, dai

[40] *Ivi*, pp. 168-169.

[41] Flusser aveva già formulato nelle sue *Vorlesungen zur Kommunikologie* della seconda metà degli anni '70 una posizione simile, ricorrendo però al registro (invero semplificato) della teoria dei giochi: «A differenza del computer, che è programmato per giochi specifici (per quanto

quali questi ultimi non sanno astrarre. Almeno questo sembra essere un modo plausibile di intendere Flusser quando sottolinea, più avanti, che la situazione da lui presentata non implica che

> i critici automatici ci destituiranno dall'essere esseri che decidono. Con il loro installarsi si presenterà infatti una nuova situazione decisionale, mai esistita prima. […] alle intelligenze umane spett[a] necessariamente il diritto di veto, perché esse solamente, e non certo le intelligenze artificiali, sono capaci di dire di "no" a tutto – non perché l'uomo ha fatto tutto ciò, ma perché egli "trascende" tutto ciò, nel senso che è capace di astrarsi da tutto. […] questo diritto di veto, questo diritto a dire "no" è proprio quella decisione negativa che chiamiamo "libertà"[42].

Sulla base di quanto visto finora è allora possibile distinguere almeno tre tipi di decisioni nel contesto della società postindustriale: (1) In primo luogo, vi sono i processi decisionali completamente esternalizzati: si tratta delle decisioni che l'apparato è in grado di calcolare in maniera autonoma e programmata, e in questo senso l'apparato sostituisce totalmente l'essere umano nel processo decisionale. È il caso della piena automazione. (2) In secondo luogo, vi sono i processi decisionali per così dire ibridi: in essi abbiamo a che fare con decisioni funzionali, in cui il soggetto decisore è rappresentato dal complesso apparato-funzionario. In questo caso l'essere umano è coinvolto nella misura in cui coadiuva il processo decisionale che in ultima istanza va ricondotto ai termini delle funzioni prescritte dal programma. (3) Infine, vi è un tipo di decisione che, pur avvenendo nel mondo plasmato e modellato dagli apparati, mantiene un carattere propriamente umano ed esistenzialmente significativo[43]. Questa decisione non riguarda

numerosi), l'essere umano è, in quanto memoria (ovvero il suo cervello, il suo sistema nervoso in generale e probabilmente tutto il suo corpo), programmato in modo tale che i vari giochi si incrociano. […] "Decidere", nel caso del computer, significa perciò soprattutto trovare la strategia migliore per vincere in un gioco, mentre ciò può significare, nel caso dell'essere umano, scegliere tra diversi giochi, assorbire dei rumori, ampliare degli universi. […] al momento sembra possibile, laddove si voglia vincere in giochi specifici, utilizzare i computer nelle decisioni, e quindi riservare alle memorie umane le decisioni nei "metagiochi"», V. Flusser, *Kommunikologie*, hrsg. von S. Bollmann und E. Flusser, Fischer, Frankfurt a.M. 1998, p. 334.

[42] V. Flusser, *Immagini*, cit., p. 169.

[43] Il che viene confermato da un passo delle lezioni francesi degli anni '70 sulla comunicologia: «Dal punto di vista della teoria dei giochi, la parola "decisione" ha due significati. Il primo sta per la possibilità di applicare, entro una data competenza, una combinazione di mosse piuttosto che un'altra, cioè: la decisione di applicare una strategia specifica all'interno di un gioco. Il secondo sta per la possibilità di applicare varie competenze alla stessa situazione, cioè: la decisione di usare giochi differenti nella soluzione dei problemi. In un contesto differente, quest'ultima è ciò che più si avvicina a quella che viene chiamata "decisione esistenziale"»; V. Flusser, *The Surprising Phenomenon of Human Communication*, cit., p. 40.

una "mossa" da fare entro un determinato gioco o programma; in quest'ultimo ambito la tendenza sarà sempre più quella di delegare le decisioni agli apparati, perché superiori in velocità, precisione ed efficienza agli esseri umani. A questi ultimi resta però quella meta-decisione che riguarda la possibilità di dire "no", ovvero di scegliere tra le diverse strutture o programmi, o di trascenderli tutti, in un atto di libertà[44].

5. *Conclusione*

Sicuramente molte delle cose che Flusser ha scritto e pensato sono state messe in questione dagli sviluppi tecnologici più recenti. È soprattutto la sua idea di programma, per certi versi, a risultare troppo rigida rispetto a ciò che oggi le intelligenze artificiali sono capaci di fare. Secondo il suo modello, gli apparati sono in grado di eseguire dei compiti, anche molto complessi, sulla base di una potenza di calcolo e computazione che supera quella degli esseri umani. Essi però sono determinati in anticipo dalle regole e dal repertorio delle loro operazioni possibili; sono, in altre parole, vincolati al programma e alle procedure in essi *embedded*, che per quanto possano essere vasti, devono prima o poi necessariamente raggiungere i propri limiti ed esaurirsi. Si tratta di un agire robotico, spesso ripetitivo e in ogni caso ripetibile, che gestisce i casi imprevisti in maniera procedurale. Le intelligenze artificiali odierne, per contro, si pongono in discontinuità rispetto a questo scenario: sono molto più flessibili, possono apprendere dall'ambiente con cui interagiscono in misura assolutamente superiore agli apparati della generazione precedente, e sono in grado inoltre di anticipare degli ambiti di probabilità[45]. Ma ciò significa forse che l'impianto filosofico

[44] È utile riportare un altro passo dalle già citate *Vorlesungen*, in cui la differenza tra le decisioni funzionali e le metadecisioni viene descritta con chiarezza: «Nel primo caso [*scil.* la decisione del computer] l'output è la conseguenza di una decisione entro una competenza, nel secondo caso [*scil.* la decisione umana] di una decisione tra competenze. Si tratta quindi di due tipi di decisione: nel primo caso viene realizzata una delle possibilità offerte dal gioco, nel secondo caso, invece, viene proposta una nuova competenza. […] il primo tipo di decisione (la strategia di gioco in senso stretto) può essere presa dalle macchine cibernetiche in modo migliore che dagli esseri umani, perché le loro memorie salvano parti più grandi delle competenze di gioco e decidono più velocemente. Per contro, il secondo tipo di decisione (la metastrategia) viene presa in modo migliore dagli esseri umani, proprio perché funzionano peggio delle memorie cibernetiche», V. Flusser, *Kommunikologie*, cit., p. 338.

[45] Si veda G. Tamburrini, *Etica delle macchine. Dilemmi morali per robotica e intelligenza artificiale*, Carocci editore, Roma 2020.

generale del pensiero flusseriano viene messo in crisi? Quello che emerge da quanto ricostruito sinora è che gli sviluppi tecnologici più recenti non abbiano fatto altro che confermare la prospettiva di fondo delineata da Flusser. Che gli apparati odierni siano molto più complessi, adattabili, penetranti e totalizzanti rispetto a quelli di 30 o 40 anni fa depone a favore delle tesi esposte in precedenza sulla natura della società postindustriale. Dal punto di vista esistenziale, che è quello che conta per Flusser, i problemi sembrano essersi solamente accentuati.

Ma c'è un'implicazione del concetto di esternalizzazione che rende il quadro più complesso, e che vale la pena segnalare in conclusione. Nel discutere la tesi della tecnica come prolungamento e simulazione degli organi del corpo, Flusser introduce la nozione di contraccolpo (*Rückschlag* in tedesco)[46]. Si tratta dell'effetto di ritorno che le tecniche, gli strumenti e le tecnologie avrebbero sull'umano nel corso della sua storia. Verso la fine del libro sulla fotografia, Flusser sostiene che «l'essere umano crea utensili prendendo se stesso a modello di questo atto di creazione – fino a quando la situazione non si inverte e l'uomo prende il suo utensile a modello per se stesso, il mondo e la società». L'apparato, abbiamo visto, non è né un utensile, né una macchina. È uno *Spielzeug* che produce simboli, simulando processi mentali. Anche in questo caso vale la dinamica del contraccolpo: specchiandosi nella sua creazione[47], l'essere umano può iniziare a prenderla come modello e assimilarne o imitarne le categorie fondamentali, e a percepire e pensare, cioè, secondo le categorie dell'apparato. Ma ciò significa che, prendendo a modello gli apparati anche nel loro modo di calcolare i processi di decisione, gli esseri umani rischiano di smarrire il senso di quella trascendenza che si manifesta nella "metadecisione", nella possibilità sempre aperta di "dire di no", e inizierebbero davvero a comportarsi roboticamente. Conservare la consapevolezza di questa dimensione decisionale diventa una delle principali sfide etiche e filosofiche della società postindustriale.

[46] Sul concetto di *Rückschlag* ha scritto pagine significative F. Restuccia, *Il contrattacco delle immagini. Tecnica, media e idolatria a partire da Vilém Flusser*, Meltemi, Milano 2021, a cui si rinvia per una trattazione esaustiva.

[47] Sulla questione della tecnologia come specchio, si veda A. Fabris, *La filosofia e lo specchio delle macchine*, cit.

English title: The Externalization of Decision-Making Processes in the Postindustrial Society. Vilém Flusser and the Functionary within the Apparatus.

Abstract

The aim of this paper is to read the increasingly invasive and transformative presence (on a social and individual level) of artificial intelligence and algorithms in decision-making processes in the light of the concept of apparatus as developed by the Czech philosopher Vilém Flusser (1920-1991). This article has a dual nature, historical and theoretical. On one hand, it offers a contribution to the reconstruction of the concept of apparatus by considering some of Flusser's most popular works, Pós-História *(1983),* Für eine Philosophie der Fotografie *(1983), and* Ins Universum der technischen Bilder *(1985). On the other hand, it proposes to use Flusser's categories to interpret some emerging phenomena of postindustrial society as forms of externalization of decision-making processes. The latter entails an increasing deresponsibilization of moral agents and is a phenomenon that, on the anthropological level, can be framed by resorting to the notion of "functionary." Still, according to Flusser, free and existentially meaningful decision-making is possible even in the world of apparatuses, given the specifically human capacity of abstraction.*

Keywords: Vilém Flusser; apparatus; functionary; decision-making; responsibility.

Marco Menon
Università di Pisa
marco.menon@cfs.unipi.it

T

Benedetta Giovanola, Simona Tiribelli

Equità e decisioni algoritmiche

Introduzione

L'enorme sviluppo, negli ultimi decenni, dei sistemi di intelligenza artificiale (IA) e, nello specifico, di algoritmi di *machine learning* (ML) e *deep learning* (DL) ha dato origine a un crescente dibattito nell'ambito dell'etica degli algoritmi[1], volto a mettere in luce, in particolare, i potenziali rischi insiti nei cosiddetti processi decisionali automatizzati, basati – appunto – su algoritmi di ML e DL. Le capacità probabilistiche degli algoritmi di ML e DL nel processare enormi quantità di dati e scoprire modelli e correlazioni preziose hanno comportato un loro utilizzo esteso, dando impulso a un inedito fenomeno di delega a sistemi algoritmici di compiti, scelte e decisioni, prima esclusivamente umani, in ambiti fondamentali, quali l'educazione, la medicina, la giustizia e la difesa nazionale. Tuttavia, questo iniziale entusiasmo, dovuto principalmente alla presunta neutralità, accuratezza e affidabilità dei modelli algoritmici, ha lasciato presto il posto a una serie di critiche sul loro uso nei processi decisionali, poiché gli algoritmi si sono spesso dimostrati difettosi e, soprattutto, iniqui nei risultati generati, piuttosto che esatti e imparziali.

Queste scoperte hanno messo in luce la centralità dell'equità nei processi decisionali algoritmici, stimolando numerose iniziative sul tema, nonché una grande produzione scientifica, di matrice sia filosofica, sia tecni-

[1] A. Tsamados *et al.*, *The ethics of algorithms: Key problems and solutions*, in «AI & Society», 37 (2022), pp. 215-230. Sul tema dell'etica degli algoritmi e, più in generale, dell'intelligenza artificiale, si vedano: A. Fabris, *Etica per le tecnologie dell'informazione e della comunicazione*, Carocci, Roma 2018; L. Floridi, *Etica dell'intelligenza artificiale*, Raffaello Cortina, Milano 2022; P. Benanti, *Human in the loop*, Mondadori, Milano 2022.

 DOI: 10.4454/teoria.v42i2.160

ca[2]. Tuttavia, nonostante l'equità sia riconosciuta come un tema centrale sia nelle riflessioni nell'ambito dell'etica degli algoritmi, sia negli studi più tecnici, il concetto di equità implicato dai processi decisionali algoritmici non è stato ancora indagato in modo adeguato e appare, anzi, piuttosto vago e opaco.

Lo scopo del nostro articolo è colmare questa lacuna, integrando la riflessione sull'equità condotta nell'ambito degli studi sull'etica degli algoritmi e della letteratura tecnica con una riflessione propriamente filosofico-morale sul tema. Nello specifico, mostreremo che un'indagine filosofico-morale sul concetto di equità è necessaria sia per chiarire il significato dell'equità nei processi decisionali algoritmici, sia per individuare i criteri che dovrebbero orientarne la progettazione.

L'articolo è suddiviso in tre paragrafi. Nel primo paragrafo ricostruiamo lo stato dell'arte del dibattito sull'equità nei processi decisionali algoritmici e mostriamo che questo presuppone un concetto di "equità negativa", ovvero di equità come assenza di discriminazione: mostriamo anche che la discriminazione, a sua volta, viene intesa come semplice assenza di distorsioni e pregiudizi (*bias*) nei set di dati con cui processi decisionali algoritmici vengono allenati; sosteniamo, infine, che il concetto di "equità negativa" non è adeguato e argomentiamo la necessità di elaborare un concetto "equità positiva" capace di andare oltre la sola non discriminazione e la considerazione dei *bias*. Nel secondo paragrafo sviluppiamo il concetto di "equità positiva" grazie agli strumenti offerti dalla riflessione filosofico-morale: in particolare mostriamo che equità e non discriminazione non coincidono e individuiamo le dimensioni e componenti costitutive dell'equità. Nella nostra rielaborazione concettuale dell'equità prendiamo le mosse da una originale riflessione sul concetto di rispetto, che riconosce il ruolo del rispetto per le persone in quanto persone, ma include anche il rispetto per le persone in quanto individui particolari. Infine, nel terzo paragrafo mostriamo come la nostra indagine filosofico-morale e la nostra rielaborazione del concetto di equità ci permettano di individuare i criteri che dovrebbero orientare il *design* degli algoritmi, rendendo i processi decisionali realmente più equi.

[2] Si vedano, tra gli altri, C. O'Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*, Crown, New York 2016; D. Shin, Y.J. Park, *Role of fairness, accountability, and transparency in algorithmic affordance*, in «Computers in Human Behavior», 98 (2019), pp. 277-284.

1. *"Equità negativa" e decisioni algoritmiche: non discriminazione e assenza di* bias

L'equità è un tema centrale nell'etica degli algoritmi e, in particolare, nelle riflessioni sui processi decisionali algoritmici[3]. Oltre a essere riconosciuta come un valore fondamentale da integrare nel *design* etico – o *value sensitive* – delle tecnologie basate su sistemi algoritmici[4], l'equità è l'unico tra i principi etici adottati nell'ambito dell'IA a essere riconosciuto in tutti i principali documenti che offrono linee giuda a livello globale per orientare in modo affidabile lo sviluppo degli algoritmi di ML e DL[5].

Questa crescente attenzione per l'equità si spiega anche in risposta a una serie di esiti iniqui prodotti dai processi decisionali algoritmici in vari ambiti, che vanno dalla pubblicità e dal marketing all'accesso al mercato del lavoro e al credito, alla giustizia e alla sanità[6]. Tra gli esempi più rilevanti si possono citare i *bias* di tipo etnico rilevati nell'algoritmo decisionale di COMPAS, un sistema di valutazione del rischio utilizzato nella *criminal justice* statunitense per prevedere il tasso di recidiva degli indagati, denunciato nel 2016 dall'agenzia giornalistica ProPublica perché profondamente discriminante nei confronti degli afroamericani. Un caso simile ha coinvolto, nel 2018, il sistema algoritmico decisionale alla base del software di assunzione utilizzato dalla *big tech* Amazon, rivelatosi discriminante nei confronti dei candidati in base al loro genere, a causa di *bias* presenti nei dati di formazione del sistema. Infine, nello stesso anno, due studiose statunitensi, Timnit Gebru e Joy Buolamwini, hanno denunciato una combinazione problematica di *bias* di genere ed etnici negli algoritmi alla base di alcuni dei software più utilizzati per l'identificazione delle persone tramite riconoscimento facciale. Distorsioni simili sono state scoperte,

[3] A. Tsamados *et alia*, *art. cit.*

[4] S. Umbrello, I. van de Poel, *Mapping value sensitive design onto AI for social good principles*, in «AI Ethics», 1, 3 (2021), pp. 1-14.

[5] Jobin A. *et al*., *Artificial intelligence: the global landscape of ethics guidelines*, in «Nature Machine Intelligence», 1 (2019), pp. 389-399.

[6] C. O'Neil, *op. cit*.; R. Benjamin, *Race after technology: abolitionist tools for the new Jim code*, Polity, Medford 2019. V. Eubanks, *Automating inequality. How high-tech tools profile, police, and punish the poor*, St Martin's Publishing, New York 2018. S.U. Noble, *Algorithms of oppression: how search engines reinforce racism*, New York University Press, New York 2019; B. Giovanola, S. Tiribelli, *Beyond Bias and Discrimination. Redefining the AI Ethics Principle of Fairness in Healthcare Machine-Learning Algorithms*, in «AI & Society», Special Issue "AI4People", (2022), https://doi.org/10.1007/s00146-022-01455-6.

più tardi, anche nel funzionamento di vari sistemi algoritmici utilizzati in ambito sanitario per compiti quali l'identificazione di patologie e l'attribuzione di priorità nell'ordine di accesso dei pazienti a programmi di cura speciali o agevolati[7].

Questi eventi hanno condotto numerosi ricercatori, tecnologi e attivisti a denunciare pubblicamente l'utilizzo dei sistemi basati su processi decisionali algoritmici, accusati di essere strumenti di ingiustizia e, nello specifico, di "discriminazione algoritmica"[8] a causa della loro propensione sia a replicare sia a esacerbare in modo invisibile e silenzioso discriminazioni e pregiudizi, incorporati nella forma di distorsioni o *bias* nei set di dati di formazione e allenamento dei modelli algoritmici.

Ad alimentare ulteriormente le preoccupazioni e le critiche sull'uso dei processi decisionali algoritmici è anche la difficoltà di rintracciare le distorsioni o i *bias* menzionati, soprattutto a causa del basso livello di scrutabilità e/o di intelligibilità degli stessi algoritmi. Alla base di questa difficoltà vi sono due fattori principali: in primo luogo, i modelli algoritmici più utilizzati sono proprietari e, dunque, coperti da segreto commerciale; in secondo luogo, questi modelli spesso includono nel loro funzionamento anche algoritmi di DL, ovvero complesse architetture di reti neurali che, pur consentendo una maggiore capacità di predizione, producono risultati non basati su nessi causali, bensì su correlazioni indotte dai dati, che spesso rendono i processi decisionali algoritmici delle vere e proprie "scatole nere"[9], ovvero modelli opachi e non esplicabili.

La necessità di contrastare e prevenire gli esiti discriminanti prodotti dall'impiego degli algoritmi a fini decisionali ha generato una crescente attenzione sul tema dell'equità, rendendo il *design* e lo sviluppo di processi decisionali algoritmici equi una delle sfide più urgenti e importanti nel settore dell'IA[10]. Tuttavia, nonostante l'importanza dell'equità sia ormai ampiamente riconosciuta[11], il concetto di equità nei processi decisionali

[7] Z. Obermeyer *et al.*, *Dissecting racial bias in an algorithm used to manage the health of populations*, in «Science», 366 (2018), pp. 447-453.

[8] O'Neil, *op. cit*.

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

[10] D. Shin, Y.J. Park, *art. cit*.

[11] J. Kleinberg *et al.*, *Human decisions and machine predictions*, in «Quarterly Journal of Economics», 133, 1 (2018), pp. 237-293; R. Overdorf *et al.*, *Questioning the assumptions behind fairness solutions*, in «NeurIPS», (2018), pp. 1-7; P. Wong, *Democratizing algorithmic fairness*, in «Philosophy &Technology», 33, 2 (2020), pp. 225-244.

algoritmici non è stato ancora indagato in modo soddisfacente e appare, anzi, piuttosto vago e opaco[12].

Analizzando le riflessioni condotte nell'ambito dell'etica degli algoritmi e degli studi tecnici sull'IA, emerge un concetto di equità come assenza di discriminazione e un'accezione di quest'ultima come assenza di *bias*. Considerando le quattro definizioni di equità principalmente adottate nella letteratura sugli algoritmi[13], l'equità è definita tramite metodi di misura matematica come: l'*anti-classificazione*, secondo cui l'equità nei processi algoritmici si ottiene evitando l'uso di termini che si riferiscono a categorie protette (quali l'etnia, la religione e il genere); la *parità di classificazione*, secondo cui un processo decisionale algoritmico è equo se le misure della sua *performance* predittiva sono uguali tra gruppi protetti; la *calibrazione*, secondo la quale l'equità di un sistema decisionale algoritmico è data dalla misura di quanto un algoritmo è calibrato tra gruppi protetti; la *disparità statistica*, secondo cui l'equità di un modello corrisponde a una stima di probabilità media uguale nei risultati per tutti i membri dei gruppi protetti.

Queste definizioni, oltre a rivelarsi incompatibili tra di loro[14], tendono a fornire metriche per misurare l'equità basate sulla considerazione del trattamento da parte del sistema decisionale algoritmico di gruppi o categorie protetti, facendo coincidere, dunque, l'idea di un sistema decisionale algoritmico equo con quella di un sistema non discriminante[15]. La discriminazione algoritmica prodotta dai sistemi algoritmici è a sua volta principalmente ricondotta alla presenza di distorsioni e, nello specifico, a due tipologie di *bias*: i *bias di automazione*, che si verificano quando i processi decisionali algoritmici riproducono su larga scala pregiudizi sociali e culturali incorporati nei dati di formazione del sistema[16]; e i *bias by proxy*, che si verificano quando determinate informazioni inferibili dai dati fungono da *proxy* per l'identificazione di caratteristiche riconducibili a gruppi protetti. Di conseguenza, l'idea diffusa è che un processo decisionale algorit-

[12] N. Saxena *et al.*, *How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness*, in «AI Ethics and Society», (2019), pp. 1-12.

[13] Wong, *art. cit.*

[14] Si veda la critica di Kleinberg *et al.*, *art. cit.*: gli autori sottolineano come la rimozione di alcuni termini che si riferiscono a categorie protette non è sempre auspicabile; si pensi, ad esempio, al settore della salute, dove fattori come il genere e l'etnia svolgono un ruolo cruciale per la predizione di determinate patologie e, di conseguenza, per il *design* di sistemi decisionali algoritmici accurati, oltre che equi.

[15] S. Barocas, A.D. Selbst, *Big data's disparate impact*, in «California Law Review», 104, 671 (2016), pp. 671-732.

[16] Noble, *op. cit.*; Benjamin, *op. cit.*

mico equo, ovvero non discriminante, sia un sistema esente da distorsioni o *bias*[17].

Questa concettualizzazione dell'equità emerge anche nei principali quadri etici elaborati a livello globale per lo sviluppo dei sistemi basati sull'IA, come mostrato da una recente analisi condotta su 84 documenti[18]. Un esempio emblematico sono gli "Orientamenti etici per un'IA affidabile" pubblicati nel 2018 dalla Commissione Europea, che costituiscono uno dei documenti cardine nel panorama internazionale per la progettazione etica dei sistemi algoritmici. Qui il principio di equità è uno tra i cinque principi chiave per il *design* etico dei sistemi basati su algoritmi e prescrive l'impegno ad assicurare l'eliminazione di *bias* che acuiscono forme di discriminazione sociale.

In sintesi, il concetto di equità che emerge dalla letteratura sui processi decisionali algoritmici e dai principali quadri etici di riferimento per l'IA può essere definito come "equità negativa"[19], poiché l'equità è concepita come assenza di discriminazione, e quest'ultima è a sua volta definita come assenza di *bias*. In altre parole, un processo è equo se i suoi risultati ed effetti non producono discriminazione nel trattamento di individui appartenenti a categorie protette e questo è possibile se vengono eliminati i *bias* del sistema algoritmico.

Tuttavia, nonostante la necessità di mitigare i *bias* nei processi decisionali algoritmici sia innegabile, dobbiamo chiederci se la rimozione di *bias*, da sola, possa garantire sistemi algoritmici equi. Per rispondere a questa domanda, dobbiamo chiederci se il concetto di "equità negativa" emergente del dibattito sul tema sia adeguato oppure se non sia necessario sviluppare un'indagine più approfondita su un concetto complesso come quello di equità[20], chiarendone i presupposti teorici[21].

Nel paragrafo seguente svilupperemo questa indagine, offrendo una rielaborazione concettuale dell'equità. Grazie alla riflessione filosofico-morale, proporremo un concetto di "equità positiva" capace di andare oltre la non discriminazione e la considerazione dei *bias* e ne individueremo le

[17] Benjamin, *op. cit.*; Noble, *op. cit.*; O' Neil, *op. cit.*

[18] Jobin *et al.*, *art. cit.*, p. 8.

[19] In questa definizione prendiamo spunto, come è evidente, dalla nota distinzione tra libertà negativa e libertà positiva introdotta da I. Berlin (cfr. I. Berlin, *Quattro saggi sulla libertà*, Feltrinelli, Milano 1989).

[20] A. Rajkomar *et al.*, *Ensuring fairness in machine learning to advance health equity*, in «Annals of Internal Medicine», 16 (2018), pp. 866-872.

[21] R. Overdorf *et al.*, *art. cit.*

dimensioni e componenti costitutive, finora trascurate nel dibattito sull'equità nei processi decisionali algoritmici.

2. *"Equità positiva": equa eguaglianza di opportunità, diritto alla giustificazione, equa eguaglianza di relazione*

Per sviluppare la nostra rielaborazione concettuale dell'equità chiariremo, in primo luogo, la differenza tra equità e (non) discriminazione; argomenteremo poi l'importanza dell'equità positiva, soffermandoci sulle sue dimensioni e componenti costitutive, e mostrando che essa consente di rispettare le persone sia in quanto persone, sia in quanto individui particolari.

Il rapporto tra equità e discriminazione è stato ampiamente riconosciuto dalla riflessione filosofica, specialmente nel contesto delle teorie della giustizia. L'argomento ricorrente è che la discriminazione ostacola l'equità, poiché si fonda sul mancato riconoscimento dell'eguaglianza morale delle persone[22] e implica che alcune di esse vengano trattate in modo crudele o umiliante[23], dunque profondamente irrispettoso. Tuttavia, grazie agli strumenti offerti dalla riflessione filosofico-morale, possiamo evidenziare che l'equità, pur essendo strettamente collegata alla (non) discriminazione, non coincide con questa e include pure altre dimensioni e componenti costitutive[24].

Un primo elemento costitutivo dell'equità è l'*equa eguaglianza di opportunità*, argomentata in modo efficace nelle teorie della giustizia di matrice liberal-egualitaria, a partire da quella Rawlsiana[25]. L'equa eguaglianza di opportunità regola la distribuzione dei benefici e degli oneri della cooperazione sociale e la gestione delle diseguaglianze socio-economiche in modo non solo da prevenire la discriminazione, ma da creare anche le condizioni che consentano l'esercizio dell'agency individuale e l'auto-realizzazione.

[22] Cfr. S. Scheffler, *what is egalitarianism?*, in «Philosophy and public affairs», 31 (2003), n. 1, pp. 5-39; E. Anderson, *What is the point of equality?*, in «Ethics», 109 (1999), pp. 289-337.

[23] A. Sangiovanni, *Humanity without dignity. moral equality, respect, and human rights*, Harvard University Press, Cambridge (MA) 2017.

[24] Per un ulteriore approfondimento di questi temi cfr. B. Giovanola, S. Tiribelli, *Weapons of moral construction? On the value of fairness in algorithmic decision-making*, in «Ethics and Information Technology», 24 (2022), n. 3. 10.1007/s10676-022-09622-5

[25] J. Rawls, *Una teoria della giustizia*, Feltrinelli, Milano 2004. Rawls, come è noto, oltre al principio di equa eguaglianza di opportunità, individua il principio di differenza e il principio di eguale libertà. Non potendoci soffermare su tali principi, in questa sede ne sottolineiamo comunque la coerenza con la nostra discussione sull'equità.

L'equa eguaglianza di opportunità mostra una dimensione distributiva della giustizia, fondata sul bisogno di rispettare le persone sia come destinatarie della distribuzione, sia come soggetti capaci di agency morale.

Un secondo elemento costitutivo dell'equità è il *diritto alla giustificazione*, rivendicato con forza da studiosi come Rainer Forst. Il diritto alla giustificazione esprime la pretesa etica che non vi siano relazioni e strutture intersoggettive "che non possono essere adeguatamente giustificate nei confronti di coloro che vi sono coinvolti"[26]; esso esprime, dunque, un principio di giustificazione reciproca, fondato sull'importanza di rispettare ogni persona in quanto persona, ovvero in quanto soggetto capace di (e titolato a) offrire e richiedere giustificazione. Di conseguenza, la questione del diritto alla giustificazione è anche una questione di potere, ovvero la questione di chi decide cosa[27]. Il diritto alla giustificazione fa emergere una dimensione socio-relazionale dell'equità, la quale mostra sia l'importanza del riconoscimento reciproco, sia la necessità di mitigare le asimmetrie di potere a livello decisionale.

Sia l'equa eguaglianza di opportunità, sia il diritto alla giustificazione sono componenti costitutive dell'equità e ne mostrano, rispettivamente, la dimensione distributiva e la dimensione socio-relazionale[28]. L'individuazione di queste componenti consente di superare l'accezione "negativa" dell'equità come assenza di discriminazione e di mostrare, piuttosto, l'importanza di un'accezione "positiva" dell'equità, fondata sul riconoscimento dell'eguaglianza e del valore morale delle persone e capace di promuovere attivamente l'eguale *rispetto per le persone in quanto persone.*

Tuttavia, il valore morale delle persone non riguarda solo la loro (astratta) capacità di agency morale. Come è stato opportunamente argomentato, esso richiede anche di tenere in considerazione le persone in quanto "individui particolari" [29], che concretamente esercitano la loro agency in modi differenti. Riconoscere questo significa superare l'attenzione esclusiva sull'eguale rispetto per le persone in quanto persone e considerare anche il *rispetto per le persone in quanto individui particolari.*

[26] R. Forst, *Two pictures of justice*, in *Justice, democracy and the right to justification. Rainer Forst in dialogue*, Bloomsbury. London 2014, pp. 3-26, qui p. 6.

[27] *Ivi*, p. 24.

[28] Per un'argomentazione più dettagliata della compresenza di dimensione distributiva e dimensione socio-relazionale dell'equità e, più in generale, della giustizia sociale, cfr. B. Giovanola, *Giustizia sociale. Rispetto ed eguaglianza nelle società diseguali*, il Mulino, Bologna 2018.

[29] R. Noggle, *Kantian respect and particular persons*, in «Canadian Journal of Philosophy», 29 (1999), pp. 449-477.

Un passo in questa direzione può essere rintracciato nella nota distinzione, introdotta da Darwall, tra rispetto come riconoscimento (*recognition respect*) e rispetto come stima (*appraisal respect*): se il primo è fondato sul riconoscimento dell'eguaglianza morale delle persone in quanto persone, il secondo consiste in un "apprezzamento positivo" delle persone "in quanto impegnate in uno specifico compito" o dotate di "caratteristiche che si ritiene manifestino la loro eccellenza"[30]. Valorizzando questa distinzione, possiamo affermare che rispettare realmente le persone richiede *anche* di considerare i "progetti fondativi" che danno senso alla loro vita[31] e li rendono dei sé concreti, piuttosto che astratti[32]. Il riferimento al rispetto per gli individui particolari esprime proprio la necessità di considerare le persone in quanto aventi specifici obiettivi, affiliazioni, valori, impegni, e non solo di trattarle come se fossero "opache", evitando di guardare dentro di loro, e astenendoci "dal guardare oltre l'esteriorità" che esse "presentano a noi in quanto agenti morali"[33]. Comporta, insomma, la necessità di considerare le persone non sono come astratti eguali morali, ma anche come individui che esercitano concretamente la propria agency in modi diversi[34].

Valorizzare il rispetto per le persone in quanto individui particolari consente di individuare un terzo elemento costitutivo dell'equità, che possiamo chiamare *equa eguaglianza di relazione*. L'equa eguaglianza di relazione mette in luce l'importanza delle relazioni nel processo di formazione di obiettivi, affiliazioni, valori, impegni degli individui particolari. Le relazioni, infatti, sono alla base delle nostre affiliazioni e impegni condivisi[35], e questi a loro volta sono centrali per definire i nostri valori e obiettivi[36]. Al contempo va rilevavo che molte relazioni, oggi, sono sempre più mediate

[30] Darwall, *Two kinds of respect*, in «Ethics», 88 (1977), pp. 36-49, qui pp. 38-39, traduzione nostra.

[31] B. Williams, *Persons, character and morality*, in *Moral luck: philosophical papers 1973-1980*, Cambridge University Press, Cambridge 1981, pp. 1-19.

[32] Cfr. M. Sandel, *The procedural republic and the unencumbered self*, in «Political theory», 12 (1984), pp. 81-96.

[33] I. Carter, *Il rispetto e le basi dell'eguaglianza*, in I. Carter, A.E. Galeotti, V. Ottonelli (a cura di), *Eguale rispetto*, Feltrinelli, Milano 2008, pp. 54-77, qui p. 66.

[34] In questa direzione cfr. L. Valentini, *Respect for persons and the moral force of socially constructed norms*, in «Noûs», (2019), pp. 1-24. https://doi.org/10.1111/nous.12319.

[35] M. Gilbert, *A theory of political obligation: membership, commitment, and the bonds of society*, Oxford University Press, New York 2006.

[36] C. Calhoun, *What good is commitment?*, in «Ethics», 119 (2009), n. 4, pp. 613-641. https://doi.org/10.1086/605564.

da tecnologie basate su sistemi algoritmici[37]. Tuttavia queste tecnologie creano spesso bolle[38] o eco camere[39]: basti pensare, a titolo esemplificativo, alle tecniche di personalizzazione alla base dei social media, che spesso tendono a restringere anziché espandere le nostre relazioni, spingendoci verso coloro che sono più simili a noi, limitando così le nostre alternative di scelta e favorendo la creazione di gruppi chiusi, con possibili rischi in termini di estremizzazione, polarizzazione e conflitto[40]. Inoltre, come mostrato dai più recenti studi sulle distorsioni cognitive ed emotive alla base delle nostre motivazioni e convinzioni, percependosi come membri di gruppi chiusi in conflitto con altri gruppi chiusi, gli individui tendono a erodere inconsapevolmente la loro capacità di percepirsi come parte di un progetto condiviso e di avere obiettivi comuni[41].

Questi esempi mostrano che le relazioni possono estremizzare i valori e gli obiettivi degli individui particolari e possono restringere, anziché ampliare, le loro affiliazioni e i loro impegni condivisi. L'equa eguaglianza di relazione è richiesta proprio affinché ciò non avvenga e consiste nel rivendicare con forza l'importanza, per gli individui, di relazioni *genuine*, ovvero radicate in una reale libertà di scelta, che esprime la nostra agency[42] e autonomia. Solo a partire da una equa eguaglianza di relazione, gli individui particolari possono riconoscersi reciprocamente come eguali eppure diversi e rispettarsi in virtù dei reciproci obiettivi e affiliazioni.

La disamina finora condotta ci ha consentito di individuare le dimensioni e componenti costitutive dell'equità, e di fare emergere un'accezione positiva, non solo negativa, di questo concetto, finora ignorata nel dibattito sui processi decisionali algoritmici. Nel paragrafo seguente mostreremo come la nostra rielaborazione del concetto di equità ci permetta di indivi-

[37] B. Giovanola, *Justice, emotions, socially disruptive technologies*, in «Critical review of international social and political philosophy», (2021), pp. 1-16. https://doi.org/10.1080/13698230.2021.18932552021

[38] E. Pariser, *The filter bubble*, Penguin, London 2011.

[39] C. Sunstein, *Democracy and the internet*, in J. van den Hoven, J. Weckert (eds.), *Information Technology and moral philosophy*, Cambridge University Press, Cambridge 2008, pp. 93-110.

[40] Parsell, *Pernicious virtual communities: identity, polarisation and the web 2.0*, in «Ethics and information technology», 10 (2008), n. 1, p. 43.

[41] B. Giovanola, R. Sala, *The reasons of the unreasonable: is political liberalism still an option?*, in «Philosophy and social criticism», (2021), pp. 1-21, DOI: https://doi.org/10.1177/01914537211040568

[42] Valentini, *op. cit.*, p. 7.

duare i criteri che dovrebbero orientare il *design* dei processi decisionali algoritmici, rendendoli realmente più equi.

3. *"Equità positiva" e decisioni algoritmiche: criteri per un* design *etico*

Come abbiamo evidenziato sopra, l'equità comprende tre componenti principali: l'equa eguaglianza di opportunità, il diritto alla giustificazione e l'equa eguaglianza di relazione.

L'*equa eguaglianza di opportunità* indica un criterio fondamentale da rispettare per far sì che i processi decisionali algoritmici garantiscano una distribuzione delle risorse e delle opportunità realmente equa. Soddisfare il criterio di equa eguaglianza di opportunità richiede, dunque, di progettare strumenti compensativi che non si limitino solo a correggere i *bias* nei set di dati di allenamento degli algoritmi, ma che siano pensati e utilizzati per mitigare le disparità sociali esistenti. Questo implica tenere conto nel *design* dei sistemi algoritmici delle profonde disuguaglianze socio-economiche esistenti, integrando tali sistemi con strumenti specifici pensati per compensarle.

Il *diritto alla giustificazione* indica un secondo criterio fondamentale per operazionalizzare l'equità nei processi decisionali algoritmici. Tale criterio richiede il rispetto di ogni persona come soggetto che può offrire e richiedere una giustificazione, ovvero richiede il rispetto dello status di eguale decisore di ogni persona. Come accennato in precedenza, i risultati dei sistemi decisionali algoritmici sono l'esito di processi probabilistici spesso opachi che operano processando enormi quantità di dati al fine di definire correlazioni che permettono la realizzazione di determinati obiettivi in modo efficiente. Sulla base di queste correlazioni, le opzioni alternative disponibili a ogni persona sono pre-determinate in modi che possono minarne le relative possibilità di scelta e azione – e dunque: lo *status* di eguale decisore. Il diritto alla giustificazione, dunque, prescrive la tutela dell'eguale diritto di ogni persona di richiedere una giustificazione per il trattamento decisionale algoritmico a cui è sottoposta e richiede che i *designer* considerino questa richiesta in modi accessibili agli utenti; richiede, cioè, di progettare sistemi in cui le inferenze e/o correlazioni utilizzate per elaborare un determinato risultato siano rese esplicabili in modo tale che le persone soggette a un risultato algoritmico possano esercitare il loro diritto di conoscere, valutare e/o contestare i parametri alla base del risultato del

processo decisionale stesso, mitigando così anche le asimmetrie di potere a livello decisionale.

Infine, l'*equa uguaglianza di relazione* indica un terzo criterio da tener presente nella progettazione di sistemi decisionali algoritmici equi. Questo criterio richiede di tutelare la possibilità di ogni persona di potersi impegnare in relazioni che esprimano la propria capacità di agire in modo genuino, che favoriscano, dunque, lo sviluppo di affetti, impegni, valori e obiettivi genuini. In realtà, la maggior parte delle tecniche attualmente impiegate nei sistemi decisionali algoritmici utilizza metodi di profilazione degli utenti che si basano su macro-correlazioni standardizzanti (quali il filtro collaborativo), ovvero sulla scoperta di macro-caratteristiche comuni tra gli utenti; questi metodi facilitano infatti la categorizzazione di individui diversi in gruppi di persone "simili" ai quali poi proporre contenuti specifici prestabiliti, pre-determinandone così l'esposizione sia informazionale sia socio-relazionale. Tali profili, tuttavia, non considerano e dunque non rispettano le persone come individui particolari, prediligendo la loro considerazione come aggregati probabilistici di caratteristiche comuni. Garantire il rispetto delle persone come individui particolari e, dunque, soddisfare il criterio di equa uguaglianza di relazione richiede un *design* dei sistemi algoritmici capace di combinare l'apprendimento continuo (tipico del ML e del DL) con strumenti volti a favorire l'interazione specifica tra i sistemi decisionali algoritmici e gli utenti, in modo che questi ultimi possano essere informati sulla loro considerazione algoritmica, ovvero sul modo in cui sono profilati e categorizzati, e a loro volta siano nella posizione di informare attivamente il sistema decisionale sui loro reali affetti, impegni, valori e fini e quindi, in altre parole, di partecipare attivamente al funzionamento del sistema nel plasmare la loro esposizione alle informazioni e alle relazioni che sono significative per sviluppare ed esprimere la loro *agency*.

Conclusioni

In questo articolo abbiamo affrontato uno dei rischi più significativi insiti nei cosiddetti processi decisionali algoritmici, ovvero il rischio di essere ingiusti e promuovere risultati iniqui.

Abbiamo preso le mosse da un'analisi critica del concetto di equità emergente nel dibattito sui processi decisionali algoritmici, per proporre poi una nostra rielaborazione concettuale dell'equità, sviluppata grazie agli strumenti della filosofia morale. Abbiamo così fatto emergere l'importanza

del concetto di "equità positiva", di cui abbiamo messo in luce dimensioni e componenti fondamentali. Infine, abbiamo mostrato le implicazioni della nostra ridefinizione dell'equità sui criteri che dovrebbero orientare il *design* di sistemi decisionali algoritmici equi, suggerendo anche alcuni strumenti per implementarli.

Speriamo di aver contribuito, così, a chiarire alcuni presupposti teorici del dibattito sull'equità nei sistemi decisionali algoritmici, che possono rivelarsi utili anche nella concreta implementazione di sistemi decisionali algoritmici equi.

English title: Fairness and algorithmic decision-making

Abstract

The paper focuses on one of the most urgent risks of artificial intelligence, and more specifically of algorithmic decision-making (ADM), that is, the risk of being unfair. In the first section we provide an overview of the discussion on fairness in ADM and show its shortcomings; in the second section we pursue an ethical inquiry into the concept of fairness, and identify its main dimensions and components, drawing insight from a renewed reflection on respect, which goes beyond the idea of equal respect to include respect for particular individuals too. In the third section we show how our conceptual re-elaboration of fairness can help identify the criteria that ought to steer the ethical design of ADM-based systems to make them really fair.

Keywords: Artificial Intelligence; algorithmic decision-making; fairness; ethical design.

Benedetta Giovanola
Università di Macerata
benedetta.giovanola@unimc.it

Simona Tiribelli
Università di Macerata
simona.tiribelli@unimc.it

T

Sofia Bonicalzi

A matter of justice. The opacity of algorithmic decision-making and the trade-off between uniformity and discretion in legal applications of artificial intelligence

1. *Introduction*

In the last few years, decisions about matters of distributive (concerning resource allocation) and retributive (concerning the punishment of lawbreakers) justice have been more and more outsourced to automated systems (A.I.), and unprecedented ethical challenges have progressively emerged. In the realm of retributive justice, the usage of A.I., usually limited to the pre-trial and post-trial phase, ranges from individuating criminals to providing companionship for inmates and allocating cases to specific judges. In the field of distributive justice, A.I. is involved, for instance, in decisions about social housing, access to health care, or career promotions (Jorgensen 2022; Rai 2020; Završnik 2020).

As compared to human adjudicators, A.I. presents, or may present in the future, concrete advantages in terms of efficiency (e.g., time and cost reduction) and uniformity of performance. However, its contribution to legal decision-making must be carefully assessed given its potential ethical drawbacks and impact on basic human rights, such as the right to be tried before an independent tribunal and to access a human rights-based criminal justice system, the presumption of innocence, the respect of privacy, or the right to equal access to public goods and services[1]. This is particularly

[1] For instance, in the U.S. legal system, the requirement that a state governs impartially, and grants people equal protection is imposed by the Equal Protection clause of the Fourteenth

 DOI: 10.4454/teoria.v42i2.161

so since current A.I. systems are progressively moving from auxiliary tools to primary decision-makers, able to impact more directly on people's life by taking decisions and making recommendations and predictions[2].

This paper aims to discuss a specific challenge – the difficult trade-off between uniformity and discretion in judicial applications of A.I. – against the backdrop of current debates in philosophy, cognitive science, and artificial intelligence. This is the gist of it: the usage of A.I. has been notoriously criticized with reference to the so-called *black box problem*, which arises in virtue of the lack of transparency and interpretability characterizing algorithmic decision-making, especially when developed through unconstrained or unsupervised deep learning. Emphasis has thus been placed on how to make the procedures more transparent through forms of explainable A.I. (§ 2). That said, it would be myopic to assume that humans alone are *de facto* best reasoners and transparent deliberators. Cognitive sciences have indeed widely shown that human reasoning is affected by multiple cognitive limitations and biases that might be likewise detrimental to the equity and fairness of the judicial processes. It is not by chance that A.I.-based products are marketed not just as up to mimicking human intelligence but also as in principle able to do better than humans by overcoming common cognitive fragilities. The implementation of A.I. technologies would thus promote a general increase in the uniformity of both procedures and outcomes and cut down pernicious excesses of discretion. A standard reply of those who warn against the potentially despicable effect of A.I. on judicial standards is that this goal is currently out of reach. Indeed, A.I.-based systems are not immune from biases analogous to those that they promise to tame. These anomalies are usually tied to the inclusion of such biases in the set of data through which the algorithm has been trained, to the *under-representativity* of the data, or to wrong assumptions involuntarily made by the programmers (§ 3).

Amendment to the Constitution (Biddle 2020). With expressions like "legal decision-making" and "judicial ruling", here I will refer in general to the activities of the various agencies dealing with distributive and retributive justice broadly conceived. Further work would be needed to discuss in detail the applications of A.I.-based systems to these different domains.

[2] As compared to previously existing supporting tools, «an AI system is a machine-based system that makes recommendations, predictions or decisions for a given set of objectives. It does so by: (i) utilising machine and/or human-based inputs to perceive real and/or virtual environments; (ii) abstracting such perceptions into models manually or automatically; and (iii) deriving outcomes from these models, whether by human or automated means, in the form of recommendations, predictions or decisions» (*Unboxing Artificial Intelligence: 10 Steps to Protect Human Rights*, 2019).

In bracketing the issue of *algorithmic bias*, here I focus on a different argumentative line that stresses the positive value of flexibility and discretion, emphasizing that sidestepping the peculiarities of human reasoning might even have some detrimental effects on the fairness of justice administration. This is particularly the case when this process is conducive to the elimination of reasonable standards of flexibility and discretion, including the ability to bend the rules when circumstances so require. Consequently, the contribution of both human and automated decision-making to social justice matters must be carefully balanced if fair results are to be obtained (§ 4)[3].

2. *The opacity of algorithmic decision-making*

A notorious problem concerning the workings of A.I. is known as the *black box problem*. This refers to the incomprehensibility or lack of transparency regarding how the system moves from the provided inputs to the produced outputs. In some situations, this opacity might be due to contingent issues, such as when the process is protected by trade secret law so that the defendants are not granted a meaningful explanation of the output. A widely debated, and infamous, case of this kind is *Loomis v. Wisconsin*. The defendant (Loomis) – categorized at high risk for recidivism and sentenced to six years in prison and five years of supervision – was denied access to the procedures and methodologies through which the predictive algorithm COMPAS (*Correctional offender management profiling for alternative sanctions*) issued the relevant risk assessment report[4]. This is especially problematic in non-autocratic systems valuing the distribution of decision-making power. Within democratic systems, the defendants, who already find themselves in a vulnerable position, tend to be

[3] Whereas the examples I will refer to belong to the common law tradition and literature, judicial discretion is central even in the systems adopting civil law, whenever general principles must be adapted to specific circumstances. Comparing the role of judges in common law and civil law systems, Yu (1999) suggests that, despite profound differences remain, there has been a progressive convergence between the two so that even «the courts in the civil law countries perform a law-making function through an extension by a flexible process of interpretation and by express legislative instruction. Judicial adaptation to changing circumstances is facilitated by the so-called "general clauses" which leave lots of discretion to the judge»

[4] Loomis appealed the sentencing for violating his *due process rights* (protected, in the U.S. systems, by the Fifth and the Fourteenth Amendment) for various reasons, including that the opacity of the algorithms prevented him from assessing its accuracy and that his right to an individualized sentence was violated (Biddle 2020).

seen as entitled to a comprehensible explanation for them to actively participate in the process and eventually question its output (Završnik 2020).

In other cases, this opacity is more radical to the extent that even programmers themselves may find the process difficult or impossible to interpret or verify, independently of whether the procedures are protected or publicly accessible. On the technical side, various causes, in isolation or combination, may explain this lack of transparency. For instance, the opacity might be related to the system's combining multiple variables that exceed standard cognitive abilities. Or it might depend on the system's relying on complex correlations, statistical modeling, or deep learning techniques that are irreducible to logical reasoning and argumentation (Re and Solow-Niederman 2019). Moreover, technical inexplicability in a descriptive sense may or may not underlie the normative indefensibility of a decision, thus creating a further level of opacity: when descriptive or technical explanations of how a decision has been made are inaccessible, it might become difficult to assess whether the ensuing normative evaluation is fair at all (Selbst and Barocas 2018).

A moment of caution is warranted to the extent that opacity is not limited to A.I. adjudicators. Traditional judicial procedures may also involve some hidden variables, such as when adjudicators rely on anonymous witnesses and undisclosed documentary evidence to make their decisions (see Binger v. King Pest Control). However, the use of undisclosed evidence or witness testimony is usually limited as much as possible and must comply with strict regulatory norms rather than being considered an almost unavoidable part of the adjudication process (Završnik 2020).

Furthermore, the kind of opacity that is typical of automated systems, where procedures are systematically subtracted from public oversight, has been criticized for bringing about a higher risk of alienation of both laypeople and experts. Indeed, a common (although sometimes misleading) assumption is that traditional judicial decisions to some extent mirror human reasoning and its everyday logic. As such, they can be discussed or even challenged via standard argumentative strategies that are within the reach of both experts and novices. From a sociological angle, a radical lack of understanding, especially affecting those who do not have the appropriate technical background, implies vulnerability to the law procedures and generates power imbalances (Re and Solow-Niederman 2019).

From a psychological angle, this issue can be understood as part of the more general discussion about how human reasoning is distinctive-

ly affected by the interaction with A.I. and technologies in general. The reliance on A.I. and other forms of automated systems to carry out daily tasks, inside and outside the legal system, is itself associated with a host of automation-induced distortions in cognition and performance, including skill degradation, automation complacency, and automation bias. Skill degradation refers to the loss of skills that might occur due to the outsourcing of some activities to A.I.[5], which often goes together with phenomena such as automation complacency (the uncritical, passive, and potentially diffident reliance on technologies that are seemingly more competent than the user) and automation bias (the preference for automated solutions that are seen as more reliable than human-based solutions in cases of mismatches or antithetical information) (Parasuraman and Manzey 2010).

Therefore, while inscrutability per se may lead to a general rejection of A.I. adjudicators (Rai 2020), overtrust or loafing may represent the opposite end of the spectrum (Zerilli, Bhatt, and Weller 2022), and even promote a worrisome self-reinforcing circle: the surge in the implementation of A.I. systems may boost skepticism about the practices and competencies of traditional human adjudicators, which may, in turn, increase the social pressure to turn to A.I. solutions to societal problems.

In the last few years, there has been more and more emphasis on how to make the procedures accessible through forms of interpretable or explainable A.I. The goal is to develop methods and processes that can be better understood, and potentially controlled, by humans[6]. On the technical side, the demand for transparency is not easy to address, especially because the improvement in functionality is often afforded by a corresponding increase in complexity. On the theoretical side, this demand requires further articulation, for instance in terms of specifying the level of understanding that is required, the content that must be communicated to different stakeholders, or the amount of information that allows subjects to make more informed decisions without exposing them to an excessive burden (Biddle 2020). Moreover, it is crucial to notice that the effort to generate understandable elucidations may produce fake but intuitive explanations that superficially satisfy the human need for a narrative but are not representative of the

[5] Research has shown that the impact of A.I. on work-related skills may vary depending on the type of job, with a potential increase in skills in high-skill jobs and an opposite effect on low-skill ones (Holm and Lorenz 2022).

[6] For an overview of interpretable models see Hall and Gill 2019; Linardatos, Papastefanopoulos, and Kotsiantis 2020; Rai 2020.

underlying processes (Perez, 2018; Re and Solow-Niederman, 2019). As such, the problem of how to balance complexity and explainability remains an open task for future research.

3. *The fragilities of human reasoning in the legal setting and the algorithmic bias*

While lack of transparency remains an issue in human-computer interaction, cognitive sciences have widely shown that human reasoning can be opaque as well, potentially affecting the equity and fairness of the decision-making process. Decades of literature in cognitive and social psychology have consistently suggested that apparently rational decisions can be surreptitiously determined by automatic or unconscious processes shirking metacognitive reflection (Bargh and Chartland 1999). Cognitive biases of various sorts have been shown to affect professional adjudicators as well, despite their tendency to consider themselves extraordinarily resistant to them in virtue of their training and expertise – and despite their ability to appear unbiased and impartial (Edmond and Martire 2019).

A most emblematic example is provided by discussions about the spread of implicit biases among legal practitioners (Rachlinski and Johnson 2009), stimulating reflections about the role that social cognition research should play in the law (Borgida and Girvan 2015) and the need to take affirmative action to counter discrimination (Kang and Banaji 2006). Implicit biases in particular are tied to forms of unintentional discrimination reflecting problematic social stereotypes that have the potential to distort the ensuing judgment (Holroyd 2015). While decision-makers themselves may fall prey to such biases, the problem is even self-reinforcing in judicial cases that deal directly with discriminatory practices: despite ubiquitous information about the existence of implicit biases[7], the likelihood that adjudicators, even informed ones, will take the offenders' implicit biases seriously when judging their conduct remains low (Girvan, 2015).

In the legal context, cognitive biases of various sorts are particularly worrisome to the extent that the integrity of the legal processes depends on adjudicators' being independent and impartial (see Ebner v Official Trustee) and on their making decisions based exclusively on admissible evidence. As discussed by Edmond and Martire 2019, evidence shows that

[7] But see Macherie 2022 for an extended critique of the science of implicit attitudes.

adjudicators are affected by common cognitive biases, including anchoring effects whereby prior exposure to arbitrary numeric information affects subsequent high or low sentencing decisions (Englich, Mussweiler, and Strack 2006). Furthermore, they are demonstrably sensitive to expectancy effects, e.g., judges' beliefs about the defendants' culpability can be passed to the jurors via mechanisms of nonverbal communication, thus affecting the final verdict (Rosenthal 2003). The impact of contingent, supposedly irrelevant, factors in the judicial ruling has been highlighted by a debated study by Danziger and colleagues (2011), testing the popular saying according to which justice is "what the judge ate for breakfast". The study shows that the percentage of favorable rulings on a judge's typical working day gradually drops during a section of sequential decisions and is restored after the judge takes a break. The suggested explanation is that the ruling effort progressively depletes the judge's executive functions and mental resources so that she is more and more inclined to avoid intervening and to accept the *status quo*, in this case by rejecting the defendant's request.

Given such difficulties and lack of impartiality, it is not surprising that A.I. systems are presented as in principle able to provide valuable support to traditional judicial processes, by overcoming the disturbing variability that is typical of human decision-making. As compared to human decision-making, A.I. is more stable and efficient: it does not decline over time, it is not subjected to contingent environmental influences, and it is never tired. These features secure advantages in terms of reduction of time and costs as well as in terms of uniformity of performance.

A major issue, however, is that it is far from being clear to which extent the A.I. systems currently available in the legal industry can make impartial and unbiased decisions (Dressel and Farid 2018). The problem of algorithmic bias, now officially tackled by some discussed legislative initiatives[8], refers to systematic errors in computer-based systems producing unfair outcomes that benefit given social groups and thus reinforce existing stereotypes (Belenguer 2022). In this respect, a wide-ranging debate was sparked off by the empirical audit run in 2016 by ProPublica (a nonprofit New York-based organization dedicated to investigative journalism), reporting that the recidivism algorithm COMPAS, implemented by Equivant

[8] See, for instance, the report issued by Human Rights Watch (*How the EU's Flawed Artificial Intelligence Regulation Endangers the Social Safety Net: Questions and Answers*) on the European proposed legislation on A.I.

and widely used in the U.S., was biased against black (Angwin, Larson, and Kirchner 2016. See also Dieterich, Mendoza, and Brennan 2016).

While the problem of algorithmic bias remains one of the main challenges for the future of human-A.I. interaction, here I will bracket this issue and focus on a different type of challenge. Indeed, even assuming that the problem of algorithmic bias can be solved, it is unclear that the neutrality and uniformity of results that A.I. can achieve are always positive. In § 4, I will thus discuss this challenge in terms of the trade-off between uniformity and discretion in judicial applications of A.I.

4. *The trade-off between uniformity and discretion in judicial applications of A.I.*

Stressing the alleged neutrality of A.I. adjudicators and the uniformity of their performance implicitly suggests that flexibility and discretion, which are more typical of human intelligence, must be curbed as much as possible. Excess of the bounds of discretion must indeed be condemned, especially so when they cross over into arbitrariness or discrimination. However, flexibility and discretion often play a positive societal role that should not be neglected, particularly when they are not detrimental to social groups that are socially dispossessed – avoiding disadvantaging those who are already disadvantaged is indeed a basic ethical principle that is accepted in most theories of justice (Biddle 2020).

If we look at the common law tradition, a historical distinction is that between *legal formalism* and *legal realism*. Legal formalism maintains that legal decision-making consists in the mechanical and logical application of legal rules and reasons. Conversely, legal realists hold that a host of psychological, contextual, and cultural factors shape the adjudicators' decisions so that they ultimately deliberate based on what seems fair in a given situation (Leiter 2005; Posner 1986). The spirit of the legal realist tradition – caricatured in the study by Danziger, Levav, and Avnaim-Pesso 2011 where the judges' ruling habits depend on when they have had breakfast – is expressed by Oliver Wendell Holmes' famous saying, according to which «the life of the law has not been logic; it has been experience» (Holmes 1881/1991). Whereas this is a descriptive claim about how judicial ruling unfolds in practice, it acquires a normative force to the extent that fairness in adjudication can be often achieved precisely through the exercise of reasonable discretion.

While A.I. systems may achieve better results in terms of uniformity of performance, a specific challenge is thus how to preserve the appropriate amount of discretion in judicial ruling. In this respect, legal scholars have observed that A.I. may affect not just the modalities of legal decision-making (e.g., in terms of time and cost reduction) but the very same values that inform and shape the existing legal culture: «by offering efficiency and at least an appearance of impartiality, AI adjudication will foster a turn toward 'codified justice', that is, a paradigm of adjudication that favors standardization above discretion» (Re and Solow-Niederman 2019, 246).

From a psychological point of view, cognitive sciences have shown that people's subjective sense of distributive and retributive justice cannot be reduced to the mechanical applications of pre-existing rules. More specifically, it cannot prescind from the flexible integration of multiple processes, including both rational and emotional factors. This is evident, for instance, in economic decision-making about equity in resource distribution. In playing the ultimatum game, people notoriously violate the standard norms of rationality when rejecting unfair offers or acting spitefully (Pillutla and Murnighan 1996). Unfair offers are known to generate a conflict between competing tendencies: the emotional one, mediated by the bilateral anterior insula, consists in resisting the offer, while the rational one, mediated by the dorsolateral prefrontal cortex, consists in accepting the offer (Sanfey et al. 2003).

Furthermore, studies about evaluations of procedural and distributive unfairness in resource allocation show a marked dissociation of activation between the two types of judgments, with unfair procedures eliciting greater activation in brain areas concerned with social cognition (ventrolateral prefrontal cortex, superior temporal sulcus) and unfair outcomes eliciting greater activation in areas involved in emotional processing (anterior cingulate cortex, anterior insula, dorsolateral prefrontal cortex) (Dulebohn at al. 2009). Analogously, the subjective sense of retributive justice in third-party scenarios results from the combination of the affective evaluation (in the amygdala, medial prefrontal, and posterior cingulate cortex) of crime severity (how much should the offender be punished?) and the more rational and categorical evaluation (in the dorsolateral prefrontal cortex) of individual responsibility (is the offender responsible or not?) (Buckholtz et al. 2008).

The involvement of affective processes in the evaluation of various forms of distributive and retributive justice should not be viewed simplistically as the result of cognitive biases and distortions affecting human ra-

tionality. Conversely, these psychological mechanisms are fundamental to the formation of our subjective sense of justice and, at least according to the realist tradition, also to how justice works (and perhaps should work) in practice, i.e., in ways that are *responsive* to the agenthood of the person who is being judged rather than being based merely on abstract inferences about one's social identity (Dworkin 1977; Jorgensen 2022).

Therefore, we are now able to better understand what difficult trade-off is at stake: on the one hand, the possibility to get rid of the *bad* variability that is typical of human cognition is extremely valuable – providing that A.I. can overcome the vexed issues of opacity and algorithmic bias. On the other hand, one should not throw the baby out with the bathwater, i.e., one should avoid giving up the good flexibility and discretion that, in humans, depend both on the joint work of different cognitive processes (emotional and rational) and on the flexible adaptation to cultural changes and specific circumstances.

In discussing this challenge, Re and Solow-Niederman 2019 note that the dichotomy between flexible humans and fixed A.I. systems should not be exaggerated[9]: it is contingent and not unavoidable, and, in any case, possible solutions are not free from specific difficulties[10]. For instance, one fascinating option could be that of coding the ability for discretion directly into the A.I. system. This kind of programmable flexibility should reflect the social, moral, and legal consensus on a given topic and evolve in relation to societal changes or unanticipated tasks. This solution, however, poses problems both at a technical and normative level insofar as it is unclear how and when this flexibility can (or should) be implemented, and to what extent forms of good and bad discretion can be so neatly disentangled – this even if one brackets reasonable concerns about the limits of an opaque, black-box based, exercise of discretion.

Alternatively, one may hypothesize a sort of division of labor between human adjudicators and A.I., to be implemented via collaborations in different phases of the judicial process relative to the same cases or by restricting the usage of A.I. to selected cases. Working in tandem with A.I., human adjudicators will play a major role whenever the ability to exert dis-

[9] For instance, in virtue of its ability to consider a higher number of variables, the A.I. could be even more sensitive than humans to the nuances of the situation.

[10] Furthermore, it might be the case that personalization leads to «being treated *worse* than otherwise and is in some tension with other weighty principles of justices, such as the generality and equal application of law, and the fair social distribution of various burdens» (Jorgensen 2022).

cretion looks particularly valuable. However, finding ways to determine the appropriate equilibrium between human and A.I. adjudicators is not simple: the situations in which one wants to avoid bad (i.e., biased) discretion are often the same in which positive (i.e., attuned with the specific situation) discretion is to be praised.

On the practical side, the mechanical and standardized application of existing rules may cause troubles particularly affecting the socially dispossessed, and further exacerbate the very same forms of discrimination it is supposed to fight. Some examples of this, in the domain of both retributive and distributive justice, can be found in Virginia Eubanks' book *Automating Inequality* (2018). Concerning retributive justice, Eubanks quoted a 2000 report of the *Leadership Conference on Civil and Human Rights* taking stock of several *mandatory minimum sentencing laws* enacted in the U.S. in the previous decades and limiting the adjudicators' discretion. Based on the acquired evidence in terms of racial disparity in the outcomes of the criminal justice system, the report states explicitly that «minorities fare much worse under mandatory sentencing laws and guidelines than they did under a system favoring judicial discretion. By depriving judges of the ultimate authority to impose just sentences, mandatory sentencing laws and guidelines put sentencing on auto-pilot». This is paradoxical to the extent that one of the main justifications for automatizing justice is usually to reduce imbalances in the treatment of different social groups.

Regarding distributive justice, Eubanks discusses the impact of algorithms used to allocate social housing based on risk profiles. An emblematic case study is offered by the social housing program *Home for Good*, implemented in 2013 to fight homelessness in the run-down area of Skid Row (Los Angeles). Having the potential to simplify pre-existing processes and potentially limit the impact of the providers' implicit bias, the system was implemented through an assessment tool that collected information and ranked the homeless based on their vulnerability score. Eubanks recognizes that the program successfully managed to help several people with a history of unstable housing. However, some of the interviewed people reported that they were automatically denied help and that the system acted as a black box since no explanation was usually provided about their prioritization score. Moreover, some major issues persist in the way the algorithms were used to track and monitor the poor: people whose behavior and lifestyle were classified (based on the Vulnerability Index Tool VI-SPDAT) as particularly risky or even illegal scored higher on the priority list while being at the same time subjected to higher scrutiny and potentially

face jail time. As such, the program is not just a tool to match the homeless to the housing resources, but a surveillance system aiming to control and criminalize the socially dispossessed while lacking the individualized attention and the ability to bend the rules that were typical of older forms of surveillance and assistance.

To conclude, when judging the functioning of A.I. systems, one should be careful not to confuse different forms of *impartiality*: one thing is to say that algorithms are (potentially) less biased (and thus more impartial) than human adjudicators towards specific social groups. A quite different thing is to wrongly assume that A.I. based decisions can then be automatically impartial also with respect to given ethical, social, and political models and values. Conversely, algorithms remain value-laden, although in ways that, once again, might remain opaque to individual citizens: its functioning reflects a specific trade-off between different interests and values, and ultimately between different societal models and conceptions of fairness, e.g., about how social cooperation must be organized. These differences are embedded, for instance, in the epistemic risks and failures the system is set to tolerate when making decisions about resource allocation or appropriateness of a certain treatment, in the judgments about what problems must be prioritized and require intervention, or in how single factors will weigh in on judicial decisions (e.g., the extent to which a person's socioeconomic background must be considered by risk assessment tools) (Biddle 2020). Therefore, as much as with human adjudicators and laws, relying on A.I. systems requires addressing ethical, and not just technical, difficulties about the societal models that are to be implemented.

5. *Conclusion*

In this paper, I have briefly discussed some ethical challenges linked with the implementation of A.I. systems in judicial decision-making. A.I. seemingly guarantees advantages in terms of uniformity and efficiency of performance, potentially overcoming the typical biases and variability of human adjudicators. However, even assuming that the problem of algorithmic bias can be somehow addressed in the future (and that reasonable choices about the overarching values embedded in A.I. decisional procedures are made), the progressive automatization of the justice system may bring along the neutralization of the good forms of flexibility and discretion that are more typical of human intelligence. This flexibility and discre-

tion have to do with the agent's *right to be treated as an individual*, which goes beyond the right to be judged by an unbiased adjudicator and rather concerns the adjudicator's «duty to be responsive to the individual's responsible agency» or to respect the separateness of persons (Jorgensen 2022). Especially if one is already in a vulnerable position, being treated as an individual includes the right to be part of the decision-making process, which might be further eroded by the opacity of the algorithmic procedures. As such, finding the appropriate balance between uniformity and discretion appears to be one of the major challenges in judicial applications of human-A.I. interaction.

Acknowledgments

I thank the members of the audience of the workshop "Nuove sfide nei processi di decisione. Bioetica, Neuroetica, Etica dell'Intelligenza Artificiale" at the University of Pisa for their insightful comments on an early version of this paper. I am also grateful to an anonymous reviewer for his/her suggestions.

Funding

S.B. has benefitted from the PRIN 20175YZ855.

References

Angwin, J., Larson, J., Kirchner, L. *Machine Bias*. *ProPublica*. May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Bargh, J.A., and Chartland, T.L. "The Unbearable Automaticity of Being." *American Psychologist* 54 (1999): 462-479.

Belenguer, L. "AI Bias: Exploring Discriminatory Algorithmic Decision-Making Models and the Application of Possible Machine-Centric Solutions Adapted from the Pharmaceutical Industry." *AI Ethics* (2022): 1-17.

Biddle, J. "On Predicting Recidivism: Epistemic Risk, Tradeoffs, and Values in Machine Learning." *Canadian Journal of Philosophy*, (2020): 1-21.

Binger v. King Pest Control, 401 So. 2d 1310 (Fla. 1981).

Borgida, E., and Girvan, E.J. "Social Cognition in Law." In *APA Handbook of Personality and Social Psychology*, *vol. 1*, *Attitudes and Social Cognition*, edited by M. Mikulincer, P.R. Shaver, E. Borgida, and J.A. Bargh, 753-774. *American Psychological Association*, 2015.

Buckholtz, J.W., Asplund, C.L., Dux, P.E., Zald, D.H., Gore, J.C., Jones, O.D., and Marois, R. "The Neural Correlates of Third-Party Punishment." *Neuron* 60, no. 5 (2008): 930-940.

Danziger, S., Levav, J., and Avnaim-Pesso, L. "Extraneous Factors in Judicial Decisions." *PNAS* 108, no. 17 (2011), 6889-6892.

Dieterich, W., Mendoza, C., and Brennan, T. *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. Performance of the COMPAS Risk Scales in Broward County*, Northpointe Inc. Research Department. July 8, 2016. https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf.

Dressel, J., and Farid, H. "The Accuracy, Fairness, and Limits of Predicting Recidivism." *Sci Adv.*, 4, no. 1 (2018): eaao5580.

Dulebohn, J.H., Conlon, D.E., Sarinopoulos, I., Davison, R.B., and McNamara, G. "The Biological Bases of Unfairness: Neuroimaging Evidence for the Distinctiveness of Procedural and Distributive Justice". *Organizational Behavior and Human Decision Processes* 110, no. 2 (2009): 140-151.

Dworkin, R. *Taking Rights Seriously*. London: Duckworth, 1977.

Ebner v Official Trustee in Bankruptcy, 205 CLR 337; 2000 HCA 63.

Edmond, G., and Martire, K.A. "Just Cognition: Scientific Research on Bias and Some Implications for Legal Procedure and Decision-Making." *The Modern Law Review* 82, no. 4 (2019): 633-664.

Englich, B., Mussweiler, T., and Strack, F. "Playing Dice with Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making." *Personality and Social Psychology Bulletin* 32, no. 2 (2006): 188-200

Eubanks, V. *Automating Inequality: How High-tech Tools Profile, Police, and Punish the Poor*. New York, NY: St Martins Pr., 2018.

Girvan, J.E. "On Using the Psychological Science of Implicit Bias to Advance Anti-Discrimination Law." *George Mason University Civil Rights Law Journal* 26, no. 3 (2015).

Hall, P., and Gill, N. *An Introduction to Machine Learning Interpretability*. Second Edition. Sebastopol, CA: O'Reilly Media, Incorporated, 2019.

Holm, J.R., and Lorenz, E. "The Impact of Artificial Intelligence on Skills at Work in Denmark." *New Technology, Work and Employment* 37, no. 1 (2022): 79-101.

Holmes, O.W. *The Common Law*. Mineola, NY: Dover Publications, 1881/1991.

Holroyd, J. "Implicit Bias, Awareness and Imperfect Cognitions." *Consciousness and Cognition* 33 (2015): 511-523.

"How the EU's Flawed Artificial Intelligence Regulation Endangers the Social Safety Net: Questions and Answers," Human Rights Watch, last modified November 10, 2021, https://www.hrw.org/news/2021/11/10/how-eus-flawed-artificial-intelligence-regulation-endangers-social-safety-net#_ftn64.

Jorgensen, R. "Algorithms and the Individual in Criminal Law." *Canadian Journal of Philosophy* 52, no. 1 (2022): 61-77.

Kang, J., and Banaji, M.R. "Fair Measures: A Behavioral Realist Revision of 'Affirmative Action." *California Law Review* 94, no. 4 (2006): 1063-1118.

Leiter, B. "American Legal Realism." In *The Blackwell Guide to Philosophy of Law and Legal Theory*, edited by W. Edmundson and M. Golding, 50-66. Oxford: Blackwell, 2005.

Linardatos, P., Papastefanopoulos, V., and Kotsiantis, S. "Explainable AI: A Review of Machine Learning Interpretability Methods." *Entropy* 23, no. 1 (2020): 18.

Macherie, E. "Anomalies in Implicit Attitudes Research." *Wires Cognitive Science* 13, no. 1 (2022): e1569.

Parasuraman, R., and Manzey, D.H. "Complacency and Bias in Human Use of Automation: An Attentional Integration." *Hum Factors* 52, no. 3 (2010): 381-410.

Perez, C.E. *Deep Learning's Uncertainty Principle*. April 6, 2018. https://medium.com/intuitionmachine/deep-learnings-uncertainty-principle-13f3ffdd15ce#:~:text=The%20uncertainty%20principle%20as%20applied,interpretable%20don%27t%20generalize%20well.

Pillutla, M.M., and Murnighan, J.K. "Unfairness, Anger, and Spite: Emotional Rejections of Ultimatum Offers." *Organizational Behavior and Human Decision Processes* 68, no. 3 (1996): 208-224.

Posner, R.A. "Legal Formalism, Legal Realism, and the Interpretation of Statutes and the Constitution." *Case Western Reserve Law Review* 37, no. 179 (1986).

Rachlinski, J.J., and Johnson, S.L. "Does Unconscious Racial Bias Affect Trial Judges." *Notre Dame Law Review* 84, no. 3 (2009).

Rai, A. "Explainable AI: From Black Box to Glass Box." *J. of the Acad. Mark. Sci.* 48 (2020): 137-141.

Re, R.M., and Solow-Niederman, A. "Developing Artificially Intelligent Justice." *22 Stanford Technology Law Review* (2019).

Rosenthal, R. "Covert Communication in Laboratories, Classrooms, and the Truly Real World." *Current Directions in Psychological Science* 12, no. 5 (2003): 151-154.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., and Cohen, J.D. "The Neural Basis of Economic Decision-Making in the Ultimatum Game." *Science* 300, no. 5626 (2003): 1755-1758.

Selbst, A.D., and Barocas, S. "The Intuitive Appeal of Explainable Machines." *Fordham Law Review* 87, no. 1085 (2018).

State v. Loomis, 881 N.W.2d 749, 760-761 (Wis. 2016), cert. denied, 137 S. Ct. 2290 (2017).

"Unboxing Artificial Intelligence: 10 Steps to Protect Human Rights," Council of Europe, last modified May 14, 2019. https://www.coe.int/en/web/commissioner/-/unboxing-artificial-intelligence-10-steps-to-protect-human-rights.

Yu, S.B. "The Role of the Judge in the Common Law and Civil Law Systems: The Cases of the United States and European Countries." *International Area Studies Review* 2, no. 2 (1999): 35-46.

Završnik, A. "Criminal Justice, Artificial Intelligence Systems, and Human Rights." *ERA Forum* 20 (2020): 576-583.

Zerilli, J., Bhatt, U., and Weller, A. "How Transparency Modulates Trust in Artificial Intelligence." *Patterns*, 3, no. 4 (2022).

ORCID ID: Sofia Bonicalzi: 0000-0003-1335-2753

Abstract

In the last few years, decisions about matters of distributive and retributive justice have been more and more outsourced to automated systems (A.I.), and unprecedented ethical challenges have progressively emerged. As compared to human adjudicators, A.I.-based systems present, or may present in the future, concrete advantages in terms of efficiency and uniformity of performance. However, striving for uniformity may also have some sizeable costs. This paper aims to focus on a specific challenge – the difficult trade-off between uniformity and discretion in judicial applications of artificial intelligence – against the backdrop of current debates in philosophy, cognitive science, and artificial intelligence. I will argue that sidestepping the peculiarities of human reasoning might have some detrimental effects on the fairness of justice administration. This is particularly the case when the emphasis on uniformity is conducive to the elimination of reasonable standards of discretion, including the ability to bend the rules when circumstances so require.

Keywords: justice; opacity of algorithmic decision-making; cognitive biases; discretion in judicial decision-making.

Sofia Bonicalzi
Università Roma Tre
Cognition, Values, Behaviour Research Group,
Ludwig-Maximilians-Universität Munich
sofia.bonicalzi@uniroma3.it

T

Veronica Neri

Intelligenza artificiale e scelte di consumo: l'immaginazione come antidoto ai processi di *behavioural bias*

1. *Premessa*

Nel contesto contemporaneo della rete il ruolo crescente dell'intelligenza artificiale (IA) impone alcune riflessioni di ordine etico circa l'orientamento delle scelte dell'individuo e l'autonomia del soggetto di fronte a tali opzioni.

In particolare il presente contributo mira a indagare quali decisioni deleghiamo consapevolmente e/o inconsapevolmente all'IA in un contesto di scelte di consumo e di acquisto (di informazioni, immaginari, valori, beni e servizi) e quali margini di autonomia l'individuo può avere, cercando di preservare la propria capacità valutativa a fronte di strategie algoritmiche che indirizzanso le nostre preferenze. Se oramai l'idea della neutralità dell'intelligenza artificiale, quale mera risultante di calcoli computazionali, appare superata, occorre esplorare se e in che modo l'IA possa chiudere in "bolle di preferenza" o, al contrario, possa consentire un maggiore coinvolgimento morale – incorporando nel proprio sistema determinati valori etici oppure stimolando gli individui a soppesare che cosa possa essere bene o male per sé e/o per la comunità.

Il contributo è strutturato in quattro parti: nella prima parte si inquadra sinteticamente il concetto di IA con particolare attenzione alla possibilità di allineare gli algoritmi attraverso i quali opera ai valori sociali odierni, come alcuni pregiudizi e stereotipi ma, soprattutto, come il consumismo, *deus ex machina* dell'agire contemporaneo; la seconda parte si incentra sulle strategie utilizzate dall'IA per incentivare i consumi e gli acquisti ba-

 DOI: 10.4454/teoria.v42i2.162

sate sul micro-targeting e sul concetto di *nudge*[1]. Si rifletterà pertanto da una parte sulla profilazione dei nostri comportamenti in rete, ponendo attenzione alla *behavioural advertising* e ai *behavioural bias*, dall'altra, sulla la c.d. "spinta gentile" fondata su una architettura delle scelte eticamente indirizzata; nella terza parte si indagano alcune problematiche morali connesse a tali sistemi di profilazione che spingono verso decisioni non usuali per il soggetto, sulla base di preferenze presunte e pregiudizi (sfociando nella c.d. personalizzazione corrotta). La profilazione degli individui in rete consente di prevedere determinati comportamenti e di sviluppare circuiti di obsolescenza (in)controllata *ad hoc*, captando 'razionalmente' le vulnerabilità che spingono l'agire del soggetto. Ciò che apre a questioni relative alla *privacy* e ad un ampliamento del concetto di responsabilità. In conclusione, nella quarta e ultima parte, si rifletterà sulle modalità attraverso le quali l'individuo può arginare tale potere di orientamento degli algoritmi. In specie si cercherà di mostrare come attraverso la capacità immaginativa gli individui possono provare ad adottare strategie di autodifesa rispetto ai sistemi di *machine learning* e di induzione ai *bias* dell'IA. La mancanza di immaginazione potrebbe costringerci in torri d'avorio in cui non siamo autenticamente noi stessi, ma individui sempre più eterodiretti.

2. *Intelligenza artificiale, algoritmi e società del consumo*

Il concetto di IA appare ampio e articolato. Se non esiste una definizione univoca della «scienza dell'artificiale», sotto il profilo etico possiamo sinteticamente considerarla «una riserva di capacità di agire a portata di mano»[2], un «ecosistema complesso, multidisciplinare, dai confini porosi non esattamente definiti, che coinvolge la specie umana e le macchine e le loro interazioni»[3]. Si tratta di una tecnologia che, attraverso calcoli algoritmici[4], decifra autonomamente i dati più salienti qualitati-

[1] R.H. Thaler, C.R. Sunstein, *La spinta gentile*, Feltrinelli, Milano 2014.

[2] Per una disamina approfondita del concetto di IA si rimanda, in specie, a L. Floridi, *Etica dell'intelligenza artificiale*, Raffaello Cortina, Milano 2022, pp. 21-63; cfr. altresì H. Simon, *The Sciences of the Artificial*, MIT Press, Massachusetts 1996, in part. p. 53.

[3] L. Lazzeretti, *L'ascesa della società algoritmica ed il ruolo strategico della cultura*, Franco-Angeli, Milano 2020, p. 26.

[4] Per algoritmo si intendere il costrutto matematico di base dell'IA ovvero «qualsiasi insieme di istruzioni matematiche per manipolare dati o per risolvere un problema». E. Finn, *Che cosa vogliono gli algoritmi. L'immaginazione nell'era dei computer*, Feltrinelli, Milano 2017, p. 23.

vamente che apprende dalla realtà, prevede rapidamente quali nuovi dati potrebbero interessare ai soggetti indirizzando così i comportamenti degli individui[5]. Tale velocità di risposta dell'IA non sempre costituisce però un aspetto eticamente positivo dal momento che indebolisce – fino a cancellare – la possibilità del soggetto di soppesare responsabilmente le ragioni che possono far propendere per determinate scelte di consumo, interrogando la propria coscienza. Il rischio nel quale si può incorrere è dunque, come si chiede Bodei, se «[non] finiremo allora per diventare eterodiretti e [se] non avremo bisogno di aumentare la nostra vigilanza nei confronti di questi cavalli di Troia mentali e di attenerci, per così dire, a una sorta di manuale di autodifesa contro le intrusioni nella sfera dei nostri pensieri, delle nostre immagini e passioni?»[6]. Se conteniamo lo spazio delle scelte dei soggetti entro i confini dettati dagli algoritmi, non esercitiamo una scelta di valore, una "valutazione forte", basata su che cosa sia buono per il sé, vagliando le alternative possibili e richiamando la questione della responsabilità del nostro agire e dei principi morali ad esso sottesi[7].

Gli algoritmi finalizzati a incentivare il consumo online motivano verso determinate azioni di scelta tramite modelli di *machine learning* (ML), la profilazione del pubblico e sistemi di filtraggio. I modelli di ML in particolare, attraverso calcoli algoritmici, elaborano i dati degli utenti sulla base dei loro percorsi di navigazione, permettendo così ai computer di imparare, apprendere ed evolversi in relazione ad analisi automatiche dei propri dati (sonori, visivi, verbali, ecc.) in entrata e in uscita, simulando la capacità cognitiva dell'uomo e, dunque, effettuando previsioni sulla base dei dati[8].

Dal punto di vista etico risulta interessante appuntare l'attenzione sul fatto che la macchina algoritmica apprende anche da dati permeati dai nostri pregiudizi – intesi quali idee e opinioni basati su convinzioni personali, semplificazioni eccessive e prevenzioni generali, a prescindere dall'esperienza diretta o dalla conoscenza, condizionando la propria capacità valutativa –, li rielabora e li restituisce sottoforma di nuove informazioni,

[5] L. Floridi, *Etica dell'intelligenza artificiale*, cit., p. 17.

[6] R. Bodei, *Dominio e sottomissione. Schiavi, animali, macchine, Intelligenza Artificiale*, Il Mulino, Bologna 2019.

[7] C. Taylor, *Che cosa è l'agire umano?*, in Id., *Etica e umanità*, a cura di P. Costa, Vita & Pensiero, Milano 2004, pp. 49-85.

[8] G. Pitruzzella, O. Pollicino, S. Quintarelli, *Parole e potere*, Egea, Milano 2017, pp. 92-93; F. Pasquale, *The Black Box Society*, Harvard University Press, Cambridge Mass. 2020.

o meglio ancora, previsioni su ciò che potrebbe essere di nostro interesse nell'alveo di una società orientata al soddisfacimento del sé e dei propri bisogni (autentici e artificiali) e al consumo.

Come scrive Bauman «La vita dei consumatori, la vita fatta di consumi, non si riduce all'acquisto, e al possesso di qualche cosa. Non si riduce nemmeno al fatto che ci liberiamo di ciò che abbiamo acquistato due giorni fa, e che ancora ieri esibivamo con orgoglio. Ciò che contraddistingue più di ogni altra cosa, semmai, è l'essere in continuo movimento. […]. È illegittimo sentirci soddisfatti»[9]. L'inadeguatezza del sé appare strumento e fine del mercato online per acquisti che immediatamente sollecitano nuovi acquisti. La durata di un bene non è più un valore, vige, di contro l'apoteosi dell'istante. Il consumismo è divenuto fondamento della vita moderna. Hochschild, al tal proposito, afferma che «[i]l consumismo si ripercuote anche sui risvolti emotivi della vita lavorativa e familiare. Esposti al bombardamento incessante della pubblicità […] i lavoratori sono persuasi ad avere "bisogno" di un maggior numero di cose»[10].

I dati sono processati dunque per rispondere ai valori sociali del consumo rapido di informazioni, beni e idee. Nuovi espedienti instillano nell'individuo la necessità di scegliere e suggeriscono che cosa scegliere, pur illudendolo di avere una maggiore padronanza del proprio tempo e della proprie decisioni. Mentre osservano e agiscono, le persone pensano a come fuggire velocemente dal senso di inadeguatezza che li pervade[11]. Il rapporto tra gli individui e gli oggetti e dell'individuo con se stesso appare dunque alterato, attraverso la «tecnostruttura della pubblicità»[12]. Il mercato online promuove una tensione interna nel soggetto, da una parte, generando il timore di essere esclusi e insoddisfatti della propria identità, dall'altra, incoraggiando nuovi bisogni (e quindi possibili soluzioni) per ridefinirla. L'agire online diviene così una fuga dal proprio sé, un rimedio all'angoscia dell'individuo odierno. Gli algoritmi agevolano il processo di obsolescenza del nostro passato per rinascere attraverso nuovi beni di consumo che ridelineano la nostra immagine e promettono (momentanea) sicurezza[13].

[9] Z. Bauman, *Homo Consumens*, Erickson, Milano 2007, p. 24.

[10] A.R. Hochschild, *The Commercialization of Intimate Life*, University of California Press, Berkeley 2003, p. 208 s.

[11] N. Aubert, *Le cultu de l'urgence. La société malade du temps*, Flammarion, Paris 2003, p. 63.

[12] J. Baudrillard, *Il sogno della merce*, Lupetti, Milano 2011, p. 73.

[13] N. Aubert, *Le culte de l'urgence*, cit., pp. 62-63.

Ma tali processi si fondano su dati e schemi di riferimento distorti in partenza, radicati nel proprio immaginario sociale, sulle relazioni deformate e semplificate tra individui e società. Si tratta di un immaginario ricco di rimozioni e di finzioni auto-interessate per rafforzare il sé ideale e, inconsciamente, credere che coincida con il proprio sé reale[14]. Il consumo diventa ludico e, sulla scia di Baudrillard, il «ludico del consumo si è progressivamente sostituito al tragico dell'identità»[15].

Nessuno individuo dichiara online le proprie autentiche preferenze. Afferma di prediligere un genere, ne osserva un altro e, sulla base di impulsi algoritmici, ne osserva un altro ancora e, alla fine, lo predilige[16]. Ciò che avviene, ad esempio, nel caso della «personalizzazione corrotta», un processo che spinge a preferire altro dal genere che rispecchia il sé ideale e il sé reale, rispondendo invece al sé algoritmico. Come scrive Sandvig «Se nel tempo vengono offerti contenuti che non sono troppo allineati con i propri interessi, gli individui possono essere ugualmente orientati rispetto a quello che desiderano. Possono cioè, *in extrema ratio*, credere erroneamente che quelli siano i loro autentici interessi e potrebbero avere difficoltà a vedere il mondo in un altro modo»[17].

3. *Strategie di micro-targeting per indirizzare i consumi*

Per rendere efficaci le proposte di consumo si adottano strategie su base algoritmica quali *behavioural advertising*, a propria volta fondate su *behavioural bias*. Sono strategie intrecciate ai processi di filtraggio e ai sistemi di clusterizzazione, nonchè alla teoria della *nudge*.

Viene meno l'*advertising* tradizionale a favore di forme più opache di orientamento, fondate sul tracciamento dei nostri comportamenti online, di cui non possiamo prevedere le conseguenze. L'identità di un individuo appare come un mosaico, in cui ogni tessera risponde a una traccia che abbiamo lasciato di noi in rete. Ed è sulla base di questa identità presunta e fluida, che l'IA produce previsioni. Al contempo, però, l'individuo consa-

[14] L. Paccagnello, A. Vellar, *Vivere online. Identità, relazioni, conoscenza*, Il Mulino, Bologna 2016.

[15] J. Baudrillard, *La società dei consumi* (1974), Il Mulino, Bologna 2021, p. 236.

[16] E. Finn, *Che cosa vogliono gli algortimi*, cit.; M. Chiriatti, *Incoscienza artificiale*, Luiss University Press, Roma 2021.

[17] C. Sandvig, *Corrupt Personalization*, «Social media collective», 26/6/2014: https://socialmediacollective.org/2014/06/26/corrupt-personalization/ (ultimo accesso 30/6/2022).

pevole non ricerca più il messaggio in sé e per sé quanto il quadro intorno al messaggio, un contesto di «connessioni tra il sistema e il mondo della vita» *à la* Habermas[18] per cercare di comprenderne il senso. In tal modo si generano nuove necessità e preferenze sulla base del variare delle ricerche online. Non si consumano solo beni e servizi, ma dati, valori, idee, opinioni e immaginari.

Ed è in questo contesto che si inscrive la *behavioural advertising*, costruita su una profilazione accurata del pubblico e finalizzata a indirizzare i propri contenuti tramite la conoscenza di abitudini, comportamenti, interessi, bisogni e vulnerabilità degli individui[19].

Per profilazione si intende, sulla scia del nuovo GDPR (2018, art. 4), «qualsiasi forma di trattamento automatizzato di dati personali consistente nell'utilizzo di tali dati personali per valutare determinati aspetti personali relativi a una persona fisica, in particolare per analizzare o prevedere aspetti riguardanti il rendimento professionale, la situazione economica, la salute, le preferenze personali, gli interessi, l'affidabilità, il comportamento, l'ubicazione o gli spostamenti di detta persona fisica»[20]. La pubblicità comportamentale indaga cioè l'agire di ciascun individuo profilato, in base altresì alle sue reazioni di fronte a determinati condizionamenti. È una strategia fondata sul *tracking* dei movimenti in rete dei navigatori, come, ad esempio, il numero di visite su certi siti, le tipologie di ricerche, gli acquisti conclusi, il tempo trascorso online e su quali siti, le attività sui social, ecc. Si recepiscono dati utili per delineare un profilo 'pubblicitario' personalizzato per ciascun individuo online con *advertising ad hoc*; attraverso *filter bubble*, ad esempio, un algoritmo che filtra le informazioni sulla base delle nostre preferenze e in coerenza con i nostri pregiudizi e modi di pensare, recepiti dai nostri movimenti (e tempi) online, indirizzato a chiuderci sempre più nell'immaginario e in quei valori veicolati da certi brand, ad esempio[21]. Ciascun individuo costruisce un ecosistema personalizzato di contenuti, una bolla autoreferenziale, in cui si è poco inclini ad accogliere prospettive diverse, perseverando nella convinzione sempre più radicale della bontà di una certa idea, o di alcuni prodotti. Il fine è l'ade-

[18] J. Habermas, *Teorie dell'agire comunicativo*, I, Il Mulino, Bologna 1981.

[19] A. Dezfouli *et al.*, *Adversarial Vulnerabilities of Human Decision-Making*, in «Proc. Natl. Acad. Sci. USA», 17,117(46), 2020, pp. 29221-29228.

[20] https://www.altalex.com/documents/codici-altalex/2018/03/05/regolamento-generale-sulla-protezione-dei-dati-gdpr.

[21] Cfr. E. Pariser, *The Filter Bubble,* Penguin Books Ltd, New York 2012; A. Bruns, *Are Filter Bubbles Real?*, Polity Press, Cambridge 2019.

sione a valori o idee, a fidelizzare gli individui, a lasciare dati personali, di contatto, ecc.[22].

Sulla base di questi studi sui movimenti degli individui online si osserva che le decisioni individuali sono spesso frutto di distorsioni cognitive e non confermano le previsioni[23]. Il codice del programma può fondarsi su pregiudizi rilevanti dal punto di vista etico, ma alcuni pregiudizi possono essere contenuti anche nei dati stessi da cui il sistema apprende in modo automatico. Tali pregiudizi (*bias*) comportamentali rappresentano deviazioni dal modello che presuppone che le persone si avvalgono di criteri razionali, con preferenze stabili nel tempo e sulla base di un susseguirsi logico di cause-effetto. Il termine *bias* del resto, dal provenzale antico *bias*, assume l'accezione di obliquo, inclinato. Si tratta cioè di una inclinazione verso una direzione piuttosto che verso un'altra. In rete i *bias* di tipo cognitivo, ovvero scorciatoie mentali dalle quali si generano credenze e da cui si traggono decisioni, sono utilizzati per indirizzare l'agire degli individui in contesti dominati dall'incertezza. In specie i *confirmation bias*, processo secondo il quale gli individui preferiscono ricevere informazioni che confermano il loro punto di vista, portandoli a negare qualsiasi evidenza che sia ad esso contraria; nel contesto delle scelte di consumo si affacciano altresì i *behavioural bias*. Si tratta di pregiudizi che inducono comportamenti non razionali da parte dei consumatori. Possono essere di tre tipologie: sulla base di preferenze non standard, quando vengono disattese le ipotesi in funzione di una certa utilità (come, ad esempio, le preferenze che possono sembrare incoerenti nel tempo); di credenze non standard, quando i pregiudizi che emergono in presenza di incertezza violano le ipotesi su come i consumatori formano le convinzioni e, infine, di processi decisionali non standard, ovvero quando le osservazioni dei comportamenti non massimizzano l'utilità di un processo[24]. Questi meccanismi corroborano la neces-

[22] Simili processi avvengono attraverso l'analisi di marcatori quali, ad esempio, i *cookies*, frammenti di dati sugli utenti memorizzati sul *browser*. I siti inviano informazioni personalizzate all'utente in base alle ricerche effettuate e si incentiva il *retargeting*, che consente di recuperare la comunicazione con l'utente che non ha concluso l'azione. M.R. Perugini, *Cookies and Consent: the New Perspectives*, in «European Journal of Privacy Law & Technologies», 1 (2021), pp. 1-37.

[23] Sulle distorsioni cognitive, cfr. R.H.Thaler, *Misbehaving. La nascita dell'economia comportamentale*, tr. it. di G. Barile, Einaudi, Torino 2016.

[24] Cfr. Dowling *et al*., *Behavioral Biases in Marketing*, 48(3), 2020, pp. 449-477; S. Della Vigna, *Psychology and Economics: Evidence from the Field*, in *Journal of Economic Literature*, 47(2), 2009, pp. 315-372. http://www.jstor.org/stable/27739926. Sulla base di questi meccanismi funzionano le chatbot, software che sfruttano l'IA e l'apprendimento automatico per simulare la

sità di una "spinta gentile", ovvero «qualsiasi aspetto dell'architettura di scelta che altera il comportamento delle persone in modo prevedibile senza vietare alcuna opzione o modificare in modo significativo i loro incentivi economici. […]. I pungoli non sono ordini. Mettere frutta al livello degli occhi conta come *nudge*. Proibire il cibo spazzatura no»[25].

Il pubblico diviene destinatario e strumento per indurre certe decisioni. La spinta gentile può assumere però, rispetto ai sistemi di profilazione poc'anzi accennati, finalità etiche, per migliorare la qualità di vita delle persone, influenzando le decisioni senza proibire. Si introduce il concetto di «paternalismo libertario» secondo il quale 'si pungolano' i cittadini ad avere una vita più soddisfacente, seguendo "spinte" che incentivino decisioni volte al loro interesse; si tratta di paternalismo dal momento che c'è una linea chiara da seguire ed è libertario poiché si è liberi di seguirla proprio in un contesto in cui l'individuo appare sempre più vulnerabile[26]. Se, da una parte, si osserva una maggiore incertezza nei confronti della vita ripiegando nella contingenza e nel piacere immediato, dall'altra, l'erosione dei legami comunitari e l'enfasi del sé sul sé rendono l'individuo più fragile e aumenta esponenzialmente la percezione dell'incertezza. La "spinta gentile" si avvale così di "ancoraggi" per indurre all'agire etico, quali la memoria, le emozioni, alcuni stereotipi, la sovrastima di sé, la contrarietà alla perdita e al cambiamento, l'aspetto economico, il conformarsi al gruppo sociale influente, il *priming*, il *feedback* immediato, ecc. Simili ancoraggi possono però promuovere altresì il c.d. "arbitraggio algoritmico", ovvero quella «capacità di organizzare il contenuto attraverso la c.d. personalizzazione corrotta che con la sua macchina culturale orienta e organizza il gusto delle persone spingendo i contenuti, non per quello che sono in sé, ma per la possibile immagine più adatta alla identificazione delle preferenze di ciascuno»[27]. Ciò che è rappresentato si sostituisce al reale senso del messaggio. Il pericolo è quello di passare online da un approccio etico a logiche meramente estetiche e del profitto.

capacità conversazionale di un individuo. Cfr. T. Numerico, *Big data e algoritmi. Prospettive critiche*, Carocci, Roma 2021, p. 33.

[25] H. Thaler, C.R. Sunstein, *La spinta gentile*, cit. p. 9.

[26] D. Cardon, *Che cosa sognano gli algoritmi*, tr. it. C. De Carolis, Mondadori, Milano 2016, p. 76 s.

[27] E. Finn, *Che cosa vogliono gli algoritmi*, cit.

4. *Aspetti morali e indirizzi di scelta*

I processi algoritmici appena delineati aprono ad alcune riflessioni ulteriori sull'etica dei dati. Relativamente alla profilazione degli individui si sviluppano circuiti di obsolescenza controllata. In un contesto in cui lo schermo fa sentire paradossalmente più protetti, l'IA capta le fragilità che spingono l'agire del soggetto indirizzandolo in tempo reale verso prodotti simili a quelli ricercati o rispondenti ai propri gusti, per instillare il desiderio di novità[28].

Un secondo aspetto riguarda, invece, i dati processati dagli algoritmi basati su pregiudizi, portatori di ulteriori pregiudizi, specchio di distorsioni cognitive sedimentatesi nel corso del tempo, che possono sfociare loro malgrado in discriminazioni – come accaduto nel 2019 con un algoritmo impiegato negli ospedali statunitensi per dare assistenza sanitaria, penalizzando persone afroamericane. Una possibile risposta a tali meccanismi è il *redress*, sistema di intervento per casi di errata valutazione da parte dei filtri automatici, come, per esempio, il blocco di contenuti legittimi e la diffusione di dati tendenziosi, offensivi, parziali, ecc.

Come scrive Testa «Per elaborare in fretta enormi quantità di dati, interpretarli e prendere decisioni il più possibile adeguate noi seguiamo delle "scorciatoie mentali" chiamate euristiche. Cioè: semplifichiamo l'elaborazione dei dati procedendo a intuito, o sulla base delle nostre esperienze pregresse. Queste scorciatoie di solito funzionano abbastanza bene. Ma, se invece di ragionare (anche sbrigativamente) partendo da dati di realtà, ragioniamo sulla base di pregiudizi o percezioni fallaci, allora le scorciatoie diventano vicoli ciechi. E le euristiche si traducono in *bias* cognitivi: interpretazioni ingannevoli di dati sballati. In sostanza, potremmo dire che anche la nostra mente si conforma alla regola *garbage in – garbage out* che vale per i computer e l'intelligenza artificiale: se si comincia a ragionare male, il risultato finale è pessimo»[29].

Se l'agire umano forgia gli *input* dell'IA ed è influenzato dai suoi *output* senza soluzione di continuità, può altresì favorire politiche di responsabilità per disincentivare i *bias*. Gli aspetti che ci sfuggono dal controllo sono invece legati al fatto che le risposte derivanti dall'IA contengono in sé una

[28] S.C. Matz *et al.*, *Psychological Targeting as an Effective Approach to Digital Mass Persuasion*, in «Proceedings of the National Academy of Sciences», 114, 48 (2017), pp. 12714-12719.

[29] A. Testa, *Prendere decisioni in un momento difficili*, in «Internazionale», 6 aprile 2020: https://www.internazionale.it/opinione/annamaria-testa/2020/04/06/prendere-decisioni-coronavirus (ultima consultazione 30/6/2022)

certa percentuale di rischio e aleatorietà rispetto alla quale non adottiamo sempre un pensiero critico. Non siamo infatti in grado di intuire i risultati di una architettura da noi programmata. Perdiamo il controllo di ciò che progettiamo e, di contro, tendiamo a deresponsabilizzarci e a delegare alcuni aspetti fondamentali. Occorre, invece, mantenere una responsabilità morale sulle nostre decisioni, ma anche una responsabilità causale e delle conseguenze di chi processa i dati e dei loro fruitori, in uno scambio costante di ruoli.

Un ulteriore aspetto significativo sotto il profilo etico è poi la relazione tra rispetto della *privacy* del consumatore, opacità dei dati e IA. Il paradigma algoritmico pone due questioni. I dati utilizzati sono dati rilasciati volontariamente (e a volte inconsapevolmente) in cambio di informazioni e servizi. La *privacy* non è qui concepita sulla base della dottrina giuridica del *right to be let alone*[30], incentrata sul lato potenzialmente negativo del "render pubblico", dalla quale discende la necessità di approntare una protezione del privato contro la sua pubblicizzazione. I dati in questo caso sono decifrabili solo dai *provider*. Occorre però che la comunità online vigile affinché i dati non vengano utilizzati per discriminare, per mancare di rispetto, per estrarre informazioni utili per mero profitto, per compiere illeciti, ecc. La mercificazione delle informazioni personali confligge con la nostra *privacy* e con lo sviluppo della nostra individualità. Si tratta di dati utilizzati come controprestazione per servizi digitali gratuiti della rete. Occorre però essere consapevoli del potere dei nostri dati, di cui siamo detentori, e, che, grazie al loro valore, anche economico, possiamo decidere se e in che modo cercare di controllare certi meccanismi[31].

Difficile imputare la responsabilità a qualcuno o a qualche cosa, soprattutto se non si tratta di dati sbagliati o offensivi, ma opachi, ambigui non comprendendo se le decisioni emergenti dai sistemi di raccomandazione siano fondati realmente su dati e motivazioni eticamente valide. Questo può ridurre, fino a cancellare, la fiducia nei confronti di tali sistemi.

La responsabilità diviene dirimente da parte dei navigatori nel momento in cui debbono sostenere una scelta, rispondendo ai valori in cui credono e delle proprie opzioni, rispondendo altresì alla tecnologia piuttosto che per la tecnologia. Ma anche da parte delle aziende che debbono comunicare

[30] S. Warren, L.D. Brandeis, *The Right to Privacy*, in «Harvard Law Review», 4 (1890), pp. 193-220.

[31] G. Malgieri, B. Custers, *Pricing Privacy. The Right to Know th Value of Your Personal Data* (2017): https://ssrn.com/abstract=3047257.

quali dati verranno re-impiegati e come verranno ri-elaborati per prendere decisioni. Solo rispettando questi criteri si può tentare di instaurare un processo di fiducia consapevole tra i soggetti in gioco. Sono sistemi che nascono per agire automaticamente, l'individuo non può esercitare un controllo diretto, ma può esercitare il proprio giudizio morale non delegandolo a un sistema che non può, per la sua stessa natura, farsene carico. Non possiamo essere responsabili di ciò che viene programmato per non essere controllato. Occorre ripensare il concetto di responsabilità, la quale agisce nel momento in cui si programma un algoritmo, sulla base del contesto di riferimento e delle persone alle quali si rivolge. La responsabilità è dunque di tutti gli attori coinvolti, fondata sulle motivazioni che inducono le tecnologie a funzionare in determinati ambienti e a cercare di avvalersene senza necessariamente promuovere il proprio allineamento ai sistemi stessi, ma sapendo prendere le distanze in senso critico[32].

5. *Immaginazione e potere algoritmico. Alcune riflessioni conclusive.*

Alla luce di quanto emerso, appare chiaro che gli algoritmi non hanno in sé un'etica quale motore del proprio agire, sono gli individui a trasmettere valore ai dati ricevuti e a ri-trasmetterli, a propria volta, alle macchine. Attraverso quali modalità potremmo allora arginare il potere di orientamento degli algoritmi e garantire scelte di consumo più consapevoli?

Un primo passo può essere la c.d. *ethics by design*, un approccio che studia i principi etici alla base dello sviluppo e del funzionamento delle macchine affinché siano incorporati e, pertanto, rispettati *di default* nei processi algoritmici. Questa approccio se necessario, non appare sufficiente poiché l'IA impara dalle sue interazioni con l'ambiente ed evolve sulla base di ciò che apprende da tali interazioni. Considerata l'elevatissima mole di dati nella quale ci imbattiamo non sempre un processo di

[32] Mantenere il controllo sugli utilizzi dell'IA e sulle sue conseguenze rimane un problema aperto. Attraverso i sistemi di *meaningful human control* (MHC) si tenta in alcuni contesti come quello della difesa, di controllare i processi attraverso i quali l'IA è progettata e opera quando, per esempio, si aprono spazi non gestibili dal sistema o quanto l'IA commette un errore. Il soggetto è solo moralmente responsabile, ma consapevole degli aspetti che può governare e della sua capacità di azione. F. Santoni de Sio, J. Van den Hoven, *Meaningful Human Control over Autonomous Systems: A Philosophical Account*, https://www.frontiersin.org/articles/10.3389/frobt.2018.00015/full (ultimo accesso 30/6/2022); M. Taddeo, *Costruire l'etica dell'intelligenza artificiale*, in G. Fregonara (a cura di) *Il potere del piferraio magico*, Pisa University Press, Pisa 2021, pp. 170-171.

IA fondata sull'*ethics by design* sviluppa soluzioni eticamente orientate[33]. Nei processi algoritmici non operano una coscienza e l'empatia, né quella capacità di immaginazione che consentirebbero di emulare in modo più autentico l'agire umano. Queste 'lacune' ontologiche del sistema non consentono una relazione di autentica reciprocità tra mente umana e 'mente' algoritmica. L'immaginazione umana potrebbe però controbilanciare i meccanismi di induzione al consumo razionalmente pianificati attraverso l'IA. Connessa al concetto di fantasia, facoltà creativa di mediazione «di conservare e riprodurre interiormente le percezioni sensibili, preparando così la memoria»[34], può costituire quella capacità conoscitiva dell'individuo di immaginare e prevedere sulla base della visione – e, in particolare, dell'appena visto – le scelte moralmente più sensate per se stessi e l'"altro", "immaginando" possibili rischi di distorsione della percezione dei dati tali da portare il soggetto a valutazioni errate. Attraverso l'immaginazione l'uomo può utilizzare eticamente l'IA e i suoi risultati – per la tutela di sé e della collettività – per limitare il potere dei sistemi di *machine learning* e i *bias* cognitivi, potendo causare nel tempo discriminazioni sempre più marcate.

Del resto le strategie volte a incentivare i consumi, centrate in modo crescente sul neuromarketing, si avvalgono di immagini, suoni e parole che interagiscono con le nostre emozioni attraverso scorciatoie mentali che possono indurre in errori di ragionamento e a valutazione sbagliate, i c.d. *bias* di scelta. Tali *bias* si sviluppano quando si è sottoposti a un *overload* informativo e non si è in grado di assumere razionalmente una decisione.

Alcuni studiosi affermano che possa esistere una forma di immaginazione dell'IA con la quale l'immaginazione umana potrebbe dialogare. Il concetto di "immaginazione algoritmica" lo si può intendere, riprendendo Finn, come un «pensiero immaginativo per cercare di prefigurare possibili futuri», ma rimane costruito algoritmicamente, tramite calcoli. Lo spazio computazionale dell'immaginazione è quello spazio in cui gli algoritmi «funzionano» e non sapremo mai «come fanno gli algoritmi a sapere quello che sanno»[35]. Più che spazio immaginativo potremmo definirlo uno spazio in cui vigono le leggi algoritmiche su stimoli esterni casuali. Non è possibile prevedere come reagirà, anche emotivamente, un individuo di

[33] Diverso è l'*ethics bluwashing*, processo volto a mettere in atto misure superficiali di indirizzo etico più formali che sostanziali. Cfr. L. Floridi, *Etica dell'intelligenza artificiale*, cit., p. 173.

[34] V. Flusser, *Immagini*, Fazi, Roma, 2009, pp. X-XI.

[35] E. Finn, *Che cosa vogliono gli algoritmi*, cit., p. 198.

fronte a certi dati e quali mosse, anche sulla base della propria immaginazione, metterà in atto e come varieranno i successivi dati in risposta ai suoi *feedback*. Ma si può realmente definire tale processo algoritmico come 'immaginazione'? Se la parola algoritmo deriva dal nome del matematico arabo Muhammad Ibn Musa al-Khuwarizmi e indica una successione di istruzioni per risolvere un problema a partire da un certo numero di dati sembra che non abbia affinità alcuna con l'immaginazione. Anche se gli algoritmi sono volti a incentivare il consumo e aspirano a emulare proprio quello che manca loro, l'immaginazione umana. Potremmo forse recuperare il concetto flusseriano di «tecnoimmaginazione», sulla capacità di decifrare però non solo immagini, ma segni più in generale[36]. I sistemi algoritmici, come gli apparati flusseriani, calcolano per emulare l'immaginazione al nostro posto, senza però riuscirci appieno. Da una parte la fantasia immaginativa non può essere emulata da calcoli numerici, dall'altra l'utilizzo di algoritmi per indurre le scelte mira ad appiattire la nostra stessa facoltà immaginativa, quasi atrofizzandola poiché non ne stimola l'esercizio. Se però adeguatamente educata la nostra capacità immaginativa può controbilanciare la «tecnoimmaginazione» algoritmica alla quale gli individui debbono essere addestrati, proiettando verso mondi alternativi.

L'impiego dell'immaginazione può aiutare ad arginare una possibile deriva della società che si consuma in quanto società stessa dei consumi. Si potrebbe forse allora ipotizzare una immaginazione etica dell'uomo per rispondere alla «tecnoimmaginazione» algoritmica e per non soggiacere al potere della tecnologia?

Il processo immaginativo appare necessario per navigare criticamente nello spazio computazionale. L'utilizzo del pensiero critico e del pensiero creativo al contempo possono mettere a sistema immaginari, immaginazione umana e «tecnoimmaginazione». Si potrebbe pensare a una sorta di immaginazione espansa, sulla scia nuovamente di Finn, che si estende oltre lo spazio della cognizione umana o, ancora, di una immaginazione collettiva che mette in comunicazione gli individui-consumatori per arginare in collaborazione i processi e i sistemi di ML. Gli individui potrebbero sperimentare di riorientare la propria immaginazione che, da soggettiva e puntiforme, può divenire intersoggettiva e sempre più collettiva e sociale. L'IA può forse, se utilizzata con consapevolezza, favorire una sorta di immaginazione «aumentata», fondendo il lavoro di trasformazione dei dati che in-

[36] V. Flusser, *La cultura dei media*, tr. it. T. Cavallo, Mondadori, Milano 2014, p. 87.

dividui e macchine computazionali svolgono insieme[37]. L'immaginazione potrà dunque essere di ausilio all'individuo per rimanere tale e non trasformarsi in «dividuo», come asseriva Deleuze, nel figurarsi quale sarebbe potuta essere l'evoluzione del soggetto nel flusso della società algoritmica[38].

English title: Artificial intelligence and consumer choices: imagination as an antidote to the processes of behavioral bias

Abstract

This contribution investigates which decisions we (un-) consciously delegate to AI in a context of consumption and purchase choices (of information, imaginaries, goods and services) and what margins of autonomy the individual may still have, while trying to preserve their own evaluative capacity. While today the idea of AI neutrality, as a mere result of computational calculations, appears to be outdated, it is necessary to explore whether and how AI can entangle us in clusterings or, on the contrary, allow greater moral involvement. The paper analyzes the micro-targeting strategies used by AI to encourage consumption and purchases and the profiling of our online behaviours, paying attention to behavioural advertising and behavioural bias systems. In conclusion, the author reflects on the ways in which the individual can stem this power of orientation of the algorithms, in particular through the capacity of imagination, typical only of human beings, which can counterbalance the mechanisms of induction to consumption rationally planned through AI.

Keywords: artificial intelligence; bias; consumption; ethics; imagination.

Veronica Neri
Università di Pisa
veronica.neri@unipi.it

[37] E. Finn, *Che cosa vogliono gli algoritmi*, cit., p. 206.

[38] G. Deleuze, *Poscritto sulle società di controllo* (1990), in Id., *Pourparlers,* tr. it. S. Verdicchio, Quodlibet, Macerata 2000, pp. 234-241.

T

Francesca Pongiglione

Trust, experts, and the potential side effects of critical thinking

1. *Becoming informed: A challenging duty*[1]

Humans are called upon daily to make decisions that may impact their own health and that of others, the environment, natural resources, and the well-being of people near or far, both in space and in time. For this reason, acquiring information about the near or distant outcomes of our actions is a civic duty that applies to citizens of the globalized world[2]. Acquiring this information also has a prior status as a moral duty, since it is essential to know the circumstances within which our actions take place in order to determine their permissibility[3]. Such information acquisition is part of what characterizes acting in an epistemically responsible manner. It thus involves an active role on the part of the individual[4].

Due to the technological development that has allowed the widespread diffusion of new media, acquiring information and thus making decisions

[1] I am very thankful to the audience of the Prin 2019-2022 seminar "Etica & Tecnologia: Nuove sfide per l'etica applicata" for their useful comments. This paper has drawn great benefit from the advice received, specifically, by professors Mario De Caro, Adriano Fabris and Massimo Reichlin.

[2] Vanderheiden, S. (2016) The Obligation to Know: Information and the Burdens of Citizenship. *Ethical Theory and Moral Practice* 19, 2: 297-311.

[3] Rosen, G. (2004) Skepticism about Moral Responsibility. *Philosophical Perspectives* 18: 295-313.

[4] Watson L. (2019) *Curiosity and Inquisitiveness*. In H. Battaly (ed.) *Routledge Handbook for Virtue Epistemology*. New York: Routledge; 155-166; Hall R.J., Johnson C.R. (1998) The Epistemic Duty to Seek More Evidence. *American Philosophical Quarterly* 35, 2: 129-139; Hookway J. (1994) Cognitive virtues and epistemic evaluations. *International Journal of Philosophical Studies* 2, 2: 211-227.

 DOI: 10.4454/teoria.v42i2.163

in a reasoned and conscious way seems to be within everyone's reach. Whoever wishes to study a subject in depth has at his or her disposal an enormous quantity of information, quickly and economically available on the internet. Some scholars have therefore suggested that it is no longer possible to justify careless conduct by appeal to ignorance. The increasing availability of information leads to an increase in the epistemic obligations of individuals, so that if before the spread of the internet ignorance or false beliefs on certain topics could be excused, now we can only speak of culpable ignorance[5].

On the one hand, then, it seems entirely reasonable to expect individuals to actively seek out information as part of meeting the "procedural epistemic obligations"[6] that allow us to determine whether our conduct is permissible. On the other hand, it is good to consider the risks inherent in loading individuals with such a burden. The amount of information to be processed, the cognitive resources required to do so, and the time that needs to be devoted to it are substantial, far beyond what is reasonable to expect of the ordinary individual—incurring the real danger that, when faced with such an onerous duty, individuals will give up[7].

However, and equally importantly, even if individuals were to try to fulfill their information acquisition duties, one must keep in mind that the wide dissemination of misinformation, especially in the new media, exposes them to the risk of being misled. The new media are, in fact, an unprecedented epistemic resource, and are one of the most widely used means of searching for information. The web undoubtedly constitutes a resource, but it is an epistemic environment full of pitfalls, where fake news, conspiracy theories, and very well-designed pseudoscientific information proliferate to the point where they can be difficult to distinguish from scientific informa-

[5] Dennett, D. (1986) Information, Technology, and the Virtues of Ignorance. *Daedalus* 115, 3: 135-153; see also Peeters W., Diependaele L., Sterckx S. (2019) Moral Disengagement and the Motivational Gap in Climate Change. *Ethical Theory and Moral Practice* 22: 425-447; and Vanderheiden, S. (2007) Climate change and the challenge of moral responsibility. *Journal of Philosophical Research* 32: 85-92, who apply this argument to the case of climate change.

[6] Rosen 2004, *op. cit*., p. 301.

[7] Hartford, A. (2019) How much should a person know? Moral Inquiry and Demandingness. *Moral Philosophy and Politics* 6, 1: 41-63; Bradford, G. (2017), *"Hard to Know"*, in P. Robichaud and J.W. Wieland (eds.), *Responsibility: The Epistemic Condition*, Oxford, Oxford University Press; 180-198; Guerrero, A. (2007) Don't Know, Don't Kill: Moral Ignorance, Culpability and Caution. *Philosophical Studies* 136, 1: 59-97.

tion[8]. Thus, a set of meta-skills must be developed to avoid being misled by misinformation on the internet.

It is therefore essential to deal with information in a critical manner, carefully assessing the reliability of the sources and in some cases extending vigilance to the content transmitted; too much trust makes us vulnerable[9]. However, critical thinking must also be exercised in the right measure to avoid falling into another frequent error, to which relatively little attention has been paid so far—one which we could define as a substantial misunderstanding of what it means to think critically and relate to information in an epistemically vigilant way. This error consists in adopting an unjustifiably critical attitude that takes the form of downplaying the testimony of experts and not giving it the weight that should be reserved for it[10].

A large part of the debate on this subject has focused on the admittedly complex problem of properly recognizing experts. Identifying experts may not be easy, and many people end up electing the wrong sources as their epistemic authorities. Furthermore, a number of vices (both epistemic and moral) lead individuals to take an attitude of preemptive distrust of experts and their testimony. Some, for example, displaying tendencies to conspiracy thinking, are convinced that experts are so driven by personal interests or corrupted by powerful institutions that their testimony simply represents the view it is convenient for them to hold; for these reasons, they distrust experts a priori. Others, manifesting a vice that I have elsewhere called "epistemic hybris"[11], think they can easily replace experts, perhaps by doing some research on the web. These individuals consistently reserve the right to investigate matters on their own even where they utterly lack the expertise to do so. In both these cases, individuals elect sources other than official experts as their epistemic authorities.

[8] Pongiglione, F., Martini, C. (2022) Climate change and culpable ignorance: the case of pseudoscience, published online: https://doi.org/10.1080/02691728.2022.2052994; Thi Nguyen, C. (2020) Echo chambers and epistemic bubbles. *Episteme* 17, 2: 141-161; Croce, M., Piazza, T. (2019) Epistemologia della fake news. *Sistemi Intelligenti*, 31, 3: 439-468; Millar, B. (2019) The Information Environment and Blameworthy Beliefs. *Social Epistemology* 33, 6: 525-537; Rini, R. (2017) Fake News and Partisan Epistemology. *Kennedy Institute of Ethics Journal*, 27, 2: 43-64.

[9] Levy, N. (2022) In Trust We Trust. *Social Epistemology*, published online: https://doi.org/10.1080/02691728.2022.2042420.

[10] Grundmann, T. (2021) *Facing Epistemic Authorities. When Democratic Ideas and Critical Thinking Mislead Cognition*. In S. Bernecker, T. Grundmann, A.K. Flowerree (eds.) *The Epistemology of Fake News. Oxford: Oxford University Press*; 134-155.

[11] Pongiglione 2022, manuscript under review.

However, it would be hasty to think that the problem lies only in the correct identification of epistemic authority. In fact, even when experts have been correctly identified, it is not a given that individuals know how to relate to their testimony in the right way. In fact, experience has shown that even once "real" experts have been identified, mistakes can still be made in how one relates to their testimony.

In what follows, I will show how the intention to exercise critical thinking sometimes leads to an excess of distrust and suspicion improperly extended even to experts recognized as such by the scientific community and by the individual herself. If a passive or compliant attitude risks allowing the subject to fall into error, so does an excessively critical attitude. Therefore, particular attention will be paid to the need to redefine the role of experts in order to establish a relationship with them that is neither one of passive subordination nor one of unmotivated distrust. Finally, it will be shown how a correct relationship with experts also passes through the exercise of a particular virtue—intellectual humility. In fact, it is this virtue that, by giving individuals the ability to recognize their own competence and epistemic limits, puts them in a position to assign the right weight to expert testimony, especially in relation to their own beliefs as non-experts.

2. *Epistemic vigilance or unjustified distrust?*

Our epistemic duties as citizens of the global world require us to seek information to ensure that our actions do not harm others or ourselves. As we do so, however, we should not passively accept everything we are told without thinking it through—without ensuring, at the very least, that the sources we rely on are trustworthy. In fact, every communicative exchange presents risks; for this reason, our relationship with information sources must always be managed with attention and a critical eye. Not only are there people who intentionally try to deceive us, but there are also many who spread false, biased, or otherwise incorrect information entirely in good faith. Accordingly, the risk of being misled in the process of exchanging and acquiring information is high[12].

[12] Levy 2022, *op. cit.*, pp. 1-2; Sperber D., Clément F., Heintz C., Mascaro O., Mercier H., Origgi G., Wilson D. (2010) Epistemic Vigilance. *Mind & Language* 25, 4: 359-393; pp. 359-360.

That blindly trusting a source, no matter how authoritative it seems or is said to be, is not an epistemically sound strategy is widely agreed upon in the literature.[13] Blindly trusting does, in fact, break the minimal rules of rationality. To be sure, we need to trust others when we make decisions in domains where we do not ourselves have expertise—something that happens on a daily basis. But this trust need not be granted blindly. It can start with a check of the reliability of the source itself via word of mouth, references, or titles[14], and it can continue with a closer examination of various elements that can confirm that our trust is well placed[15].

We often lack expertise on the topics about which we need information and are therefore unable to evaluate the quality of that information. In these cases, we can focus on the reasons for trusting a specific source, assessing for conscientiousness and accuracy; this assessment should provide us with prima facie reasons for trust[16]. We then need to consider the reliability of the source in the specific context of the information we need[17]: for example, a doctor may be a good source of information about a vaccine but not about repairing a washing machine.

Some scholars emphasize the need to be vigilant as well with respect to the content of the information itself, even when we lack the expertise to make a sound evaluation. We can be vigilant about information content by assessing it both for internal consistency and for consistency with our prior beliefs[18]. This leads to a more critical attitude toward the testimony of others. Not all scholars agree that the latter type of evaluation is necessary when testimony comes from experts; according to some, once experts

[13] Baghramian, M., Panizza, F. (2022) Scepticism and the Value of Distrust. *Inquiry: An Interdisciplinary Journal of Philosophy*; Grundmann 2021, *op. cit.*; Lackey, J. (2018) Experts and Peer Disagreement. In M.A. Benton, J. Hawthorne, D. Rabinowitz (eds) *Knowledge, Belief, and God. New Insights in Religious Epistemology*. Oxford: Oxford University Press, 228-245; Lynch, M.P. (2016) *The Internet of Us. Knowing More and Understanding Less in the Era of Big Data*. New York: Liveright Publishing Corporation; Zagzebsky, L. (2012) *Epistemic Authority: A Theory of Trust, Authority, and Autonomy in Belief*. Oxford: Oxford University Press; Sperber et al. 2010, *op. cit*.

[14] Lynch 2016, *op. cit.*, pp. 42-43.

[15] Martini, C. (2020) *The Epistemology of Expertise*. In M. Fricker, P.J. Graham, N.J.L.L. Pedersen, D. Henderson (eds.) *The Routledge Handbook of Social Epistemology*. New York: Routledge, 115-122.

[16] Zagzebski 2012, *op. cit.*, pp. 57-58.

[17] Sperber et al. 2010, *op. cit*.

[18] Lackey 2018, *op. cit*; Sperber et al. 2010, *op. cit.*

have been identified[19], they can be trusted even if the information they convey conflicts with our non-expert opinions[20].

However, the great attention devoted to preventing the epistemic misstep of blind trust has somewhat overshadowed the substantial risk of falling into the opposite excess and ending up adopting an unjustifiably distrustful[21], or at least overly skeptical, attitude towards information, with the result that one extends doubt even to valid content, thus losing sight of the truth (Pritchard 2021, p. 63). A form of distrust can certainly be part of an appropriately critical approach to information. However, once reliable sources have been identified through the exercise of epistemic vigilance, one must know how to relate to their testimony properly—something that is entirely compatible with the exercise of vigilance even over the content transmitted, if this is considered appropriate by the subject.

The relationship of the ordinary person to the expert is not one of epistemic parity, such that the individual reasons of one are equivalent to the testimony of the other. For this reason, the opinion of experts cannot be regarded as just another source of reasons and opinions, to be compared on an equal basis with other opinions coming from non-experts and finally with one's own. To do this would be tantamount to not giving the expertise itself any weight. If it is irrational to trust blindly, it is equally irrational to treat epistemic authority as just another source[22].

Sometimes, in fact, the subject confronting the opinion of experts with respect to a given issue has prior beliefs or a personal inclination regarding the issue. This is the case, for example, for individuals who, although aware of the existence of climate change, have a strong interest in downplaying the threat it poses so as not to feel compelled to change their dai-

[19] It is assumed here that the individual has correctly identified the epistemic authority through the practice of vigilance. Of course, one can also be mistaken in this identification process, and this can occur for a variety of reasons —unwarranted skepticism toward official sources (as suggested in Cassam, Q. (2016) Vice Epistemology. *The Monist* 99: 159-180), pseudoskepticism (an epistemic vice close to conspiracy thinking, analyzed in Torcello, L. (2016) The Ethics of Belief, Cognition, and Climate Change Pseudoskepticism: Implications for Public Discourse. *Topics in Cognitive Science* 8 (1): 19-48), or even the adoption of a novice-oriented conception of epistemic authority, which disposes one to assign authority more readily to those who do not deserve it (Croce, M. (2019) On What it Takes to Be an Expert. *Philosophical Quarterly* 69, 264: 1-21).

[20] Constantin, J., Grundmann, T. (2020) Epistemic authority: preemption through source sensitive defeat. *Synthese* 197: 4109-4130; Grundmann 2021, *op. cit*.; Zagzebski 2012, *op. cit.*

[21] Audi, R. (2011) The Ethics of Belief and the Morality of Action: Intellectual Responsibility and Rational Disagreement. *Philosophy* 86: 5-29; p. 9.

[22] Constantin and Grundmann 2022, *op. cit.,* p. 4110; Grundmann 2021, *op. cit.,* p. 140.

ly choices or actions, and who therefore begin to collect information from the most disparate sources, including unofficial ones, in order to reinforce their view of things (as in the instance described in Robichaud)[23]. In their evaluations, such individuals make the error identified by Grundmann: they treat the opinion of experts as one voice among many, which they then compare on an equal basis to other opinions, not necessarily from experts, and finally to their own, thus indulging in the *Principle of Democratic Reasoning*, the principle that says not to exclude, diminish, or marginalize the weight of any rational person's reasons, including one's own, in assessing the truth of a given proposition—even when one has no expertise to comment on it[24]. The same mistake is made by those who, having learned from the medical community that vaccines against Covid-19 are safe and bring more benefits than harm even to the individual, prefer to follow their own inclination not to be vaccinated, thus assigning a greater weight to their own opinion than to that of the experts (even while recognizing them as such)[25].

This attitude of excessive and misleading criticism of experts can be traced back essentially to two issues: first, a lack of understanding of the role of experts and how they can, and in some cases should, guide our choices; and second, intellectual dispositions such as presumption or overconfidence in one's own abilities and skills. It is therefore necessary to define the correct way to relate to experts and their testimony and to specify the intellectual virtues that are useful in fostering this relationship.

3. *Trust in experts and intellectual humility*

While identifying incorrect ways of relating to expert testimony may seem relatively simple, determining the proper weight to give it is much more complex. One must keep in mind that even experts make mistakes,

[23] Robichaud, P. (2017) Is ignorance of climate change culpable? *Science Engineer Ethics* 23: 1409-1430.

[24] Grundmann 2021, *op. cit.*, p. 137.

[25] The reason for this specification is that there are individuals who, skeptical of official institutions and the information they convey, do not regard official experts as epistemic authorities at all, instead assigning epistemic authority to others according to non-objective criteria such as personal inclination or sympathy. Although this too constitutes an epistemically incorrect attitude, it has its own distinctive characteristics, which this essay does not address. Here we are focusing instead on attitudes towards experts the individual recognizes as such, thus assuming that experts have been correctly identified (clearly no small assumption).

that they often have opposing views on the same issue, and that therefore a strategy of deference cannot always ensure epistemic success.

Several theoretical proposals have been advanced to address this problem. Some scholars argue that expert testimony should provide preemptive reasons to trust it in preference to, for example, one's own beliefs, opinions, or intuitions. This proposal, called by Grundmann the *Preemptive View* (PV), is rooted in the idea that the proper attitude to reserve for epistemic authority is indeed one of deference because of the greater likelihood of arriving at the truth by relying on those with objective expertise in a given domain[26]—the so-called *Track Record Argument*[27].

The Preemptive View has received several criticisms. One is that deference to epistemic authority cannot occur unless reliable experts are available; yet finding reliable experts can be challenging and raises additional problems in domains where experts disagree with each other[28]. This objection has motivated the *Total Evidence View* (TEV), which treats expert opinion as one more piece of evidence to be added to and weighed against the others available to the subject, without preemption (this is the proposal of Lackey, *op. cit.*). But even the TEV is not immune to criticism. Suppose we have individuals who are not only incompetent in a certain domain but also unaware of their incompetence. In this case, by adopting TEV they would end up assigning their dubious judgment a weight that it should not have, being likely to lead to erroneous conclusions[29].

Recently, Levy and Savulescu have made a theoretical proposal that can be considered a moderate version of the PV, avoiding some of its main weaknesses. The idea is to recommend deference to epistemic authorities only in cases where there is evidence on which the scientific community converges. Their suggestion is that when the opinions of multiple experts tend to produce consensus within scientific institutions at a certain level, such as the National Academy of Sciences or the British Medical Associ-

[26] It is, of course, not easy to establish what this "objective expertise" consists of. The use of this terminology reflects implicit adherence to a "research-oriented" concept of expertise, which is based precisely on the presence of objective criteria, such as the possession of more evidence in a certain domain; better reasoning skills and expertise in the same; and, finally, the formation of correct beliefs (see Grundmann, T. (2022) *Experts: What Are They and How Can Laypeople Identify Them?* In J. Lackey & A. McGlynn (eds.), *Oxford Handbook of Social Epistemology*. Oxford University Press, who supports this definition of expertise, as well as Croce 2019, *op. cit.*).

[27] Constantin and Grundmann 2022, *op. cit.*; Grundmann 2021, *op. cit.*; see also Zagzebski 2012, *op. cit*, who supports a similar view.

[28] Lackey 2018, *op. cit*., pp. 233-234.

[29] Grundmann 2021, *op. cit.,* pp. 144-145.

ation, deference is the most epistemically responsible strategy[30]. Indeed, non-experts have no basis of expertise from which to challenge the scientific consensus, which is why we criticize anti-vaxxers or climate change deniers, who in most cases speak from outside the scientific community without any disciplinary expertise[31]. To be sure, deference is not always the right choice, and it is not always what epistemic responsibility would prescribe, because sometimes experts have divergent opinions on the same issue. In such cases, even people with different expertise can responsibly try to form their own opinions, if they have the minimum skills to do so. This is what happens, for example, in evaluating the greater or lesser effectiveness of public policies on which there is no consensus. In cases like these, an individual may listen to several voices, compare them, and even try to take part in the debate if possessed of skills of some use. For this reason, the attitude to be recommended toward expert testimony varies depending on the context and the skills of the individual[32].

However, one of the prerequisites for Levy and Savulescu's proposal to work is that individuals respect the limits of their own competence. This can be done by carefully evaluating the weight they attribute to their own opinions so as not to presume to equate their opinions with those of experts. This means adopting an attitude of intellectual humility (also called *epistemic* humility). If, in fact, presumption and overconfidence are the vices whereby individuals tend to act without recognizing their own limitations and generally overestimate their abilities and knowledge[33], humility is the virtue that allows people to understand who they are and what their position is in relation to others[34]. Intellectual humility operates in the same way, referring to individuals' attitudes toward their own epistemic condition.

The context of the Covid-19 pandemic has particularly highlighted the risks created by overconfidence and presumption, with various institutions and members of society speaking out on medical issues without having any

[30] Levy, N., Savulescu, J. (2020) Epistemic Responsibility in the Face of a Pandemic. *Journal of Law and the Biosciences*, Advance Access Publication 28 May 2020: 1-17; pp. 5-6.

[31] Ivi, p. 7.

[32] Ivi, p. 17.

[33] Cassam Q. (2017) Diagnostic error, overconfidence and self-knowledge. *Pelgrave communications*: 1-8; Roberts R.C., Wood J.W. (2003) *Humility and Epistemic Goods*. In M. De Paul, L. Zagzebski (eds.) *Intellectual Virtue. Perspectives from Ethics and Epistemology*. New York: Clarendon Press, 257-279.

[34] Zagzebski 2012, *op. cit*., p. 246; see also Bommarito N. (2018) Modesty and humility. *Stanford Encyclopedia of Philosophy*.

expertise to do so. As Erik Agner noted, there have been numerous displays of "supreme confidence" by people with no expertise on issues on which the most experienced scientists were expressing themselves with the utmost caution. Hence the call for the exercise of epistemic humility, the virtue that makes human beings aware of the provisional and incomplete nature of their beliefs[35]. Ian Kidd, drawing on Confucianism, has defined intellectual humility as the virtue that empowers one to be aware of one's limitations, to recognize what capacities one does not possess, and to rely on the teachings of "sages"[36]. Humble people are aware of the fragility of their own certainties[37] and act accordingly, avoiding overconfidence[38].

Described in this way, intellectual humility seems to be the virtue that, if exercised, leads individuals to seek out and trust the testimony of experts in contexts in which they realize they are not competent enough to attribute value to their own beliefs. Accordingly, ordinary people who must decide whether to be vaccinated against Covid-19 and wish to know whether and to what extent vaccine prophylaxis is risky, if they adopt an attitude of epistemic humility, will not give undue weight to their own prior opinions but will rely on the advice of those who have the expertise to speak on the topic. Pritchard also noted that intellectual humility, along with other epistemic virtues such as conscientiousness and honesty, can aid in the difficult task of recognizing and debunking fake news. The intellectual virtues are characterized by the search for a right middle ground, a balance between opposing attitudes, an excess of either of which constitutes a vice. Epistemic humility allows one to identify the right way to relate to sources of information: with a critical eye, yet moderating one's skepticism and thus preventing an excess of it from leading one to discredit reliable sources[39].

Although expressed in different words, Baghramian and Panizza's call for "moderated skepticism" also involves a form of humility and constitutes an invitation to achieve the right attitude toward information. They advocate practicing control and vigilance to avoid ending up in a condition

[35] Agner, E. (2020) Epistemic Humility – Knowing your Limits in a Pandemic. *Behavioral Scientist*, accessed online at https://behavioralscientist.org/epistemic-humility-coronavirus-knowing-your-limits-in-a-pandemic.

[36] Kidd I.J. (2015) *Educating for Intellectual Humility*. In J. Baehr (ed.), *Educating for Intellectual Virtues: Applying Virtue Epistemology to Educational Theory and Practice*. London: Routledge, 54-70; p. 62.

[37] Ivi, p. 58.

[38] Ivi, p. 62.

[39] Pritchard, D. (2021). *Good News, Bad News, Fake News*. In S. Bernecker, T. Grundmann, A.K. Flowerree (eds.) *The Epistemology of Fake News*. Oxford: Oxford University Press; 46-76; p. 63.

of subordination that involves sacrificing critical thinking—yet balancing this with trust in those we ourselves recognize as being in a condition to judge better than we can in a given domain[40].

Conclusions

This contribution does not pretend to pronounce definitively and in full on such a complex and intricate theme as a person's relationships with information and with the testimony of experts.

What I wanted to emphasize is the need to maintain a balance in the exercise of the necessary epistemic vigilance. If not properly calibrated, vigilance can turn into an overly critical attitude, and if fueled by presumption or arrogance, this can lead to the epistemic errors of devaluing the testimony of experts and overvaluing one's own opinion. Exaggerated and unmotivated skepticism towards expert testimony can derive from a misunderstanding of what it means to relate critically to information. This error has received less attention in the literature than its opposite counterpart, the exercise of blind trust. Hence the decision to deepen its analysis. It can also derive from overconfidence and arrogance; hence the call to epistemic humility.

The present reflection was partly inspired by a recent news event. During a demonstration by kindergarten teachers opposed to the Covid-19 vaccine, a national newspaper reported an interview with a teacher who had decided not to be vaccinated, thus losing her job. When asked why she chose not to protect herself, the interviewee said that "a drug whose effectiveness drops so quickly is not a vaccine. And then in my opinion there have been too many adverse events." For the interviewee, "this serum should be a personal health treatment, not an obligation that impairs our rights to health and work." Giving up her salary "is a strong choice, but I do it for my children: it is my duty to educate them to critical thinking. But I feel so bitter; in recent years I have had only praise for my professionalism. Now we are treated like this without having done anything wrong except refusing to do something that affects our health, not that of others" (12/24/2021, G. M. Fagnani, *Corriere Della Sera*).

The epistemic errors in this brief excerpt are numerous: the way she expressed herself on the effectiveness of the drug (with what competence?), the excessive emphasis given to her own opinion ("in my opinion there

[40] Baghramian and Panizza 2022, *op. cit.*

were too many adverse events"), the lack of understanding of the concepts of the rights to work and health, and the (false) belief that the choice to decline vaccination does not affect the health of others. What is most striking is that the interviewee interprets her vaccine refusal as an expression of critical thinking. The case reported is just one example of the many people who refuse the vaccine on the grounds of their purported exercise of critical thinking (useful in this regard is Hobson-West's analysis of anti-vax groups predating the Covid-19 era)[41], showing that they have not understood what it consists of. The aim of this work was therefore to highlight that what is sometimes mistaken for critical thinking is actually an epistemic error that consists in marginalizing the opinion of experts, combined with a lack of intellectual humility.

Abstract

Our epistemic duties as citizens of the global world require us to seek information to ensure that our actions do not harm others or ourselves. As we integrate that information, we should not passively accept everything we are told without thinking it through—without ensuring, at the very least, that the sources we rely on are reliable. This avoidance of excessive trust is the counsel of an epistemically vigilant attitude. However, the intention to exercise critical thinking sometimes translates into the opposite excess: distrust and suspicion improperly extended even to experts recognized as such by the scientific community and by the individuals themselves. If a passive or compliant attitude risks leading individuals into error, so does an excessively critical attitude. We need to redefine the role of experts in order to establish a relationship with them that is neither one of passive subordination nor one of distrust. It will be shown how a correct relationship with experts also passes through the exercise of a particular epistemic virtue—intellectual humility.

Keywords: trust in experts; epistemic vigilance; intellectual humility.

Francesca Pongiglione
Università Vita-Salute San Raffaele
pongiglione.francesca@unisr.it

[41] Hobson-West P. (2007) 'Trusting blindly can be the biggest risk of all': organized resistance to childhood vaccination in the UK. *Sociology of Health & Illness* 29, 2: 198-215.

T

Sarah Songhorian, Francesca Guma, Federico Bina, Massimo Reichlin[1]

Moral Progress: *Just* a Matter of Behavior?

1. *Moral improvement as a requisite of moral progress*

One of the most relevant challenges, for empirically informed ethics, is to understand whether and how moral progress is feasible, given human beings' natural equipment (Klenk & Sauer 2021; Buchanan & Powell 2018). To answer this question, we need to clarify what is meant by "moral progress" and to suggest how it can be measured. The recent discussion on this topic mainly identified moral progress with the institution of collective moral practices which are considered better ones in virtue of their outcomes (Sauer 2019; Sauer et al. 2021); for example, because they better promote the well-being of people and/or sentient individuals. According to this approach, to see whether any moral progress has come about, we need to consider the outcomes that those practices and people's observable behavior actually produce.

However, if on the one hand moral progress refers to changes in collective behavior, on the other we believe that it must encompass individuals' moral improvement as well. We suggest the existence of a bijective relationship between a society's moral progress and the moral improvement of the individuals who are part of it. More precisely, we suggest a) that any progress in collective institutions and practices requires the active contribution of some individuals who have developed a sensitivi-

[1] Faculty of Philosophy, Vita-Salute San Raffaele University, Milan. Corresponding author: songhorian.sarah@unisr.it. Although all authors have collaborated in discussing and revising all parts of the paper, MR is mainly responsible for writing § 1, FG for § 2, SS for § 3 and § 5, and FB for § 4.

 DOI: 10.4454/teoria.v42i2.164

ty for the values at stake; b) moreover, that moral progress is inherently unstable unless enough individuals with better moral capacities promote and strengthen the new ideals and values through their beliefs and behaviors. These claims are partly in line with Buchanan and Powell's interest in highlighting the links between individual and socio-institutional moral change. However, despite touching on potentially interesting implications for an account of individual moral progress, Buchanan and Powell basically refer to morality and moral progress as social phenomena, not merely as individual ones; and they are mostly interested in individual changes in moral beliefs and attitudes «only insofar as these occur in sufficiently large numbers of people to effect social change» (Buchanan & Powell 2018, p. 47).

On the contrary, we want to single out the importance of the moral improvement of individuals who become sensitive to certain values. We agree that it is only with the spread of the new beliefs and values in a sufficient number of individuals that societal moral progress – i.e., a progressive change in common sense morality – is realized; and that such a progress is a factor in determining progressive changes in laws, or other established social practices and formal institutions. There are, of course, complex relationships among the three levels. And yet, the improvement of individuals is an important condition of societal moral progress, which in turn contributes to institutional progress. This must not be meant to exclude that institutional moral progress may also be accomplished independently, nor that it usually has feedback effects on both individual and societal moral progress.

If this picture is plausible, then it is reasonable to say that human moral progress depends at least in part on the possibility for individuals to improve their moral capacities, e.g., by reducing the influence of epistemically defectives biases and other distorting influences. Based on empirical research, some believe that the pervasiveness of morally irrelevant influences on moral judgments prevents moral progress (see Klenk & Sauer 2021, pp. 947-956). Quite the opposite, we believe that moral progress is possible – among other things – by enhancing our capacities to consciously control our moral judgments. Improving the capacity to produce consistent, accurate, and informed moral judgments may make individual improvement possible, thereby causing effects also in the social sphere (Campbell & Kumar 2012). While societal moral progress may be accomplished independently from individual moral improvement, improving the individual capacity for moral judgment is a relevant contribution to the promotion and

strengthening of progressive changes in institutional moral practices: better moral agents adopt better moral behaviors that may eventually be institutionalized, and they may actively promote progressive changes in social institutions.

2. *Moral progress beyond behavior*

As anticipated, many have identified moral progress in the outcomes that social structures and institutions produce, with particular reference to people's observable behavior. While this conceptualization might make its measurement easier, we will show that it overlooks relevant aspects of moral progress. To better illustrate our point, let us consider a few interesting experimental works that, although not primarily intended as contributions to the debate on moral progress, have nonetheless implications for it, since they consider changes in behavior in a direction the authors deem positive.

Schwitzgebel, Cokelet, and Singer (2020) have empirically tested whether ethics classes positively influence students' moral behavior, especially by looking at their meat consumption before and after an educational intervention. Since students attending ethics classes increased their vegetarian choices (as compared to a control group), they can be conceived as having improved, since reducing meat consumption is deemed a morally positive change by the authors. This study was extended and replicated, confirming that it is possible to influence students' attitudes and daily behavior through standard methods of university-level philosophy instruction (Schwitzgebel, Cokelet & Singer 2021).

In both cases, two factors are crucial for a moral improvement to occur. On the one hand, a change in the subjects' behavior or opinion is essential: individual moral improvement is, thus, substantive – i.e., it focuses on the observable behavior or on the content of normative judgments. On the other, the change has a precise direction decided from the beginning by the researchers. The goal is already clear in the title of the first paper: the authors want to study whether it is possible to influence students' behavior in a direction that is assumed beforehand as positive.

These studies have not been explicitly meant as a contribution to either define or promote moral progress, but simply as a test of the ability of ethics lessons to influence students' behavior. However, it is our opinion that, when placed in the context of identifying ways to stimulate moral progress,

this approach is exposed to several criticisms. First, such interventions can result in forms of indoctrination. Given that the experimenters aim at obtaining a certain response, can one really consider it an authentic moral improvement of the subject? If only the outputs are observed, it seems difficult to evaluate whether a change in moral opinion and/or behavior is the result of the acritical (and perhaps not fully conscious) assumption of an external point of view or the effect of a new personal way of thinking about the matter. Moreover, can these modifications be considered as stable improvements? The follow-up conducted in 2021 shows a fair amount of stability. However, if such results are the product of suggestion or indoctrination, the question remains not only whether the subjects would be able to personally formulate good reasons in favor of their new opinion, or whether they would merely repeat remarks that impressed them, but also how long these changes can last. Finally, is it possible to identify what produced the change and how it occurred? Focusing exclusively on the substantive component does not allow one to understand how the changes occurred. Considering only behavior prevents us from ascertaining whether these results are the effect of an increase in the individual capacity for moral judgment, a different personal way of judging, or just the effect of indoctrination. Influencing subjects to produce a particular behavior does not help illuminating the real factors producing that change.

In their second study, Schwitzgebel, Cokelet, and Singer refine the experiment to test whether behavioral change is caused mostly by elements of the instruction (e.g., by introducing non-vegetarian professors and by allowing only half of the students to watch a vegetarianism advocacy video). These modifications could decrease the likelihood of students' suggestibility. However, since these studies do not focus on the reasoning and justification processes that lead subjects to accept certain judgments, such considerations remain only hypotheses.

Focusing on a substantive component of moral improvement seems unsatisfactory. In particular, this approach ignores that a change in behavior or moral judgment, while being an indicator of moral improvement, should not be considered the only possible one, nor the most suggestive. For these reasons, and since we believe that philosophy is one of the means – among others – to achieve moral improvement, in this paper we suggest an alternative approach to the issue.

3. *A procedural moral improvement*

Given the difficulties of an account focused uniquely on behavior, we argue that moral improvement should be considered first and foremost as having a procedural rather than a substantive character (Schaefer & Savulescu 2019; Rawls 1951). On this account, we should not look at humans' actual behavior nor at the content of their moral judgments – although both are certainly relevant –, but rather at the abilities and faculties needed to ground them. What is relevant in this perspective is not what individuals do, judge, or believe; but rather the reasons and justifications they can provide in support of their actions, judgments, and beliefs. Thus, our goal is to underline the role of how a moral output is reached rather than simply focus on what that output actually is. Regardless of the content of a given moral output, we agree with Schaefer and Savulescu (2019) who provide a set of features that would make a moral justification a good one – i.e., logical, empirical, and conceptual competence; openness to the revision of one's opinions; sympathetic imagination; and the attempt to reduce one's biases. Once an output has been reached by making the best of these (and possibly other) abilities, then it should count as an improved one as opposed to an outcome that does not involve them at all.

We argue that there are at least two reasons for a procedural account to be preferable. First, it enables one to hold a pluralistic stance «thus avoiding many question-begging moral assumptions» (Schaefer & Savulescu 2019, p. 75). Several moral disputes are, in fact, so controversial that it is problematic to believe that one solution is certainly the true one, that everyone has reasons to accept. Second, a procedural account – especially one that is concerned primarily with how people justify their behaviors, decisions, judgments, and beliefs – is more suited to account for instances in which an individual might have come to a moral conclusion because of external or internal drives that would not count as an appropriate moral justification. Let us now delve a little bit more into these two issues.

As far as the latter is concerned, to say that individual moral improvement only consists in performing – or complying with – practices judged by a third party as "morally better" overlooks the possibility that behavior can be influenced or causally determined by manipulation or indoctrination. Since the latter are hardly considered appropriate sources of moral education, or of individual moral improvement, accounts that focus uniquely on behavior should show how they can be excluded. How can

we distinguish between someone who is getting rid of her biased behavior towards a social group because she has understood that it was grounded on faulty bases so that she now believes it was morally wrong to have that behavior to begin with, from someone else who does exactly the same just because it is fashionable to be seen as open-minded? In this case, the behavior change will certainly be relevant to account for a person's improvement, but it will not be sufficient. Indeed, it is difficult to say whether a change is determined by an effective, stable, and authentic moral improvement by only observing behavior: people could act in a certain way because they are influenced by internal or external stimuli, by their desire to be socially approved rather than by that of deserving approbation (Smith 1759, III.2.32), by morally irrelevant factors rather than by the morally pivotal ones. On the assumption that an authentic moral action involves a strong sense of agency of the subjects, focusing on how judgments are made and on how moral behavior is grounded can reveal a way to increase the agents' real moral capacity and the conscientiousness of their moral responses (Schaefer 2015).

Coming to the first issue, the adoption of a pluralistic stance drives us clearly away from accounts measuring moral change and moral improvement only in terms of their behavioral outputs. Although, as noted in § 2, Schwitzgebel, Cokelet, and Singer (2020, 2021) are not explicitly concerned with moral progress or improvement, their focus on the reduction of meat consumption after studying meat ethics is a concrete example of the substantive view that we deem problematic (especially when applied to more debated or controversial issues). There are, in fact, many contexts of choice where the issues involved are so disputable, and/or where no action is clearly recommended, that believing one particular behavior represents the right way to go means assuming a specific normative outlook, one that might not be universally shared. Is there a clear set of actions that we can universally conceive as the right one when dealing with issues such as, say, the scarcity of health care resources or global poverty? Since the answer is negative, it seems reasonable to focus on how people justify their often-divergent beliefs and behaviors; endowing people with a sensitivity for the reasons at stake and a capacity to respond to them helps reducing moral plurality by excluding those moral stances that do not pass the test of justification. Thus, improving the abilities and faculties that are involved in an appropriate moral justification should be the starting point to promote moral improvement and moral change. Schaefer and Savulescu's list (2019) is an interesting example of how one can and

should proceed, although we do not claim it is the only one nor it is necessarily complete.

By aiming to avoid the imposition of a substantive normative standpoint as the only right or best one, a procedural account lowers the risk of indoctrination, manipulation, and paternalism in the promotion and assessment of moral improvement and aims to track the path to enhancing moral agency. Focusing on individuals' abilities to provide reasons according to logical, empirical, and conceptual competence, openness to the revision of one's opinions, sympathetic imagination, and bias reduction – i.e., the abilities Schaefer and Savulescu focus on – is a good starting point to ascertain whether one is actually improving her moral stance. Thus, while a behavior or a judgment for which the subject can provide (convincing) reasons is certainly better than one for which no justification seems to be available to her, this clearly is not the end of the story nor a solution for every moral dispute. Much is yet needed for a complete account of individual moral improvement to be in place.

To pave the way for it, though, a procedural account like the one we have gestured towards here is required. In § 4, we will consider one of the most challenging objections to such an account: how can we be sure that improving someone's ability to provide reasons for her actions leads to a moral progress and not a regress? How can we be sure that a procedural account of moral justification has the resources to distinguish proper justification (or moral reasoning) from mere post-hoc confabulation (Haidt 2001; Greene 2008)?

4. *Acceptable moral justifications*

As mentioned, by avoiding any substantive commitment, our proposal risks considering an amelioration in the formal ability to rationalize any moral (or immoral) conclusion as a proper instance of moral improvement. In this section, we offer some replies to this concern by suggesting that not every reason-giving account counts as a proper moral justification, and by adding some considerations about the empirical and theoretical assumptions which may ground this worry.

This objection may stem from views sympathetic to Haidt's influential model of moral judgment (Haidt 2001). Drawing on empirical research, Haidt concludes that moral judgment is not the product of conscious reasoning, but the expression of automatic, unconscious, and affectively-laden

"intuitions" shaped by evolutionary, cultural, and social pressures[2]. Within this model, conscious reasoning intervenes only ex post by concocting reasons to support and socially justify fast and automatic reactions: «one feels a quick flash of revulsion […] and knows intuitively that something is wrong. Then, when faced with a social demand for a verbal justification, one becomes a lawyer trying to build a case rather than a judge searching for the truth» (Haidt 2001, p. 182). According to Haidt, the function of moral reasoning is to socially justify intuitive responses, but it has no power in shaping their content ex ante. In this framework, increased proficiency in the ability to provide socially acceptable justifications would just better perform the function of convincing others about the acceptability of conclusions that are essentially insensitive to rational scrutiny and revision.

However, not all justifications are equal. Following Schaefer and Savulescu (2019), we believe that satisfying certain procedural requirements makes certain reasons or justifications more intersubjectively acceptable than others, without committing to any substantive normative or metaethical view[3]. In particular, some justifications can be more consistent, more sensitive to empirical evidence and to others' perspectives and interests, and more open to revision than others. Acceptable moral justifications do not merely confirm one's opinions, intuitions, or feelings by effectively convincing other people about their soundness; they also express the effort of considering a broader spectrum of information, such as non-moral facts, or the interests and preferences of the individuals involved (including the agent's ones).

If, as we believe, Schaefer and Savulescu's criteria are reasonable and sensible, one can discriminate between different levels of reliability or appropriateness of moral justifications, distinguishing between confabulations (and the correlative phenomenon of moral dumbfounding), motivated or confirmatory rationalizations, and appropriate moral justifications.

In light of Haidt's work, a confabulation is the attempt to fabricate justifications for moral conclusions with clear fallacious results (e.g., blatant logical contradictions), pushed by the desire to hold and confirm one's feelings, intuitive judgments, and beliefs, even when put in front of inconsistencies and contrasting rational arguments (Festinger 1957; Kunda

[2] Haidt's idea of "intuition" differs radically from traditional (rationalist) conceptions of the term in the history of ethics.

[3] See §3 and below in this section – as well as Schaefer and Savulescu (2019) – for a more extensive discussion and defense of these criteria.

1990). In Haidt's famous experiments, some subjects try to rustle up support for their intuitive conclusions by offering fallacious and unsatisfactory justifications which, for example, patently clash with relevant information or just restate intuitive conclusions without justifying them at all (Haidt 2001, 2012). Therefore, we can conceive confabulation as a vicious kind of reason-giving, which lacks several features of an acceptable justification (such as empirical and logical consistency and openness to revision).

Rationalization can be conceived, more broadly, as the justification of behavioral outputs by offering reasons in their support "that would have made it rational" (Cushman 2020, p. 183), even if such reasons do not match the actual processes that led to that output. Many rationalizations can be more consistent and sensitive to logical reasoning and evidence than moral confabulation. However, providing reasons in favor of a moral judgment does not guarantee providing acceptable moral reasons because what is rational, e.g., from a self-interested point of view may not be so from a moral point of view. For example, a rationalization may be grounded on an astute selection of data, aimed to make the preferred conclusion plausible, while a proper moral justification does consider more morally relevant factors, such as the interests of other individuals involved. Also, while rationalization does not require critically examining one's own moral preferences, a good moral justification does. Moreover, even though rationalization requires paying attention to possible influences of biases on argumentation, it does not require taking seriously, for instance, the main moral reasons for and against available stances or lines of action. Supporting moral conclusions with acceptable moral justifications does not simply require a generic capacity to provide any kind of reasons in their favor, but to provide a much more specific kind of reason-giving account.

An acceptable moral justification, thus, requires adequately knowing the context of the situation under evaluation, along with one's and others' perspectives. Reasons for and against different conclusions should be balanced in light of available information, showing the attitude to evaluate potential alternatives with an open mind, and being disposed to reconsider one's opinions. The potential influences of biases or prejudices that might affect the evaluation should also be considered. To achieve this goal, it is important to avoid considering one's preferences as the right evaluative standard for the situation at hand, acknowledging and balancing the different interests at stake. Finally, acceptable moral justifications should satisfy standards of logical consistency. Improvement in these capacities would not only enhance the formal ability to justify any

possible moral judgment or behavior – as the objection we are addressing states – because if these requirements are satisfied the spectrum of reasonably acceptable moral conclusions shrinks significantly.

A couple of final remarks are in order. A strength of our view is that it stands even if Haidt's model of moral judgment is plausible. Even if in isolated, specific circumstances of choice explicit moral reasoning intervenes only after quicker psychological responses, improved justificatory abilities would not just better support a-rational outputs, but can be sensitive to independent relevant information. Nonetheless, there are several reasons to reject Haidt's thesis according to which moral reasoning has no causal influence on moral feelings, intuitions, and judgments. Critics have stressed the limits of Haidt's model, denouncing its rigid lack of interaction between controlled and automatic processes, as well as its blindness about the diachronic dimension of moral judgment (Campbell & Kumar 2012; Railton 2014).

Even if it does not come into play immediately before the expression of a moral conclusion at the time of decision, explicit moral reasoning can feedback on, inform, and improve people's future moral responses (Sauer 2017). If this is true, an appropriate moral justification can also reliably point out some of the reasons that informed one's intuitive judgment or behavior (Cushman 2020).

All these are not necessary requirements of motivated (or confirmatory) rationalizations. Therefore, we conclude that acceptable moral justifications can be distinguished from other reason-giving accounts. This allows us to reject the objection accusing our position of considering mere improvements in the capacity to rationalize as proper moral improvements.

5. *Conclusion*

The aim of this paper was to argue in favor of the need – within the empirically informed debate on moral progress – to focus on an individual procedural moral improvement. We have argued that moral improvement is an essential condition for moral progress and that it should be understood not in substantive terms, but rather in procedural ones. A change in behavior or in moral judgment, while being an indicator of moral improvement, should not be considered the only possible one, nor the most indicative.

For this reason, we have gestured towards a procedural account of the abilities required to reason and to justify one's actions and beliefs as the

first necessary step to truly understand the contribution made by individual moral improvement to the debate on moral progress.

Finally, we have considered a challenging objection to our account – i.e., whether the abilities a procedural account proposes to improve allow us to distinguish appropriate moral justifications from mere post-hoc confabulations; we have argued that such a distinction can in fact be drawn and that not any reason-giving account counts as a proper form of moral justification.

References

Buchanan, A., Powell, R. 2018, *The Evolution of Moral Progress: A Biocultural Theory*, Oxford University Press, Oxford.

Campbell, R., Kumar, V. 2012, "Moral Reasoning on The Ground", Ethics, vol. 122, n. 2, pp. 273-312.

Cushman, F. 2020, "Rationalization is Rational", Behavioral and Brain Sciences, vol. 43, pp. 1-16.

Festinger, L. 1957, *A Theory of Cognitive Dissonance*, Stanford University Press, Stanford.

Greene, J.D. 2008. "The Secret Joke of Kant's Soul". In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Vol. 3. The Neuroscience of Morality: Emotion, Brain Disorders, and Development* (pp. 35-80). MIT Press, Cambridge.

Haidt, J. 2001, "The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgment", Psychological Review, vol. 108, pp. 814-834.

Haidt, J. 2012, *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. Pantheon Books, New York.

Klenk, M., Sauer, H. 2021, "Moral Judgement and Moral Progress: The Problem of Cognitive Control", Philosophical Psychology, vol. 34, n. 7, pp. 938-961.

Kunda, Z. 1990, "The Case for Motivated Reasoning", Psychological Bulletin, vol. 108, n. 3, pp. 480-498.

Railton, P. 2014, "The Affective Dog and Its Rational Tale: Intuition and Attunement", Ethics, vol. 124, n. 4, pp. 813-859.

Rawls, J. 1951, "Outline of a Decision Procedure for Ethics", The Philosophical Review, vol. 60, n. 2, pp. 177-197.

Sauer, H. 2017, *Moral Judgments as Educated Intuitions*. MIT Press, Cambridge (MA).

Sauer, H. 2019, "Butchering Benevolence Moral Progress beyond the Expanding Circle", Ethical Theory and Moral Practice, vol. 22, n. 1, pp. 153-167.

Sauer, H., Blunden, C., Eriksen, C., and Rehren, P. 2021, "Moral progress: Recent developments", Philosophy Compass, 16(10), e12769.

Schaefer, G.O. 2015. "Direct vs. Indirect Moral Enhancement", Kennedy Institute of Ethics Journal, vol. 25, n. 3, pp. 261-289.

Schaefer, G.O., Savulescu, J. 2019, "Procedural Moral Enhancement", Neuroethics, vol. 12, n. 1, pp. 73-84.

Schinkel, A., de Ruyter D.J. 2017, "Individual Moral Development and Moral Progress", Ethical Theory and Moral Practice, vol. 20, n. 1, pp. 121-136.

Schwitzgebel, E., Cokelet, B., and Singer, P. 2020, "Do Ethics Classes Influence Student Behavior? Case Study: Teaching the Ethics of Eating Meat", Cognition, vol. 203, p. 104397.

Schwitzgebel, E., Cokelet, B. and Singer, P. 2021, "Students Eat Less Meat After Studying Meat Ethics", Review of Philosophy and Psychology. https://doi.org/10.1007/s13164-021-00583-0

Smith, A. (1774), *The Theory of Moral Sentiments: An Essay Towards an Analysis of the Principles by which Men Naturally Judge Concerning the Conduct and Character, First of their Neighbours, and Afterwards of Themselves*. The Fourth Edition, Millar, London; Kincaid and Bell, Edinburgh.

Abstract

The aim of this paper is to argue in favor of the need – within the empirically informed debate on moral progress – to focus on an individual procedural moral improvement. We argue that moral improvement is a prerequisite for moral progress and that it should be understood in procedural (rather than substantive) terms.

Thus, we gesture towards a procedural account of the abilities required to reason and to justify one's actions and beliefs as the first necessary step to understanding the contribution individual moral improvement makes to the debate on moral progress.

Finally, we consider a challenging objection to our account – i.e., whether the abilities a procedural account proposes to improve allow us to distinguish appropriate moral justifications from mere post-hoc confabulations – arguing that not any reason-giving account counts as a proper form of moral justification.

Keywords: moral progress; individual moral improvement; procedural improvement.

T

Francesca Guma

Becoming Better Moral Agents by Strengthening Free Will. A Possible Prospect?

1. *The limits of judgment*

Several theories and experimental proposals in contemporary ethics show an intent to study various aspects of moral reasoning and a desire to investigate whether and how it is possible to improve and/or enhance the ability to make moral choices (Klenk & Sauer 2021; Songhorian, Guma, Bina & Reichlin 2022). This is a relevant issue, especially as a result of studies in cognitive psychology, neuroscientific research, and reflections on self-control that psychologists, philosophers, and decision theorists have made in recent years (Bermùdez 2018).

Referring primarily to Kahneman's studies (Kahneman 2011), many authors point out that decisions are often made in absence of awareness. Individuals frequently give automatic responses that do not result from thoughtful, informed reflection. This happens not only in tasks considered entirely solvable by resorting to fast and intuitive thought processes but also in those which seem important to appeal to slower but also more logical and reflective processes. Similar remarks have led to more inquiry into the origin – rational or emotional – of moral judgments. In an attempt to integrate empirical knowledge with philosophical knowledge, several vastly different theoretical proposals have arisen in ethics (Greene 2013; Nichols 2004; Sauer 2017). Regardless of the position considered, the presence of automatic and unconscious reasoning, even in the elaboration of moral judgments leads to problematizing the subject's capacity to make conscious ethical choices. Similar remarks can also be made by taking psychological theories other than Kahneman's model as a reference. Considering, for example, a psychodynamic approach that takes the psyche in its

 DOI: 10.4454/teoria.v42i2.165

topical, dynamic, and economic relationships into account, the difficulties related to the agent's control and awareness remain evident: a thoughtful, rational, conscious choice is an intricate and complex matter. Stressing that the human will cannot be understood in its totality if confined to the level of consciousness, these theories show that it is easy to find both situations in which an unconscious will can be traced and situations in which conscious psychic acts experienced as willed and free although, in reality, are not (Guma 2021).

For these reasons, regardless of the theoretical frame of reference, it is important to ask how much control of action and choice the agent really has, to what extent it is possible to make conscious and thoughtful moral choices and if so, whether one can call upon greater commitment and concentration to arrive at more accurate reflections. These questions are important because, especially in ethics, the will must depend on us and must be in our power. Generally, agents are exempt from responsibility for effects that are not caused by them and that they cannot avoid. Recognizing the contribution scientific research has offered, it becomes difficult to evaluate an individual who acts driven by internal influences that she does not know and cannot control. This problem emerges, for example, in actions caused by implicit biases, which undermine the subject's agency. Several authors question whether – and how – we can hold ourselves accountable for our implicit biases, or how we should structure society to counterbalance them (Beeghly & Madva 2020, part 3). As we recognize the presence of automatic and unconscious modes of functioning, it becomes crucial to ask how much control we can actually exercise over the choices and moral judgments we make.

2. *Free will and moral judgment*

Theories developed from the observation of the limits of rationality and conscious control that human beings exercise over their choices, decisions, and actions are closely related to the question of free will. To understand this relationship, it is essential to start from a notion of free will that can dialogue with empirical research. Free will can, thus, be identified with the opportunity and capacity to will otherwise. To do so, it is necessary to give an empirical interpretation to the two conditions deemed necessary to define free will: the existence of alternative possibilities and the agent's conscious control of their will. In this view, freedom of will on the one hand

requires the existence of actual alternative possibilities to be understood in both a negative and a positive sense: to choose and act freely, a subject must suffer neither the effect of exogenous forces that constrain from outside (she must have the actual opportunity to act and will), nor the effect of endogenous forces that limit or prevent her from being able to act otherwise (she must have the actual ability to act and will). On the other hand, to speak of a free choice or action, it becomes essential to ascertain the presence of actual conscious control of the will by the subject, that it is crucial to assess the subject's agency: the individual must subjectively perceive that she is in control of her behavior, must feel endowed with the opportunity and ability to act and will (Magni 2019).

This naturalistic conception of free will reveals the connections between the considerations seen in §1 and freedom of will: as List points out, the subjects' sense of agency is intimately connected to the idea of being able to make decisions independently; individuals want to be their own masters, to achieve their own ends, to act and choose consciously (List 2019); however, empirical findings undermine or even disprove the very agency of the subject (Soon, He, Bode & Haynes 2013). Considering psychic determinism and the related difficulties an individual may face when confronted with her own mental resistances, cognitive deficits and motivational constraints coincide with admitting humans' concrete difficulties to cope with endogenous and exogenous forces that hinder them in effectively choosing freely.

If judgments are believed to be at least in part «a form of measurement in which the instrument is a human mind» (Kahneman, Sibony & Sunstein 2021, p. 361) and the results of experimental research are noted, it becomes important to ask whether and how it is possible to develop strategies that can make moral judgments more competent, more rational and more solid. The possibility of becoming better moral agents is related to the possibility of increasing the effective opportunity and ability to will otherwise. Reflecting on strategies that can improve and/or enhance the capacity to make moral choices can be seen as reflecting on strategies that can improve and/or enhance free will. Assuming that for human beings nothing is truly neutral and that, therefore, depending on the characteristics of the observed objects one is pushed to produce unconscious/automatic inferences that condition the results of reasoning, some proposals developed to improve moral reasoning can be considered attempts born out of reflection on the possibility of expanding the subject's positive freedom, of increasing control over action and choice to make her less exposed to automatisms.

But is it possible to strengthen free will to become better moral agents? If so, how?

In the next two paragraphs, I will present two possible ways to achieve an improvement in individual moral action, showing how, while both acknowledging free will natural and empirical limits, these proposals arrive at extremely different solutions.

3. Nudge and suggestion

Assuming the ineradicable presence of endogenous and exogenous conditions that make it difficult for the agent to control her action and choice, those advocating for nudges and suggestion believe it is impossible to increase the subject's agency. Thus, the only possibility to achieve individual moral improvement, lies in introducing specific exogenous conditions capable of directing individual's actions and choices: to improve people's moral judgments it is appropriate to induce them, through more or less gentle nudges and suggestion, to make objectively better choices. Thaler and Sunstein's proposal (2008) is one of the most famous examples of such an approach. The authors believe choice architects must influence individuals' behavior to improve their lives. Subjects are left free to choose, meaning that the existence of actual alternative possibilities is not affected. However, because they are judged to be fallible in making their own decisions, individuals are prodded through information disclosure, warning, and making rules about default situations. Clearly, from the negative aspect of freedom point of view, there are no restrictions: choices are not prevented, blocked, or made overly burdensome; no constraints or prohibitions are placed, and the subject can choose among possible alternatives. On the positive aspect of freedom, although the authors start with the idea of providing measures that protect or increase freedom of choice (Thaler & Sunstein 2008, p. 5), the proposed interventions are not designed to maintain or increase self-actualization and non-hetero-directed action.

Such an approach offers, at least in the first instance, guidelines that appear attractive also to be applied to the increasing moral capacities context. This theory not only considers the objective and incontrovertible evidence that humans will never be fully rational but also emphasizes that they will always and in any case be conditioned in their decisions. Starting from this acknowledgment, it suggests practical interventions that produce

concrete results, allowing agents to make choices that are deemed objectively better (John, Smith & Stoker 2009; Bhargava & Loewenstein 2015). Extending the theory of nudging to the field of reflection interested in improving and/or enhancing moral capacities is quite straightforward, especially since nudging is also applied in contexts that require ethical choices and reflections. Consider, for example, *Green Nudge*, which was created precisely to encourage individuals to engage in environmentally responsible behavior (Schubert, 2017).

However, applying these conceptions focused on nudges to the enhancement of moral capacities leads to some difficulties. Firstly, approaches based on nudging (or suggestive interventions) are not born to achieve an individual's true improvement. As much as they assume that humans would be better moral agents if they better control their moral reasoning and become more aware of it, they conclude that given the impossibility or difficulty in achieving such improvements, it is essential to develop a way to achieve objectively better, concrete results. To do that: they define (more or less explicitly) a scale of values that they consider preferable to the one that an individual might have; they devise interventions that play on what they have identified as cognitive weaknesses; and finally, they nudge/suggest the individual to make a certain type of judgment. Such interventions cannot achieve effective individual moral improvement because they are not aimed at increasing the subject's capacity. By choosing the path that leverages the agent's weaknesses, they aim only for outward behavior modification. The person undergoing such an intervention is not stimulated to produce an improvement in moral reasoning, but rather to give a particular response. The new way of acting or judging may be better considering the social context, but it cannot be considered an actual improvement of the agent. How much can a change achieved by this route be worth and how long can it last? Being the result of a suggestion, the subject has not acquired any ability, and the effect remains linked to the strength of the input given from outside.

Secondly, they do not seem to safeguard a morally relevant characteristic: the subject's agency. The moral actions we consider authentic are those that involve a strong sense of agency: our judgments of responsibility, praise, and blame are stronger if we can attribute agency to the individual. These approaches inserting external elements with the purpose of piloting judgments and decisions do not recognize the value of agency. The possibilities for moral improvement through the development of interventions that influence behavior, attitudes, dispositions, and motivation raise

important questions of freedom and responsibility that not only affect our sense of who and what we are, but also whether we are, or can remain, creators and masters of our decisions and actions (Harris 2016). Such proposals seem to imply that it is permissible to subjugate the autonomy of an individual who does not appear to behave rationally. Taking up a critique developed by Quong concerning paternalism, it is possible to note that starting from people's psychological deficits to direct their choices does not seem a good way to develop their individual moral capacities, but neither is it a good way to respect them as persons. Underneath these settings lies a «judgemental definition»: agent A attempts to increase agent B's values toward a particular decision or situation that B faces; A's action is motivated by a negative judgment regarding B's capabilities to make the right decision, or to handle the particular situation in such a way that she can effectively increase her own values. In making a negative judgment, A may have considered and examined three different abilities of B: practical reasoning, willpower, and emotional management. In identifying these three capacities as relevant, Quong rules out the possibility of considering actions as paternalistic solely because they aim to make up for physical or informational deficits. The heart of paternalistic action is always in negative judgment. The one who performs the paternalistic act, in each case, believes that she knows better than the other how the latter should act; she is convinced that the other does not possess the necessary level of rationality, willpower, or emotional management to accomplish what is best for her. To treat a person paternalistically is yes to treat her as a child, but in a specific sense: it is an attempt to act in her best interest because it is believed that such a person lacks the ability to do that for herself (Quong 2010). Considering this analysis, it is possible to say that choosing this first way to achieve moral improvement in the individual means judging one's mental abilities negatively, disbelieving her deliberative capacities, deeming her inferior in the faculty of decision-making and/or choice, and arrogating the right to direct her, more or less kindly, to the option that is deemed best.

Thirdly, these proposals risk generating coercive fallout in the social freedom area. Assuming the impossibility of increasing conscious control of the agent's will, the architects of choice, essentially, pose as directors of conscience: it is true that they always leave the possibility of alternative choices, but it is also true that they assume that individuals will be more likely to go toward what is suggested to them. Such a scenario does not seem very different from the one described by Berlin in *Two Concepts of*

Liberty (1969). Reflecting in the political sphere on conceptions that identify the notion of freedom with self-determination, Berlin asserted that it is possible to embark on a slippery slope capable of leading to a conclusion that is at least singular: the total denial of the freedom that, on the contrary, was intended. Following his argument, asserting the existence of an ego split into a rational part of an elevated nature (the true Ego) and an irrational, desiring part of a lower nature (the empirical Ego) can lead to the claim that a person is free if and only if she acts rationally, following her true Ego. Since it is commonly agreed that some people are more rational than others, if one or more people were to convince themselves that they know the true end to which actions should tend, they could construct a good argument for coercing other less rational individuals. The reasoning would be based on the consideration that the desires of the less rational subjects would be equal to those of the more rational ones if they were not distracted by the empirical Ego. From here the step to a totalitarian state would be a short one: leaning on the consideration that the many would desire what they are forced to if they were not at the mercy of their lower natures, such oppression would come to take the form of liberation. This perspective, however exaggerated, does not seem so surreal for positions that favor nudges and suggestions.

In conclusion, I agree with the starting point of these accounts and believe that it is important to consider the exogenous forces that every human is subjected to daily. Indeed, it is interesting to point out that such conceptions highlight that a good choice architecture system could help make information more comprehensible and could allow individuals to refine their capacity to map decisions. Noting that subjects are always inevitably subjected to events and rules that influence their judgments can help develop strategies that, by calculating exogenous forces, stimulate agents to behave more consciously. In this sense, nudges could be thought of as tools aimed at increasing the subject's positive freedom, or as tools that respect the decision-making autonomy of the individual and enhance reflective decision-making (Baldwin 2014). However, these would be different interventions because they are not designed to obtain a specific response or behavior from the agent. For these reasons, I believe it is appropriate to seek an alternative way to achieve an improvement in individual moral action.

4. *Increasing agency*

Always starting from the observation that individuals defect in moral capacities due to a lack of conscious control of will, I argue for the possibility of identifying ways to increase the subject's agency, focusing, for example, on knowing one's own psychic dynamics, augmenting or elucidating information, and providing spaces to reflect on one's logical and argumentation capacities. Acknowledging humans' natural, automatic, and unconscious component does not necessarily lead to the conclusion that the subject has, as an individual, no margin for improvement: the individual certainly has a limited capacity to consciously control her will, but this does not eliminate the possibility of considering her capable of increasing it. If one considers agency a crucial aspect of ethics, then it seems essential to develop interventions that aim to track ways to increase the effective possession of the opportunity and capacity to want otherwise. Indeed, this approach does not have the advantage of achieving definite and obvious concrete results, but it appears preferable for at least two reasons.

Firstly, it aims for effective and stable individual moral improvement because it preserves the interest in identifying ways that can effectively expand the subject's possibilities for conscious choice and decision-making. Accepting that humans would be better moral agents if they could better control their moral reasoning and acquire greater awareness of it, this approach aims to increase the subject's effective possession of the opportunity and capacity for conscious control. The focus is not on the content of moral judgments, but on the ability to develop them, be aware of them, justify them, recognize them, argue for them, and provide good reasons in their defense. Maintaining the goal of increasing agency, one seeks not a change in outward behavior (in the output), but a change in the procedure underlying the capacity for moral reasoning, an increase in the awareness of the reasons for one's moral judgments. Although this approach recognizes the impossibility of generating automatons with perfect morality, it allows effective individual moral improvement. Developing procedures that can strengthen the subject's free will indeed makes it possible to think of genuine and stable moral improvements because there would be changes and improvements not in any specific outward behaviors but in the individual's general moral attitude. This view also avoids possible coercive effects on social freedom: by focusing not on the content of judgments, but rather on how they are made, this approach is not committed to a specific nor-

mative framework, nor does it presume to list what can be judged as good or right, leaving the agent free to work out the judgment she deems most appropriate.

Secondly, keeping the goal of increasing agency, safeguards the morally relevant characteristic that previous positions do not value. As much as the literature on moral responsibility provides different perspectives about what makes a subject responsible for an action, it is fairly common to believe that being morally responsible is deeply connected to what it takes for that action to be an expression of the agent's will. Thinking of ways to increase people's agency, thus seems a good way both to respect and to increase their moral capacities (Reichlin 2017). The project is certainly ambitious, but some studies suggest it is feasible. For example, recent studies show that in some contexts our implicit biases can be changed easily (Beeghly & Madva 2020, part 1), while some authors suggest developing strategies enabling indirect control over such biases, such as through the development of a set of long-term habits or certain social policies (Beeghly & Madva 2020, part 3). In view of this, it is also useful to consider the problem of adaptive preferences highlighted by Elster and Sen: individual preferences are influenced by the social and environmental conditions in which humans are embedded, which is why their choices may often not be the ones they would make if they were more aware of their situation. Sometimes, by increasing information, the individual shows that she acts differently than she would have done by ignoring certain factors. Another interesting proposal, and not far from these considerations, comes from Gigerenzer, who stresses the possibility and importance of educating individuals to make the best possible decisions for themselves. Reconfirming the impossibility of Olympic rationality, Gigerenzer points out that intuitive, quick, and immediate mental processes are useful, often necessary, and capable of leading to optimal decisions if one has the right tools and knowledge to avoid falling victim to bias and to the way information is presented. Indeed, for the author, it is not only heuristics that lead us to erroneous conclusions and limit our evaluative ability, but also poor statistical education. Gigerenzer's proposal can be read from the perspective of developing interventions aimed at increasing subjective agency: the author creates teaching methods that enable even elementary school children to learn how to recognize and solve some Bayesian statistical problems that often underlie bad decisions (Gigerenzer 2008). In this framework, the goal is not to steer the mind, but to empower its cognitive and deliberative tools (Hertwig & Grüne-Yanoff 2017). Strengthening the subject's capacity for

conscious control enables her to consider her moral choices and actions in a more authentic dimension, reflexively increasing judgments of responsibility, praise, and blame. In this sense, interventions that act not only on moral behavior constitute true moral enhancements, because they improve either the individual's capacity for moral insight or his or her ability to weigh the reasons for and against a certain course of action and decide accordingly.

5. *Conclusion*

Considering a naturalistic conception of free will in relation to a theory of individual moral improvement makes one think about interventions that help individuals increase their opportunity to act, will, and choose consciously when exercising their moral capacities. It also highlights the importance of the agent reinforcing her own power or ability to judge something good or right, without having external judges deciding for her, regardless of her inclinations. There are great differences among people, which is why it does not seem attractive to advocate a position aimed at providing a guide that applies to everyone: such a guide would destroy some of the conditions necessary for freedom and, ultimately, for achieving real individual moral improvement. Moreover, such a position would leave open the question as to who might be entitled to establish what is right or wrong, good or bad.

Making information more comprehensible, educating individuals to recognize their own limitations and mistakes, and helping them improve their ability to map decisions, can be considered some of the ways to increase the positive aspect of the individuals' free will. As this is related to being able to control one's own choices and will, giving information, increasing the opportunity to receive feedback, providing spaces for discussion, and developing *ad hoc* education programs can be interventions aimed at helping the individual make more informed decisions. In this sense, it seems that «the best, more promising methods we have of moral enhancement are [...] traditional ones: education, parental and peer group guidance, social and personal example, and indeed reflection on what's rights, namely ethics» (Harris 2016, p. 117).

Focusing on the agency does not exclude the possibility of educating subjects, even wanting to try to persuade them to make some choices rather than others; however, it leads one to reflect on what to consider moral

enhancers. Viewing moral empowerment as indirect (Schaefer 2015), from a formal, procedural rather than a substantive perspective (Songhorian, Guma, Bina & Reichlin 2022), concerned with the capacities of individuals can lead to defining such empowerment in terms of an enhancement of free will. This would not wish to see situations realized in which individuals would be incapable of doing evil, but rather it would become possible to observe better moral agents because they are more capable of consciously choosing their judgments and actions.

References

Baldwin, R. 2014, "From regulation to behaviour change: giving nudge the third degree", *The Modern Law Review*, vol. 77, n. 6, pp. 831-857.

Bhargava, S., Loewenstein G.A. 2015, "Behavioral Economics and Public Policy 102: Beyond Nudging", *American Economic Review, Papers & Proceedings*, vol. 105, n. 5, pp. 396-401.

Beeghly, E., Madva, A. (Ed.) 2020, *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind*, Routledge, New York.

Bermùdez, J. (Ed.) 2018, *Self-Control, Decision Theory, and Rationality: New Essays*, Cambridge University Press, Cambridge.

Berlin, I. 1969, *Two Concepts of Liberty*, in I. Berlin, *Four Essays on Liberty*, Oxford University Press, London.

Gigerenzer, G. 2008, *Rationality for Mortals: How People Cope with Uncertainty*, Oxford University Press, New York.

Greene, J.D. 2013, *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*, Penguin, New York.

Guma, F. 2021, "Determinato ma possibilmente libero. La libertà del volere nella teoria psicanalitica", *Metapsychologica – Rivista di psicanalisi freudiana*, vol. 1, pp. 95-124.

John, P., Smith, G., Stoker, G. 2009, "Nudge Nudge, Think Think: Two Strategies for Changing Civic Behaviour", *The Political Quarterly*, vol. 80, n. 3, pp. 361-370.

Harris, J. 2016, *How to be Good: The Possibility of Moral Enhancement*, Oxford University Press, New York.

Hertwig, R., Grüne-Yanoff, T. 2017, "Nudging and Boosting: Steering or Empowering Good Decisions", *Perspectives on Psychological Science*, vol. 12, n. 6, pp. 1-14.

Kahneman, D. 2011, *Thinking, Fast and Slow*, Penguin Books, London.

Kahneman, D., Sibony, O., Sunstein, C.R. 2021, *Noise. A Flaw in Human Judgment*, Little Brown Spark, New York.

Klenk, M., Sauer, H. 2021, "Moral Judgement and Moral Progress: The Problem of Cognitive Control", Philosophical Psychology, vol. 34, n. 7, pp. 938-961.

List, C. 2019, *Why Free Will Is Real*, Harvard University Press, Cambridge.

Magni, S.F. 2019, *L'etica tra genetica e neuroscienze: libero arbitrio, responsabilità, generazione*, Carocci, Roma.

Nichols, S. 2004, *Sentimental Rules: On the Natural Foundations of Moral Judgment*, Oxford University Press, New York.

Quong, J. 2010, *Liberalism without Perfection,* Oxford University Press, Oxford.

Reichlin, M. 2017, "The Moral Agency Argument Against Moral Bioenhancement", Topoi, vol. 38, n. 1, pp. 53-62.

Sauer, H. 2017, *Moral Judgments as Educated Intuitions*, MIT Press, Cambridge (MA).

Schaefer, G.O. 2015. "Direct vs. Indirect Moral Enhancement", Kennedy Institute of Ethics Journal, vol. 25, n. 3, pp. 261-289.

Schubert, C. 2017, "Green Nudges: Do They Work? Are They Ethical?", *Ecological Economics*, vol. 132, pp. 329-342.

Songhorian, S., Guma, F., Bina, F., Reichlin, M. 2022, "Moral Progress: *Just* a Matter of Behavior?", *Teoria. Rivista di Filosofia*, vol. 2, pp. 175-187.

Soon, C.S., He, A.H., Bode, S., Haynes, J.D. 2013, "Predicting free choices for abstract intentions", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, n. 15, pp. 6217-6222.

Thaler, R.H., Sunstein, C.R. 2008, *Nudge: Improving Decisions about Health, Wealth, and Happiness*, Yale University Press, Yale.

Abstract

A relevant challenge in contemporary ethics is to understand whether and how individual moral improvement is feasible. Assuming the ineradicable presence of endogenous and exogenous conditions that make it difficult for the agent to control her action and choice (§1), I argue that theories developed from the observation of the limits of rationality and conscious control that human beings exercise over their decisions and actions are closely related to the question of free will (§2). I present two possible approaches to achieve individual moral improvement, showing their strengths and weaknesses. One proposal advocates nudges and suggestions to enhance people's

moral judgments (§3), whereas the other identifies ways to increase the subject's agency (§4). I conclude by arguing that developing procedures that can strengthen the subject's free will makes it possible to think of genuine and stable moral improvements because it generates would be enhancements not in any specific outward behaviors but in the individual's general moral attitude (§5).

Keywords: moral improvement; moral enhancement; free will; agency; nudging.

Francesca Guma
Vita-Salute San Raffaele University
guma.francesca@hsr.it

T

Federico Bina

Models of moral decision-making: Recent advances and normative relevance

1. *Dual-process models and the normative challenge*

Decades of experimental research have been regarded by many as supporting dual-process theories of human cognition, according to which two types of processes – one automatic (type 1), the other controlled (type 2) – are involved in the psychology of judgment and choice (Kahneman 2011; Evans & Stanovich 2013). Dual-process frameworks, however, are controversial, both in descriptive terms and for their potential normative implications. Specifically, disagreement persists about the interactions between type 1 and type 2 processes and their relative reliability. I will refer to the problem of drawing normative conclusions from a better understanding of decision processes as the *normative challenge*[1].

According to dual-process views, type 1 processes provide quick and efficient solutions to ordinary problems. However, these responses are often statistically inaccurate, biased, and unreliable in front of new and complex problems and decisions, due to their inflexible dependence on limited information and insensitivity to new and/or relevant ones (Kahneman 2011)[2]. On the contrary, type 2 operations are more flexible and sensitive to new and relevant information and changes in the decisional environment; they are also responsible for hypothetical thinking, simulation of alternatives,

[1] Evans calls unjustified inferences from description to normative conclusions about reasoning "normative fallacies" (Evans 2019).

[2] At least at the time of decision. As discussed below, type 1 processes are not completely inflexible, since they can significantly learn over time; the point is that they cannot be updated in real time.

 DOI: 10.4454/teoria.v42i2.166

and cost-benefit analyses (CBA). This, of course, requires higher computational costs.

The idea that these differences render type 2 more reliable than type 1 processes has been widely criticized. In particular, critics have emphasized a greater interaction between processes, suggesting that the dual-process image is not accurate (Kruglanski 2013) and that type 1 processes can be subject to sophisticated learning mechanisms, made sensitive to relevant information, and attuned to considered normative standards. Controlled processes can in fact be translated into automatic ones both implicitly and through exercise, as it happens for skill-acquisition and expertise in several domains (Hogarth 2001; Kahneman & Klein 2009). In light of their flexibility, penetrability, and ability to learn, it has been argued that type 1 processes should be considered very reliable in guiding decisions (Gigerenzer 2007).

In what follows, I will explain why these reasons are not sufficient to consider type 1 processes reliable, especially to address new and complex problems, and specifically in the moral domain. This claim is based on a vindicatory etiological and procedural reply to the normative challenge: the reliability of decision strategies is assessed in light of new (non-normative) understanding of the basic processes underlying their functioning, combined with relevant features – e.g. novelty, uncertainty, stakes – of the problems at hand.

2. *Dual-process moral cognition (beyond the reason/emotion divide)*

Dual-process models have been very influential also in recent (neuro) psychological research and empirically-informed ethical debates on moral judgment and decision-making. In the past two decades, empirical studies on (in)famous moral dilemmas have found correlations between characteristically deontological (D) responses and type 1 processes, while characteristically consequentialist (C) judgments correlate with type 2 reasoning (Conway & Gawronski 2013; Greene 2014; Patil et al. 2020).

A few scholars have concluded that these data support consequentialism as a normative theory (Greene 2014; Singer 2005). In sections 4 and 5, I suggest that this conclusion is problematic. Nonetheless, I will argue that empirical research and updated dual-process frameworks can still support significant conclusions for moral theory, though the nature of these conclusions is *procedural* rather than *substantive*.

A big part of the recent scientific and philosophical debate has questioned both Greene's dual-process account and the normative implications that he drew from it. Many critics have stressed that type 1 and 2 processes interact much more than Greene acknowledges; that empirical evidence does not show strong correlations between D judgments–type 1 processes and C judgments-type 2 reasoning; and that type 1 processes can learn and be reason-sensitive, attuned, educated, or trained. For these reasons, critics conclude, type 1 processes are more reliable than Greene maintains (Cecchini 2021; Sauer 2017; Railton 2014, 2017).

Although these claims are true from a descriptive point of view, inferring from them that type 1 processes are reliable in moral decision-making is problematic. As I formulated it, the normative challenge consists in understanding whether we are justified to infer normative conclusions from an increased understanding of the processes underlying moral judgments and decisions[3]. A more detailed description of these processes, therefore, might be of help.

In the past decades, dual-process frameworks have been characterized in several ways: fast vs. slow, automatic vs. controlled, unconscious vs. conscious, habitual vs. goal-oriented, affective vs. rational. I will focus here on a dual-process framework for morality which I believe to be more promising than others for several reasons (see section 3). First of all, this framework denies the problematic – though extremely common and influential – emotion/reason divide. Although this distinction has (historically) been a favorite way of philosophers to understand moral psychology, both critics and advocates of dual-process models have recognized that positing a clear distinction between emotions and reason (or affective and "cognitive" processes) is incorrect, since both type 1 and 2 processes always involve integrative information-processing as well as affective and motivational components (Saunders 2016)[4].

[3] Note that the same strategy is adopted by those who defend the higher reliability of type 1 processes: since they can learn and be sensitive to reasons – they argue – type 1 processes can be reliable.

[4] For instance, processes leading to C judgements do not just elaborate the factual information "5 is more than 1", but also affective elements leading to endorse, or choose, that "saving 5 lives is *better* than saving 1". Moreover, both D and C judgements involve factual information processing: D judgements and emotional reactions are always driven by a clear representation of structural features of the situation, such as personal interaction, the exercise of bodily force (Greene et al. 2009), or direct vs. indirect harm (Royzman & Baron 2002; Cushman et al. 2006).

Denying the emotion/reason distinction, however, does not mean leaving *any* dual-process accounts of moral cognition behind. Experimental research shows that two types of processes can be distinguished in moral as well as in non-moral decision-making, although framed in different ways, and portrayed as deeply interacting and cooperating.

3. *Action-outcome and computational frameworks*

A promising strand of dual-process models (Crockett 2013; Cushman 2013), relatively under-considered in the philosophical literature, frames moral cognition by stressing the distinction between:

1) Attributing value *directly to actions* by associating positive or negative value to them on the basis of a history of feedback (e.g. rewards or losses);
2) Attributing value to expected *outcomes* on the basis of a causal model (a "cognitive map") representing options, values, and transition functions.

These frameworks have two immediate advantages. First, they account for the presence of affective and cognitive information-processing in both types of processes; second, their reliance on learning models account for the diachronic dimension of moral cognition significantly more than first-wave dual-process models did.

These models are also consistent with several studies in moral psychology reporting a preference for indirect over direct harm (Rozyman & Baron 2002), strong aversion to typically harmful actions even when fake or victimless (Cushman et al. 2012; Haidt et al. 1993), and the systematic presence of moral norms across history and societies prescribing the wrongness of specific action-types independently of outcomes (e.g. rituals, food and sexual taboos) (see Graybiel 2008). In these cases, characteristically deontological responses are elicited by the value directly associated with actions, regardless of other relevant information, such as expected outcomes or empathic concern for the subjects involved.

In addition to this evidence, action-outcome frameworks are supported by recent research in computer science and computational neuroscience, reflecting the difference between two basic kinds of reinforcement learning: *model-free* and *model-based* algorithms (Dolan & Dayan 2013).

3.1. Model-free learning and decision-making

Model-free (MF) algorithms work by associating positive or negative value to specific and immediately available actions after a history of rewards, independently of a causal representation of the environment. Imagine an agent A who, when turning right in a state *r* (*round*), gets a reward. If this association occurs a significant number of times, A will associate a positive value to the option "turn right" when in *r* states. Now imagine that A reaches state *r* after turning left in a state *s* (*squared*). Since A associates positive value to state *r*, A will also associate positive value to the option "turn left" when in *s*; and so on, creating adaptive chains of actions.

This mechanism brings A to associate value to the available actions in each particular state on the track leading to a reward, treating each of them as if it was itself a reward. The main advantage of this algorithm is that it is computationally cheap: at each step, it decides on the basis of the value associated with the immediately available action, avoiding costly simulations of future or hypothetical states and comparisons between them. However, and precisely for this reason, MF algorithms are not farsighted. They cannot be goal-oriented – nor prospective in general – because they lack a causal representation of the relation between possible actions and outcomes. This precludes them from any chance to make plans at all: MF algorithms are fundamentally retrospective.

Moreover, although very efficient, MF algorithms are inflexible. They cannot use information to adjust values associated with states, actions, and outcomes (and, consequently, preferences and behavior) because they lack a global representation of them. Value representations can be updated, but this requires time, trial-and-error learning, or interference of strong opposing values (Dickinson et al. 1995).

3.2. Model-based learning and decision-making

By contrast, model-based (MB) algorithms choose by considering available courses of action on the basis of a causal representation – a model, or a cognitive map – of the environment. The model includes causal relations between events (actions, outcomes, rewards, and transition functions) to which A attributes different values; the expected values of the available options are compared, and choices are taken by exploring the decision tree and via CBAs (Dolan & Dayan 2013).

The main downside of this algorithm are its computational costs. Nonetheless, MB strategies can be very flexible, because the model can be updated at any moment by integrating new information and changes in the environment. Imagine that agent A has identified the optimal strategy to reach a reward. Knowing that an obstacle is obstructing the optimal policy (e.g. the fastest route) can make A choose the preferred alternative option in the most efficient way (e.g. without having to face the obstacle on the fastest route before finding an alternative). MB algorithms can be very farsighted, because they can identify clear and complex policies made of long chains of actions, simulating and evaluating consequences of consequences, and modulating value representation accordingly.

In human (moral) cognition, these two types of algorithms interact deeply (Cushman & Morris 2015; Kool et al. 2018). MF mechanisms do not only regulate motor habits or personal harm-aversion, but also the application of rules, principles, and concepts (Dayan 2012); they also facilitate MB decision-making by proposing limited sets of possibilities, thus avoiding the consideration of potentially infinite options in deliberative planning (Phillips & Cushman 2017). But to what extent can the differences between these algorithms – and/or their interaction – be normatively significant?

4. *Addressing the normative challenge*

Greene (2017) argued that the MF-MB distinction provides further support for consequentialism[5]. Like fast-and-frugal heuristics, MF decision-making is generally reliable in front of ordinary contexts and problems, but «it would be a cognitive miracle if we had reliably good moral instincts about unfamiliar moral problems» (Greene 2014, 715). New, complex, and controversial moral problems require MB reasoning. Since empirical research shows strong correlations and similarities between MB thinking and consequentialism, Greene concludes that the latter is the best normative theory to address those kinds of problems.

[5] Greene (2014) illustrates this idea through the analogy with a camera's automatic vs. manual settings. As he noticed later, however, this analogy can be misleading because the automatic settings of standard cameras do not change after they leave the factory, whereas «people's "automatic settings" are constantly evolving through learning [...] The key point, however, is that at the time of decision one is stuck with the automatic settings that one has, regardless of how circumstances might have changed» (Greene 2017, 5).

Note that according to Greene – as for many other advocates of consequentialism – this does not mean that agents should engage in CBA *all the time* (Hare 1981; Brink 1989). MF decision-making can work well in many circumstances, but MB reasoning is more reliable when we have to decide about complex cases, as well as about moral principles, rules, procedures, decision strategies, and whether or not to trust our intuitions. Advocates of deontological and virtue theories, Greene argues, deny this, favoring forms of MF thinking such as reliance on norms or the moral perception of virtuous agents.

These conclusions are partly convincing, but also partly problematic. On the one hand, Greene addresses the normative challenge in a promising way. Consider the following characterization that Railton (2017) recently gave of moral inquiry. Unlike other domains (but similarly to science) the moral discourse aspires to overcome subjective, tribal, elitist, or esoteric points of view and interests by following procedures, and looking for understanding and justification that are impartial, general, consistent, authority-independent, shareable, thinking- and action-guiding, and non-instrumentally concerned with interests and reasons of those actually or potentially affected (Railton 2017, p. 173).

If this characterization is plausible, then the only decision strategy able to accomplish these tasks cannot but be MB reasoning. Consistency, for instance, would be impossible without a model representing the value associated with principles, actions, and outcomes. MB reasoning is also the only strategy allowing us to consider the interests and reasons of others beyond our natural and cultural inclinations, and to evaluate them critically in light of relevant information and alternative possibilities. Moreover, consistent and intersubjectively acceptable moral justifications (Songhorian et al., this volume) cannot but be MB. Referring to a model – models are non-perspectival by definition – is the only way to make one's reasons intelligible to others. Finally, MB reasoning is necessary to link immediately available actions with distant goals, and to consider alternative courses of action (Railton 2017).

On the other hand, however, the idea that the higher reliability of MB reasoning supports consequentialism is problematic. The empirical literature is partly inconsistent on this matter; there are, nonetheless, at least four reasons to doubt such a bold normative conclusion.

1) Studies on confidence and decision-time in moral decision-making suggest that non-C judgments might be the result of MB reasoning *also at the time of decision* (Koop 2013; Gürcay & Baron 2017; Bialek & De Neys 2017);

2) MB reasoning should not be identified uniquely with CBA in act-utilitarian terms, but rather as a broader reflective operation that considers i) information, potential courses of actions and outcomes, ii) intuitions, feelings, rules and principles, and iii) reasons, testing their reciprocal consistency and discarding recalcitrant options (Brink 1989; Campbell & Kumar 2012; Bazerman & Greene 2010).
3) D/non-C judgments can be justifiable even when they are the proximate output of MF processes. First of all, they can be the (distal) output of previous MB reasoning or rationalization. In some cases, justificatory reasons can even track some processes that led to the new "educated" intuition, even if these processes did not intervene at the time of decision (Sauer 2017; Kumar 2017).
4) Finally, also C judgments can be the result of MF processes (Bago & De Neys 2019). For instance, Trémolière and Bonnefon (2014) have shown that the higher the number of lives involved in sacrificial dilemmas, the more intuitive C judgments are. This suggests that C responses can be model-free too, requiring MB reasoning when they are more counterintuitive (Kahane 2012).

To sum up, empirical research and the MF-MB framework support important normative conclusions, though mostly in "procedural" terms, i.e. suggesting how we should think in front of complex or new decisions, and how to justify them. This, however, has no clear direct implications for normative ethical theory in a more substantive way.

5. *Facilitators, conflict detectors, and metacognition*

Some readers might still be unconvinced about the procedural normative conclusion that MB moral reasoning is more reliable than MF mechanisms to address new and complex moral problems. I will briefly consider two possible reasons in favor of this skepticism:

i) In a recent paper, Cecchini argued that default-interventionist models of moral cognition – according to which type 2 (MB) processes intervene to control, endorse, or reject type 1 (MF) outputs – are inaccurate because (MB) moral reflection *fundamentally depends* on (MF) intuitions (Cecchini 2021, 301). In fact, recent research suggests that:

i.i) MF mechanisms often *facilitate* MB reasoning, providing by default limited sets of options within potentially infinite ones (Phillips & Cushman 2017);

i.ii) MF mechanisms *detect conflicts* between intuitions, reasons, and non-moral information, signaling the need for further reflection (De Neys 2014).

Although these claims are descriptively true, by no means they constitute an objection to the normative conclusion defended here. Operations such as cognitive filtering and conflict detection are not intrinsically reliable: they might be based on, and lead to, either reliable learning histories and actions, or biased and unjustifiable ones[6].

Consider these two cases. First (i.i), agent A might not even consider being fair or kind to a member of a discriminated group, or engaging in sustainable behaviors, because these options might not be included in the default set provided by MF processes as a result of her learning history. Her habits are different and pretty inflexible; she can contemplate different possibilities, but she does not consider *those* actions since the value associated to them is significantly lower than alternatives available at the time of decision. Second (i.ii), intuitive conflict detection and resolution might result in discarding reasonable options (e.g. the less harmful, or the more supported by evidence) because too costly to hold; the conscious reasoning process called upon by intuitive conflict detection might be merely confirmatory of pre-reflective intuitions (Kunda 1990; Haidt 2001).

There is hence no reason to hold MF mechanisms trustworthy in the moral domain just because of their causal role: decisions are often driven by intuitive (MF) processes, but in no way this justifies them. On the contrary, the aforementioned limits of MF algorithms cast doubt on their outputs if no specific convergent support is provided by MB reasoning. In both the aforementioned cases, only MB strategies can critically evaluate whether to endorse the input provided by MF default options or to consider alternative ones. Moreover, only MB reasoning can test whether intuitions are reciprocally consistent and supported by reasons, independently of pre-reflective confidence about their rightness. Deciding uniquely based on the strength of "feelings" or "seemings" is not a defensible strategy (Brink 1989, ch. 5; Harris 2012, 294).

[6] In order to respect Railton's criteria for non-perspectival moral inquiry mentioned above – i.e. for being intersubjectively communicable, understandable and justifiable –, the normative standards needed to assess the reliability of cognitive processes and behavioral outputs cannot but be model-based.

ii) Finally, MF-type 1 mechanisms have been indicated as responsible for the meta-cognitive task of deciding whether MF or MB strategies should be implemented to address specific problems (Cecchini 2021; Thompson et al. 2011)[7]. However, recent studies suggest that when facing a problem, people often engage in CBA weighing the expected outcomes of each strategy (including, in the calculation, the computational costs of MB reasoning), rather than relying on heuristics. Specifically, data show that engagement in MB reasoning – both as metacognitive arbitrator and as the ultimate decision strategy – is proportional to the stakes and levels of uncertainty involved (Kool et al. 2017, 2018). These results are consistent with previous research suggesting that at each time point agents estimate the expected costs and rewards from engaging in a full MB estimation of action-outcome values (Keramati et al. 2011). Although MF processes do play a role in this arbitration, there is no reason for holding them reliable detectors of the right decision mode for specific and complex problems (Bazerman & Greene 2010).

6. *Conclusions*

In this paper I argued that dual-process models of moral cognition are plausible, though they should not be framed in terms of the problematic emotion/reason dichotomy. I also suggested that the distinction between model-free and model-based learning and decision-making algorithms can lead us to draw important normative conclusions. Specifically, in light of a) how they function, and b) the problems we have to face, this framework supports the higher reliability of model-based moral decision-making in front of new, uncertain, and/or complex scenarios. Reliability can be conceived of in terms of justifiability: people would more likely provide – and freely accept – good moral justifications based on non-perspectival model-based reasons, rather than on the subjective "feeling" or "smell" of what is right (although this latter strategy can give rise to effective *post-hoc* rationalizations; see Songhorian et al., this volume).

These conclusions, however, are procedural rather than substantive. Indeed, model-based moral reasoning should not be seen as merely eval-

[7] Evans (2019) hypothesizes a 'type 3' process for this task, presenting aspects of similarity with both type 1 and type 2 processes.

uating outcomes (Cushman 2013), nor as a kind of purely consequentialist form of thinking (Greene 2017), since it can be open to the consideration of several non-consequentialist reasons, norms, intuitions and evaluations (Białek & De Neys 2017). The coherentist mechanism needed to balance all these considerations is a form of model-based reasoning, though it looks closer to a reflective equilibrium than to a pure cost-benefit analysis.

References

Bago B., De Neys W. 2019, The intuitive greater good: Testing the corrective dual process model of moral cognition, *Journal of Experimental Psychology: General*, 148(10), 1782.

Bazerman M.H., Greene J.D. 2010. In favor of clear thinking: Incorporating moral rules into a wise cost-benefit analysis, *Perspectives on Psychological Science*, 5(2), 209-212.

Białek M., De Neys W. 2017, Dual processes and moral conflict: Evidence for deontological reasoners' intuitive utilitarian sensitivity, *Judgment and Decision Making*, 12(2), 148.

Brink D.O. 1989, *Moral realism and the foundations of ethics*, Cambridge University Press.

Campbell R., Kumar V. 2012, Moral reasoning on the ground, *Ethics*, 122(2), 273-312.

Cecchini D. 2021, Dual-process reflective equilibrium: rethinking the interplay between intuition and reflection in moral reasoning, *Philosophical Explorations*, 24(3), 295-311.

Conway P., Gawronski B. 2013, Deontological and utilitarian inclinations in moral decision making: a process dissociation approach, *Journal of Personality and Social Psychology,* 104(2), 216.

Crockett, M.J. 2013, Models of morality, *Trends in cognitive sciences*, 17(8), 363-366.

Cushman F. 2013, Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, 17(3), 273-292.

Cushman F., Gray K., Gaffey A., Mendes W.B. 2012, Simulating murder: the aversion to harmful action, *Emotion*, 12(1), 2.

Cushman F., Morris A. 2015, Habitual control of goal selection in humans, *Proceedings of the National Academy of Sciences*, 112(45), 13817-13822.

Cushman F., Young L., Hauser M. 2006, The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm, *Psychological Science*, 17(12), 1082-1089.

Dayan P. 2012, How to set the switches on this thing, *Current Opinion in Neurobiology*, 22(6), 1068-1074.

De Neys W. 2014, Conflict detection, dual processes, and logical intuitions: Some clarifications, *Thinking & Reasoning*, 20(2), 169-187.

Dickinson A., Balleine B., Watt A., Gonzalez F., Boakes R.A., 1995, Motivational control after extended instrumental training, *Animal Learning & Behavior*, 23(2), 197-206.

Dolan R.J., Dayan P. 2013, Goals and habits in the brain, *Neuron*, 80(2), 312-325.

Evans J.S.B., 2019, Reflections on reflection: the nature and function of type 2 processes in dual-process theories of reasoning, *Thinking & Reasoning*, 25(4), 383-415.

Evans J.S.B., Stanovich, K.E. 2013, Dual-process theories of higher cognition: Advancing the debate, *Perspectives on Psychological Science*, 8(3), 223-241.

Gigerenzer G. 2007, *Gut feelings: The intelligence of the unconscious*. Penguin.

Graybiel A.M. 2008, Habits, rituals, and the evaluative brain, *Annual Review of Neuroscience*, 31(1), 359-387.

Greene J.D. 2014, Beyond point-and-shoot morality: Why cognitive (neuro) science matters for ethics, *Ethics*, 124(4), 695-726.

Greene J.D. 2017, The rat-a-gorical imperative: Moral intuition and the limits of affective learning, *Cognition*, 167, 66-77.

Greene, J.D. Cushman F.A., Stewart L.E., Lowenberg K., Nystrom L.E., Cohen J.D. 2009, Pushing moral buttons: The interaction between personal force and intention in moral judgment, *Cognition*, 111(3), 364-371.

Gürçay B., Baron J. 2017, Challenges for the sequential two-system model of moral judgement, *Thinking & Reasoning*, 23(1), 49-80.

Haidt J. 2001, The emotional dog and its rational tail: a social intuitionist approach to moral judgment, *Psychological Review*, 108(4), 814-834.

Haidt J., Koller S.H., Dias M.G. 1993, Affect, culture, and morality, or is it wrong to eat your dog?, *Journal of Personality and Social Psychology*, 65(4), 613.

Hare R.M. 1981, *Moral thinking: Its levels, method, and point*. Oxford University Press.

Harris J. 2012, What it's like to be good, *Cambridge Quarterly of Healthcare Ethics*, 21(3), 293-305.

Hogarth R.M. 2001, *Educating Intuition*, University of Chicago Press.

Kahane G., Wiech K., Shackel N., Farias M., Savulescu J., Tracey I. 2012, The neural basis of intuitive and counterintuitive moral judgment, *Social Cognitive and Affective Neuroscience*, 7(4), 393-402.

Kahneman D. 2011, *Thinking, fast and slow*. Macmillan.

Kahneman D., Klein G. 2009, Conditions for intuitive expertise: a failure to disagree, *American Psychologist*, 64(6), 515.

Keramati M., Dezfouli A., Piray P. 2011, Speed/accuracy trade-off between the habitual and the goal-directed processes, *PLoS Computational Biology*, 7(5), e1002055.

Kool W., Gershman S.J., Cushman F.A. 2017, Cost-benefit arbitration between multiple reinforcement-learning systems, *Psychological Science*, 28(9), 1321-1333.

Kool W., Cushman F.A., Gershman S.J. 2018, Competition and cooperation between multiple reinforcement learning systems, in Morris, R.W., Bornstein, A., & Shenhav, A. (Eds.), *Goal-directed decision making: Computations and neural circuits*, Academic Press, 153-178.

Koop G.J. 2013, An assessment of the temporal dynamics of moral decisions, *Judgment and Decision Making*, 8(5), 527.

Kruglanski A.W. 2013, Only one? The default interventionist perspective as a unimodel, *Perspectives on Psychological Science*, 8(3), 242-247.

Kumar V. 2017, Moral vindications, *Cognition*, 167, 124-134.

Kunda Z.1990, The case for motivated reasoning, *Psychological Bulletin*, 108(3), 480-498.

Patil I., Zucchelli M.M., Kool W., Campbell S., Fornasier F., Calò M., Cikara M., Cushman, F. 2021, Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures, *Journal of Personality and Social Psychology*, 120(2), 443-460.

Phillips J., Cushman F. 2017, Morality constrains the default representation of what is possible, *Proceedings of the National Academy of Sciences*, 114(18), 4649-4654.

Railton P. 2014, The affective dog and its rational tale: Intuition and attunement. *Ethics*, 124(4), 813-859.

Railton P. 2017, Moral learning: Conceptual foundations and normative relevance, *Cognition*, 167, 172-190.

Royzman E.B., Baron J. 2002, The preference for indirect harm, *Social Justice Research*, 15(2), 165-184.

Sauer H. 2017, *Moral judgments as educated intuitions*. MIT Press.

Saunders L.F. 2016, Reason and emotion, not reason or emotion in moral judgment, *Philosophical Explorations*, 19(3), 252-267.

Singer P. 2005, Ethics and intuitions, *The Journal of Ethics*, 9(3), 331-352.

Songhorian S., Guma F., Bina F., Reichlin M. 2022, Moral progress: Just a matter of behavior?, *forthcoming*.

Thompson V.A., Turner J.A. P., Pennycook G. 2011, Intuition, reason, and metacognition, *Cognitive psychology*, 63(3), 107-140.

Trémolière B., Bonnefon J.F. 2014, Efficient kill–save ratios ease up the cognitive demands on counterintuitive moral utilitarianism, *Personality and Social Psychology Bulletin*, 40(7), 923-930.

Abstract

In the last decades, research in cognitive psychology and neuroscience fueled a rich debate about i) the main mechanisms underlying human (moral) decision-making and ii) their reliability. In this paper, I first make clear that the emotion/reason distinction should be set aside, although this does not imply casting doubt on dual-process models in general. To support this idea, I discuss a dual-process framework for moral decision-making informed by computational models of reinforcement learning. I finally consider some normative implications of this research, stressing their procedural, rather than substantive, nature.

Keywords: dual-process; moral cognition; reinforcement learning; intuition; consequentialism

Federico Bina
Vita-Salute San Raffaele University
f.bina@studenti.unisr.it

Edizioni ETS
Palazzo Roncioni - Lungarno Mediceo, 16, I-56127 Pisa
info@edizioniets.com - www.edizioniets.com
Finito di stampare nel mese di dicembre 2022