Etica, diritto e scienza cognitiva

Ethics, Law, and Cognitive Science

TEORIA

Rivista di filosofia
fondata da Vittorio Sainati
XXXVII/2017/2 (Terza serie XII/2)

Edizioni ETS

«Teoria» è indicizzata ISI Arts&Humanities Citation Index e SCOPUS, e ha ottenuto la classificazione “A” per le riviste di Area 11, settori 11/C1, 11/C2, 11/C3, 11/C4, 11/C5.

La versione elettronica di questo numero
è disponibile sul sito: www.rivistateoria.eu

Direzione e Redazione: Dipartimento di civiltà e forme del sapere
dell’Università di Pisa, via P. Paoli 15, 56126 Pisa,
tel. (050) 2215500 - www.cfs.unipi.it

Direttore: Adriano Fabris

Amministrazione: Edizioni ETS, piazza Carrara 16-19, 56126 Pisa,
www.edizioniets.com, info@edizioniets.com - tel. (050) 29544-503868

Abbonamento: Italia € 40,00, Iva inclusa; estero € 50,00,
(Iva inclusa, spese di spedizione incluse)
da versare sul c.c.p. 14721567 intestato alle Edizioni ETS.
Prezzo di un fascicolo: € 20,00, Iva inclusa.
Prezzo di un fascicolo arretrato: € 30,00, Iva inclusa.

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.

Iscritto al Reg. della stampa presso la Canc. del Trib. di Pisa
n° 10/81 del 23.5.1981. Direttore Responsabile: Adriano Fabris.
Semestrale. Contiene meno del 70% di pubblicità.

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.

TEORIA

T

Rivista di filosofia
fondata da Vittorio Sainati
XXXVII/2017/2 (Terza serie XII/2)

Etica, diritto e scienza cognitiva

Ethics, Law, and Cognitive Science

Edizioni ETS

Indice/Contents

Premise / Premessa

Since the second half of the 20th century, the evolution of the theoretical reflection on ethics has gone through several stages. A first phase, between the 1950s and 1960s, was characterized by the prevalence of metaethics, namely, the investigation of the metaphysical, epistemological and semantic aspects of moral concepts and properties. During the 1970s and 1980s, the focus shifted to normative ethics, namely, the attempt to provide a general theory that tells us how we ought to live. Then, during the 1990s, problems of practical or applied ethics became the main focus: bioethics, environmental ethics, communication ethics, animal ethics, business and professional ethics, and so forth. Finally, in the new millennium, applied ethics became less capable of giving substance to theoretical reflection, enabling a return to new forms of metaethics, with questions concerning the nature of moral concepts, the criteria of soundness of the argumentation in ethics, and the implications of the recent scientific findings for moral matters once again taking centre stage (cf. Lecaldano, 2012).

The upsurge of interest in metaethics gave rise to a series of research programs opposing the previous lack of interest in experimental sciences and the demand for a synthesis among ethics, biology, psychology and neuroscience. The transformation of moral psychology is a case in point. Until recently, philosophers and psychologists tended to cultivate moral psychology independently from each other: the former considered it a branch of metaethics, while the latter viewed it as a branch of scientific psychology. As a result, the type of moral psychology cultivated in philosophy was almost entirely speculative, even when the issues addressed (e.g., the existence of the global character traits apparently posited by virtue ethicists, or the role of reason versus emotion in moral judgment and motivation) unquestionably turned on empirical assumptions about the mind. Conversely, the type of moral psychology developed by psychologists, although endowed with experimental evidence, rarely took into account the complexity and subtleties of the categories via which philosophers had investigated ethics in the past. As previously said, however, there have been many recent attempts to overcome such methodological barriers. Increasingly philosophers have joined forces with psychologists in pursuing the project of a moral psychology that is simultaneously experimental and philosophical (cf. Doris, 2012; Doris-Stich, 2014; May, 2017).

The synergy between ethics and science has been enhanced by a wealth of data from cognitive neuroscience. Over the last three decades, cognitive science has expanded "vertically" into the brain, placing neuroscience at its forefront. A powerful engine of this expansion has been cognitive neuroscience, i.e., the attempt to link psychological functions with neural structures by developing mechanistic explanations of cognitive processes (cf. Marraffa-Paternoster, 2017). However, neuroscientific studies of moral beliefs, emotions, and decisions were not possible until the 1990s. Consequently, it was only in the early 2000s that cognitive neuroscience started to influence ethics, gradually resulting in a brand-new field termed "the neuroscience of morality (or ethics)" (cf. Sinnott-Armstrong, 2008c). The latter concerns what neuroscientific research tells us about core questions in metaethics, normative ethics and philosophy of law (e.g., Can neuroscience support moral emotivism and undermine moral rationalism? Does neuroscience undermine free will or moral and legal responsibility?). According to Roskies (2016; cf. also Farah, 2010; Glannon, 2011; Illes-Sahakian, 2011; Clausen-Levy, 2015), the neuroscience of morality should be considered a part of *neuroethics* together with the *ethics of neuroscience*, which instead encompasses questions similar to those in the field of applied ethics (e.g., Do brain reading technologies violate privacy? Do we have an obligation to enhance ourselves by altering our brains?).

We can thus conceive empirical moral psychology and the neuroscience of morality as two disciplines under the umbrella of the emerging research area of "the cognitive science of morality" (cf. Sinnott-Armstrong, 2008b). This new area incorporates findings from neuroscience, developmental psychology, evolutionary psychology, social psychology, evolutionary biology, experimental economics, cross-cultural anthropology, and even primatology, allowing moral philosophers to theorize on issues such as moral development (e.g., Bloom, 2013), the nature of character (e.g., Doris, 2014), the kinds of neurocognitive processes that generate our moral intuitions (e.g., Greene, 2013), the evolutionary origins of our moral capacities (e.g., Joyce, 2006; Sinnott-Armstrong, 2008a), cross-cultural differences in moral norms (e.g., Haidt, 2012), and so on.

An undeniable contribution of the cognitive science of morality is the afore-mentioned new understanding of metaethics. A metaethical theory that is informed by research in cognitive science (or, more radically, that it is itself part of cognitive science) very positively switches the focus from the more traditional analysis of the language of morals to the workings of the moral mind. However, whereas the contribution of the cognitive science of morality to metaethics is clear and substantial, the role of cognitive science with regard to *normative* ethics is more difficult to assess.

Even if the ban on deriving an "ought" from an "is" should not be taken as philosophical dogma, the normative use of cognitive science findings about human moral minds is a slippery task. In our view, the pursuit of a synthesis between philosophical reflection and scientific inquiry does not force us to endorse the ideal of a complete naturalization of normative ethical questions. Rather, a reasonable naturalistic perspective on normative ethics should aim to assess the admissibility of moral theories against what science can tell us about how real people think, feel, and behave. This assumption has been characterized as a *principle of minimal psychological realism*: "Make sure when constructing a moral theory or projecting a moral ideal that the character, decision processing, and behavior prescribed are possible, or are perceived to be possible, for creatures like us" (Flanagan, 1991, p. 32), In other words, the principle asserts that possible moral theories must have possible moral psychologies.

An overview of this issue

The purpose of this issue is to explore the relevance of the cognitive science of morality for a variety of topics in metaethics, normative ethics, applied ethics, and philosophy of law. A first group of articles is concerned with how recent cognitive science findings affect our practices of attributing moral and legal responsibility.

What Neuroscience will tell us about Moral Responsibility by Daniel C. Dennett is an elegant reflection on determinism, moral and legal responsibility and punishment from the perspective of neuroscience. Dennett concisely but effectively argues that compatibilist free will gives us everything we need to be morally responsible and allows us to maintain a moderately retributivist line of thinking.

In *Neurolaw and Punishment: A Naturalistic and Humanitarian View, and its Overlooked Perils*, Andrea Lavazza argues for a position that is similar to Dennett's. The author scrutinizes a progressive naturalization of criminal law that builds on data from neuroscience. These data are interpreted as challenging our capacity of knowingly, voluntarily and consciously undertaking a course of action by choosing between alternatives. As such, they can be seen as a demonstration of the illusoriness of moral responsibility that opens the way to more humane forms of punishment, which are justified on purely utilitarian grounds. Taking a stance against this purely consequentialist view of punishment, the author suggests a more nuanced assessment of the relevant neuroscientific data and defends a moderate form of retributivism.

In *Responsibility and Control in a Neuroethical Perspective*, Elisabetta Sirgiovanni grapples with the so-called "Frail Control Hypothesis" in the same anti-radical, reformist vein as Lavazza. This hypothesis has it that people are far less in control than they suppose, given the influence of unconscious situational factors. The hypothesis is a threat for the folk notion of responsibility, to the extent that folk psychology sees conscious control as the *sine qua non* of responsible agency. The author considers possible solutions to this threat and discusses objections to all of them. She then provides some suggestions for building an account of responsibility that unifies the benefits of the different solutions while taking their limitations into consideration.

In *Responsibility and the Relevance of Alternative Future Possibilities*, Felipe De Brigard shifts the focus onto the blossoming empirical investigation of "folk morality", especially as manifested in people's beliefs about free will and responsibility. For the most part, these vignette-based studies are exclusively focused on participants' judgments of the causal history of the events leading up to an agent's action and considerations about what the agent could have done differently in the past. However, recent evidence suggests that, when judging whether or not an individual is responsible for a certain action – even in concrete, emotion-laden and fully deterministic scenarios – considerations about alternative future possibilities may become relevant. The author reviews this evidence and suggests a way of interpreting the nature of these effects as well as some consequences for experimental philosophy and the psychology of free will and responsibility.

In recent decades, evolutionary biology, sociology and behavioral economics have productively interacted with psychological sciences, making it increasingly clear that human beings are *naturally* inclined to competition, and sometimes destructivity, but also to forms of sociality, cooperation and even altruism. In *Social Justice, Individualism, and Cooperation*, Mario De Caro and Benedetta Giovanola explore the contribution that this literature can offer to the field of political philosophy. In particular, the authors argue that, in order to make the reflection on *social justice* more reliable and effective, political philosophers must take into account the anthropological model emerging from what cognitive sciences tell us about self-assertiveness, egoism, competition, pro-sociality, cooperation and altruism.

In *Biology, Ethics and Moral Reflection*, Simone Pollo suggests a specific way to link the cognitive science of morality and normativity. Rather than being a direct source of norms and values, the understanding of moral psychology carried out by cognitive science contributes to the task of moral reflection insofar as it is a form of self-understanding. Part of the practice of moral reflection – i.e., critically weighing up and evaluating our own habits, attitudes and moral responses – is the understanding of our own nature, both as a specific individual and as a member of the human species. The author's aim is to discuss whether the cognitive science of morality could be regarded as a modern answer to the ancient exhortation "know thyself", and whether scientific advancements in this area could lead to moral progress.

According to contemporary (neuro)cognitive science, moral beliefs are decisively dependent on emotions. Our moral beliefs both actively influence and appear as necessary and sufficient conditions for moral judgments. Results have suggested an original view of the nature of ethics, according to which moral concepts are essentially related to emotions (epistemic emotionism); and moral properties consist of emotional facts (metaphysical emotionism). This view, coupled with the cultural relativity of human emotions and sentiments, generates a powerful argument in favor of ethical relativism. In *Emotions and Morality: Is Cognitive Science a Recipe for Ethical Relativism?* Massimo Reichlin argues (1) that, as far as epistemic emotionism is concerned, this account does not demonstrate that the right order of causation proceed in all cases from emotions to judgments; does not disprove the possibility of dispassionate judgments; has no persuasive explanation of the distinction between moral and conventional rules; cannot account for autistic morality; and 2) that, as far as metaphysic emotionism is concerned, this account offers a much too deflationary account of moral disagreement. The latter can be best understood within a realistic account of the facts (including pro-attitudes such as emotions and sentiments) that provide the best reasons for action.

In *Lockean Persons, Self-Narratives, and Eudaimonia*, Rossella Guerini and Massimo Marraffa explore the ethical import of a naturalistic form of narrative constructivism that distances itself from both the non-naturalistic and antirealist strands in theorizing on the self. Their criticism builds on William James' theory of the self. Against this Jamesian backdrop, the claim that we constitute ourselves as morally responsible agents (as "Lockean persons") by forming and using autobiographical narratives is combined with the realist claim that the narrative self is not an idle wheel but a layer of personality that serves as a *causal* center of gravity in the history of the human psychobiological system. This alliance between narrative constructivism and self-realism takes shape in the context of a tradition of thought that views the synthesis of the various strata of personality as the highest developmental point of the selfing process – a viewpoint that aligns with an ethic that hinges on the idea of *eudaimonia*: the discovery and actualization of our unique potentials and talents.

"Implicit bias" is a term of art referring to the relatively unconscious and relatively automatic features of prejudiced judgment and social behavior. In *Category Matters: The Interlocking Epistemic and Moral Costs of Implicit Bias*, Lacey J. Davidson rejects the claim that social categories are or should be irrelevant to our evaluations of individuals. She provides evidence against this claim by denying its empirical plausibility, emphasizing the epistemic and moral benefits that may come from social categories. Throughout the paper, she emphasizes the unique interlocking of epistemic and moral considerations with respect to implicit bias. The author's hope is that this analysis may lay the groundwork for an account of the right ways in which social categories can impact our judgments – i.e., the ways in which such impacts may improve rather than diminish our epistemic and moral situations.

Recently, philosophers have appealed to empirical studies to argue that whenever we think about a proposition *p*, we automatically believe *p*. (This view of belief formation is often called "the Spinozian theory", as Spinoza is thought to be the first who defended it.) Levy and Mandelbaum (2014) have gone further in claiming that the automaticity of believing has implications for the ethics of belief, in that it creates epistemic obligations for those who know about their automatic belief acquisition. Uwe Peters, in *On the Automaticity and Ethics of Belief*, uses theoretical considerations and psychological findings to raise doubts about the empirical case for the Spinozian theory.

In *The Ethical Convenience of Non-Neutrality in Medical Encounters: Argumentative Instruments for Healthcare Providers*, Maria Grazia Rossi, Sarah Bigi and Daniela Leone explore ethical questions regarding communication by considering the asymmetry of the doctor-patient relationship in an institutional setting. In this well-defined professional context, there is considerable debate about the ethically relevant topic of healthcare providers' neutrality. The authors argue that it is possible and desirable to adopt and manage non-neutral communication strategies to safeguard patients' freedom and autonomy in making decisions. To deal with the topic of neutrality on the communicative level, they use a normative argumentative model of communication, focusing on its effectiveness as a communicative instrument for healthcare providers.

In *"Publicity", Privacy and Social Media. The Role of Ethics above and beyond the Law*, Veronica Neri begins by noting that social media plays an increasingly important role in the relationship between ethics and the law. The author raises new issues regarding the concepts of both "publicity" (in the etymological sense of "making public"), and privacy. The limits of both the law and of deontology are becoming more and more evident in relations established via the social media, with a resultant need for ethical reflection, focusing on the motivation that leads users to convey certain information – beginning with the desire to "spectacularize" their lives – as well as the possible principles that may help guide informed choices. Among these, the concept of "responsible freedom", and consideration of the possible consequences arising as a result of certain choices – consequences for both ourselves and other individuals, on social media as well as in our off-line day-to-day lives – appear fundamental.

These brief summaries cannot, of course, do justice to the rich empirical detail, careful philosophical arguments, and variety of profound issues that arise in these articles. Taken together, however, they reveal the extent to which contemporary cognitive science is able to contribute to moral theory.

Mario De Caro, Massimo Marraffa

Dalla metà del Novecento a oggi l'evoluzione della riflessione teorica sull'etica ha attraversato varie fasi. Una prima fase, tra gli anni Cinquanta e Sessanta, è stata caratterizzata dalla prevalenza della *metaetica* ossia la riflessione sugli aspetti metafisici, epistemologici e semantici dei concetti e delle proprietà morali. Negli anni Settanta e Ottanta l'attenzione si è invece spostata sull'*etica normativa*, vale a dire la disciplina che si propone di delimitare l'ambito dell'agire moralmente corretto. Negli anni Novanta si è poi attraversata una fase in cui sono venuti in primo piano i problemi dell'*etica pratica o applicata*: bioetica, etica ambientale, etica della comunicazione, etica degli animali, etica degli affari e delle professioni e così via. Infine, in questo secolo, ridottasi la capacità delle etiche applicate di infondere concretezza alla riflessione teorica, si è avuto un ritorno, da diverse prospettive, ai problemi della metaetica. Hanno così riguadagnato centralità gli interrogativi sulla natura dei concetti morali, sui criteri di correttezza delle procedure argomentative in campo etico e sulle ricadute delle più recenti acquisizioni scientifiche sugli interrogativi morali (cfr. Lecaldano 2012).

Il ritorno di interesse per la metaetica ha generato una serie di programmi di ricerca che al precedente disinteresse verso le scienze sperimentali hanno opposto l'esigenza di una sintesi fra etica, biologia, psicologia e neuroscienza. Ciò è ben illustrato dalla traiettoria della *psicologia morale*, la disciplina che indaga genesi e sviluppo delle credenze e delle motivazioni su cui si basa l'agire morale. Fino non molti anni fa, filosofi e psicologi tendevano a sviluppare questa disciplina in completa indipendenza gli uni dagli altri: mentre i primi, infatti, intendevano questa disciplina come una branca della metaetica, per i secondi era una branca della psicologia sperimentale. Di conseguenza, se in ambito filosofico la psicologia morale aveva un carattere quasi esclusivamente speculativo – e ciò anche quando i temi trattati (per esempio, il nesso tra moralità e carattere oppure la natura del ragionamento morale) avevano un evidente contenuto empirico. Viceversa, la psicologia morale sviluppata dagli psicologi, benché fondata empiricamente, ben raramente teneva conto della complessità e della sottigliezza delle categorie con cui per secoli i filosofi hanno indagato l'ambito etico. Tuttavia, come detto, negli ultimi anni sono stati sviluppati molti tentativi per superare queste barriere metodologiche: e così, sempre più spesso, filosofi e psicologi lavorano assieme al progetto di una psicologia morale che sia allo stesso tempo empirica e filosofica (cfr. Doris, 2012; Doris-Stich, 2014; May, 2017).

Nel frattempo, inoltre, la sinergia tra etica e scienza si è ulteriormente arricchita, grazie ai copiosi e rilevanti risultati che provengono dalle neuroscienze. Nel corso degli ultimi tre decenni la scienza cognitiva si è espansa “verticalmente”, verso il cervello, collocando le neuroscienze in una posizione di assoluta centralità. Un potente motore di questa espansione è stata la *neuroscienza cognitiva*, vale a dire il progetto di istituire un nesso tra le funzioni psicologiche e le strutture neuronali sviluppando spiegazioni meccanicistiche dei processi cognitivi (cfr. Marraffa-Paternoster, 2017). E tuttavia non vi è stata la possibilità di condurre indagini neuroscientifiche su temi quali le credenze, le emozioni e le decisioni morali fino alla fine degli anni novanta; di conseguenza solo all’inizio del nuovo millennio la neuroscienza cognitiva ha iniziato a influenzare l’etica, dando luogo gradualmente a un settore completamente nuovo: la *neuroscienza dell’etica* (cfr. Sinnott-Armstrong, 2008c). Oggetto di questa nuova area di ricerca è ciò che l’indagine neuroscientifica ci dice in merito alle domande fondamentali della metaetica, dell’etica normativa e della filosofia del diritto (per esempio, “la neuroscienza fornisce dati in favore dell’emotivismo morale e revoca in dubbio il razionalismo morale?”; oppure “la neuroscienza mette in discussione il libero arbitrio e la responsabilità morale e legale?”). A giudizio di Roskies (2016; cfr. anche Farah, 2010; Glannon, 2011; Illes-Sahakian, 2011; Clausen-Levy, 2015), la neuroscienza dell’etica è parte della *neuroetica*, nell’ambito della quale si affianca all’*etica della neuroscienza*, che si occupa invece di problemi simili a quelli sollevati nell’ambito dell’etica applicata (per esempio, “le tecnologie che consentono la lettura dell’attività cerebrale violano la privacy?”; oppure “abbiamo l’obbligo di potenziare noi stessi modificando il nostro cervello?”).

Psicologia morale empirica e neuroscienza dell’etica possono allora essere viste come due discipline costituenti la fiorente area di ricerca della “scienza cognitiva dell’etica” (cfr. Sinnott-Armstrong, 2008b). In questo settore di ricerca entriamo in contatto con dati provenienti dalla biologia evoluzionistica, la neuroscienza, la psicologia evoluzionistica, la psicologia dello sviluppo, la psicologia sociale, l’economia sperimentale, la psicologia culturale e la primatologia cognitiva; questi dati consentono ai filosofi di teorizzare su problemi quali lo sviluppo morale (per es. Bloom, 2013), la natura del carattere (per es. Doris, 2014), i processi neurocognitivi alla base delle intuizioni morali (per es. Greene, 2013), le origini filogenetiche delle nostre capacità morali (per es. Joyce, 2006; Sinnott-Armstrong, 2008a), le differenze interculturali nelle norme morali (per es. Haidt, 2012); e così via.

Un contributo innegabile della scienza cognitive dell’etica è la nuova concezione della metaetica a cui si è già accennato. Una teoria metaetica informata dalle indagini degli scienziati cognitivi (o, più radicalmente, che è essa stessa parte della scienza cognitiva) sposta assai opportunamente l'attenzione dall'analisi tradizionale del linguaggio della morale al funzionamento della mente morale. Tuttavia, se il contributo che la scienza cognitiva dell'etica apporta alla metaetica è chiaro e innegabile, il suo ruolo nei riguardi dell'etica normativa è più difficile da valutare.

Anche se non si considera il divieto di derivare le norme dai fatti come un imperativo filosofico, l'uso normativo di ciò che le scienze cognitive ci insegnano in merito alla mente morale è un'impresa piena di insidie. A nostro parere, il perseguimento di una sintesi fra riflessione filosofica e indagine scientifica non implica che si debba aspirare a una completa (e oltremodo ipotetica) naturalizzazione del piano etico-normativo. Piuttosto, a nostro avviso, l'obiettivo di una prospettiva naturalistica *ragionevole* sull'etica normativa deve essere di vincolare l'accettabilità delle teorie morali a ciò che è stato definito "principio di realismo psicologico minimo": chi costruisce una teoria morale deve essere ben sicuro che quanto prescrive sia possibile per «creature come noi» (Flanagan, 1991, p. 32). Tale principio non sottintende una posizione riduzionistica, che porterebbe ad assimilare l'elaborazione teorica dell'etica filosofica alle acquisizioni delle discipline scientifiche sulla natura della morale. Questo principio, piuttosto, si limita a formulare l'ineludibile esigenza di incardinare le teorie morali su una psicologia che non sia il mero prodotto dell'incontenibile immaginazione di filosofi e teologi, ma sia congruente con ciò che oggi la scienza ci dice su di noi.

Una visione d'insieme del presente fascicolo

Il presente fascicolo di *Teoria* si è proposto di prendere in esame le ricadute della scienza cognitiva dell'etica per una varietà di temi di metaetica, etica normativa, etica applicata e filosofia del diritto.

Un gruppo di articoli esamina criticamente le scoperte della neuroscienza cognitiva concernenti la responsabilità morale e legale. *What neuroscience will tell us about moral responsibility* di Daniel C. Dennett è un'elegante riflessione sul determinismo, la responsabilità morale e legale e il retributivismo alla luce dei dati della neuroscienza. L'autore riassume concisamente ma incisivamente la sua posizione secondo la quale la nozione compatibilista di libertà è tutto ciò di cui si ha bisogno per formulare giudizi di responsabilità morale e consente altresì di mantenere una linea moderatamente retributivistica.

In *Neurolaw and Punishment: A Naturalistic and Humanitarian View, and its Overlooked Perils*, Andrea Lavazza difende una posizione non dissimile da quella di Dennett. Lavazza analizza un progetto di naturalizzazione del diritto penale incardinato su dati neuroscientifici che alimentano lo scetticismo nei confronti della nostra capacità di intraprendere deliberatamente, volontariamente e consapevolmente un'azione scegliendo tra alternative. Questi dati sono perciò considerati una prova dell'illusorietà della responsabilità che apre il margine per forme più umane di punizione, giustificate su basi puramente utilitaristiche. L'autore prende posizione contro questa concezione puramente conseguenzialista della punizione, propone una valutazione più sfumata dei dati neuroscientifici in questione e difende una forma moderata di retributivismo.

In *Responsibility and Control in a Neuroethical Perspective* Elisabetta Sirgiovanni affronta la cosiddetta "ipotesi del controllo debole" nello stesso spirito antiradicale e riformista di Lavazza. L'ipotesi del controllo debole asserisce che gli agenti hanno molto meno controllo sulle proprie azioni di quanto sono normalmente portati a supporre, è questo a causa dell'influenza di fattori situazionali inaccessibili alla consapevolezza cosciente; e nella misura in cui la psicologia del senso comune definisce il controllo cosciente come condizione necessaria dell'agire responsabile, tale ipotesi rappresenta una minaccia per la nozione ordinaria di responsabilità. L'autrice prende in esame una serie di possibili soluzioni a tale minaccia e discute le obiezioni che ad esse sono state rivolte; quindi avanza alcune proposte per la costruzione di una teoria della responsabilità capace di unificare i punti di forza delle differenti soluzioni pur tenendo conto dei loro punti deboli.

Responsibility and the Relevance of Alternative Future Possibilities di Felipe De Brigard ci conduce alla fiorente industria delle ricerche empiriche sull'etica del senso comune, con particolare attenzione rivolta agli studi sulle credenze ordinarie in materia di libertà e responsabilità. Nella maggior parte dei casi questi studi basati su casi ipotetici si focalizzano esclusivamente sui giudizi dei partecipanti in merito all'eziologia degli eventi che hanno portato all'azione di un agente e sulle considerazioni relative a ciò che l'agente avrebbe potuto fare diversamente nel passato. Tuttavia, ci fa notare l'autore, prove recenti inducono a ipotizzare che quando si giudica se o meno un individuo è responsabile per una determinata azione (in scenari concreti, carichi emotivamente e pienamente deterministici) possono divenire rilevanti anche considerazioni circa possibilità future alternative. De Brigard esamina questi dati sperimentali, delinea un modo di interpretare la natura di questi effetti e indica alcune conseguenze per la filosofia sperimentale e la psicologia della libertà e della responsabilità.

Negli ultimi decenni, la biologia evolutiva, la sociologia e l'economia comportamentale hanno interagito produttivamente con le scienze psicologiche, rendendo sempre più chiaro che gli esseri umani sono *naturalmente* inclini alla competizione, e talora persino alla distruttività, ma anche alla socialità, alla cooperazione e all'altruismo. In *Social Justice, Individualism, and Cooperation*, Mario De Caro and Benedetta Giovanola esplorano il contributo che questa letteratura può offrire nel campo della filosofia politica. In particolare, gli autori sostengono che, al fine di rendere la riflessione sulla *giustizia sociale*, più affidabile ed efficace, i filosofi politici dovrebbero tenere conto del modello antropologico che emerge da quanto le scienze cognitive ci dicono sull'autoaffermazione, l'egoismo, la competizione, la pro-socialità, la cooperazione e l'altruismo.

In *Biology, Ethics and Moral Reflection*, Simone Pollo suggerisce uno specifico modo di collegare le scienze cognitive della morale alla normatività. L'interpretazione della psicologia morale offerta dalla scienza cognitiva non ci fornisce direttamente norme e valori ma, nella misura in cui è una forma di autocomprensione, contribuisce alla riflessione morale. Una parte della riflessione morale – che indaga e valuta i costumi, gli atteggiamenti e le risposte morali – consiste nella comprensione della nostra natura, sia in quanto individui sia in quanto membri della specie umana. Pollo si chiede poi se la scienza cognitiva della morale possa essere considerata una risposta moderna all'antica esortazione "conosci te stesso" e se, dunque, gli sviluppi di tale scienza possano contribuire al progresso morale.

Secondo la scienza (neuro)cognitive contemporanea, le credenze morali dipendono in modo cruciale dalle emozioni, e ciò soltanto perché essi influenzano attivamente i giudizi morali, ma anche perché paiono esserne condizioni necessarie e sufficienti. E ciò ha suggerito una visione originale della natura dell'etica, per la quale (a) i concetti morali sono legati in modo essenziale alle emozioni (*emotivismo epistemico*) e (b) le proprietà morali consistono di fatti emotivi (*emotivismo epistemico*). Questa concezione, accoppiata con la relatività culturale delle emozioni e dei sentimenti umani, genera un potente argomento in favore del relativismo etico. In *Emotions and Morality: Is Cognitive Science a Recipe for Ethical Relativism?* Massimo Reichlin sostiene che (1) l'emotivismo epistemico non dimostra che l'imputazione causale procede sempre dalle emozioni ai giudizi; non invalida l'idea che siano possibile giudizi spassionati; non offre una spiegazione persuasiva della distinzione tra regole morali e convenzionali; e non riesce a dar conto per la moralità autistica; (2) l'emotivismo metafisico offre un modello troppo deflazionistico del disaccordo morale – un fenomeno che può essere invece ben compreso se si assume un visione realistica rispetto ai fatti (inclusi i pro-atteggiamenti come le emozioni e i sentimenti) che offrono le migliori ragioni in favore di una determinata azione.

In *Lockean Persons, Self-Narratives, and Eudaimonia*, Rossella Guerini e Massimo Marraffa si interrogano sulla dimensione etica di una forma di costruttivismo narrativista che prende le distanze tanto dalle tendenze non naturalistiche che da quelle antirealiste nella riflessione sull'identità personale. La riflessione degli autori prende le mosse dalla teoria del *self* di William James. Su questo sfondo, l'idea che noi costituiamo noi stessi in quanto agenti moralmente responsabili (come "persone" nel senso di Locke) mediante narrazioni autobiografiche si congiunge con la tesi realista che l'identità narrative non è una ruota che gira a vuoto ma uno strato della personalità che funge da baricentro *causale* nella storia del soggetto (= un sistema psicobiologico). Questa alleanza tra una forma di costruttivismo narrativista e una posizione realista riguardo la realtà del *self* trova radicamento in una tradizione di pensiero che concepisce la sintesi fra i vari strati della personalità come il punto culminante del processo di autocostruzione del soggetto – ciò in accordo con un'etica incardinata sull'idea di *eudaimonia*, la scoperta e realizzazione delle proprie potenzialità e dei propri talenti.

"Bias implicito" è un termine tecnico che si riferisce alle caratteristiche relativamente inconsce e automatiche dei pregiudizi e dei comportamenti sociali. In *Category Matters: The Interlocking Epistemic and Moral Costs of Implicit Bias*, Lacey J. Davidson rifiuta la tesi che la responsabilità epistemica e morale presuppone che le categorie sociali non impattino sulle nostre valutazioni degli altri individui e delle loro azioni. Davidson nega la plausibilità empirica di questa tesi, sottolineando i benefici epistemici e morali che possono venire dalle categorie sociali. L'autrice evidenzia inoltre la peculiare interconnessione delle considerazioni epistemiche e morali con i *bias* impliciti. L'auspicio è che un'analisi di questo tipo possa contribuire a chiarire come le categorie sociali possano influire in modo corretto sui nostri giudizi, contribuendo a migliorare la nostra vita morale invece che a degradarla.

Negli ultimi anni i filosofi hanno iniziato a richiamarsi agli studi empirici per sostenere che quando noi pensiamo a una proposizione *p*, la crediamo automaticamente vera (questa concezione è spesso detta "teoria spinoziana", perché si pensa che Spinoza sia stato il primo a difenderla). Levy and Mandelbaum (2014) hanno spinto ancora oltre questa tesi, sostenendo che l'automaticità del credere abbia implicazioni per l'etica della credenza in quanto crea obblighi epistemici a quanti sono consapevoli di aver acquisito automaticamente una determinata credenza. Uwe Peters, in *On the Automaticity and Ethics of Belief*, fa uso di considerazioni teoriche e di risultati della psicologia per sollevare dubbi sulla sostenibilità empirica della teoria spinoziana.

In *The Ethical Convenience of Non-Neutrality in Medical Encounters: Argumentative Instruments for Healthcare Providers*, Maria Grazia Rossi, Sarah Bigi e Daniela Leone indagano le questioni etiche che riguardano la comunicazione considerando in particolare l'asimmetria della relazione dottore-paziente negli ambienti istituzionali. Importanti discussioni sono oggi dedicate alla questione, eticamente rilevante, della neutralità degli operatori sanitari. Le autrici di questo articolo argomentano che per salvaguardare la libertà e l'autonomia dei pazienti nella presa delle decisioni che li riguardano è possibile, e anzi desiderabile, adottare e gestire strategie non-neutrali di comunicazione. Per trattare il tema della neutralità nell'ambito della comunicazione, le autrici usano un modello normativo di argomentazione, sottolineandone l'efficacia come strumento di comunicazione per gli operatori sanitari.

In *"Publicity", Privacy and Social Media. The Role of Ethics above and beyond the Law*, Veronica Neri indaga il sempre più rilevante ruolo che i social media giocano all'incrocio tra etica e diritto, sia rispetto al tema della pubblicità (nel senso etimologico di "rendere pubblico") sia rispetto a quello della privacy. In questo ambito sono sempre più evidenti i limiti della legge e della deontologia: e per questo è necessario sviluppare una riflessione etica che si concentri sia sulle ragioni che motivano i comportamenti degli utenti – in primo luogo, rispetto alla spettacolarizzazione delle proprie vite – sia sui principi che potrebbero essere d'aiuto nel compiere scelte informate. Di particolare rilevanza, in questo senso, appare il concetto di "libertà responsabile", riferito alle possibili conseguenze che le nostre scelte potrebbero avere sia per noi stessi e sia per gli altri, e ciò tanto nell'ambiente dei social media quanto nella vita off-line.

Questi brevi riassunti non possono, naturalmente, rendere adeguatamente conto della ricchezza dei riferimenti sperimentali, dell'accuratezza delle argomentazioni filosofiche e della varietà di temi di grande profondità degli articoli qui raccolti che, nel loro complesso, mostrano quanto la scienza cognitiva contemporanea possa contribuire oggi allo sviluppo della teoria morale.

Mario De Caro, Massimo Marraffa

References

Bloom, P. (2013), *Just Babies: The Origins of Good and Evil*, Crown, New York.

Clausen, J. - Levy, N. (eds., 2015), *Springer Handbook for Neuroethics*, Springer, New York.

Doris, J. (ed., 2012), *The Moral Psychology Handbook*, Oxford UP, Oxford.

Doris, J. (2014), *Lack of Character: Personality and Moral Behavior*, Cambridge UP, Cambridge.

Doris, J. - Stich, S. (2014), *Moral psychology: empirical approaches*, in Zalta, E.N. (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2014 Edition), URL=<https://plato.stanford.edu/archives/fall2014/entries/moral-psych-emp/>.

Farah, M.J. (2010), *Neuroethics. An Introduction with Readings*, MIT Press, Cambridge (MA).

Flanagan, O. (1991), *Varieties of Moral Personalities. Ethics and Psychological Realism*, Harvard UP, Cambridge (MA).

Glannon, W. (2011), *Brain, Body, and Mind: Neuroethics with a Human Face*, Oxford UP, Oxford.

Greene J. (2013), *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*, Penguin, London.

Haidt, J. (2012), *The Righteous Mind*, Pantheon, New York.

Illes, J. - Sahakian, B.J. (eds., 2011), *Oxford Handbook of Neuroethics*, Oxford UP, Oxford.

Joyce, R. (2006), *The Evolution of Morality*, MIT Press, Cambridge (MA).

Lecaldano, E. (2012), *Etica*, in *Lessico del XXI secolo*. URL=<http://www.treccani.it/enciclopedia/etica_%28Lessico-del-XXI-Secolo%29/>.

Levy, N. - Mandelbaum, E. (2014), *The powers that bind: Doxastic voluntarism and epistemic obligation*, in Matheson, J. (ed.), *The Ethics of Belief*, Oxford UP, Oxford, pp. 12-33.

Marraffa, M. - Paternoster, A. (2017), *Models and mechanisms in cognitive sciences*, in Magnani, L. – Bertolotti, T. (eds.), *Springer Handbook of Model-Based Science*, Springer, Berlin.

May, J. (2017), *Moral psychology, empirical work in*, in *The Routledge Encyclopedia of Philosophy Online* (REP Online), URL=<https://www.rep.routledge.com/articles/thematic/moral-psychology-empirical-work-in/>.

Roskies, A. (2016), *Neuroethics*, in Zalta, E.N. (ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), URL=<http://plato.stanford.edu/archives/spr2016/entries/neuroethics/>.

Sinnott-Armstrong, W. (ed., 2008a), *Moral Psychology Vol. 1: The Evolution of Morality: Adaptations and Innateness*, MIT Press, Cambridge (MA).

Sinnott-Armstrong, W. (ed., 2008b), *Moral Psychology Vol. 2*: *The Cognitive Science of Morality: Intuition and Diversity*, MIT Press, Cambridge (MA).

Sinnott-Armstrong, W. (ed., 2008c), *Moral Psychology Vol. 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, MIT Press, Cambridge (MA).

Sinnott-Armstrong, W. (ed., 2014), *Moral Psychology Vol. 4: Free Will and Moral Responsibility*, MIT Press, Cambridge (MA).

What Neuroscience Will Tell Us About Moral Responsibility*

Daniel C. Dennett

There has been a lot of speculation recently about how advances in the neurosciences are going to oblige society in general, and lawmakers in particular, to reform or even overthrow our current understanding of the law, citizenship and, particularly, punishment. The theme that unites these speculations is the suggestion, sometimes explicitly endorsed, that science has shown that we human beings do not have free will after all, and hence are not morally responsible agents. Punishment is therefore unjustifiable, and should be replaced by a non-punitive system of treatment, with restraint only to the extent that it protects the public from dangerous individuals. More moderate proposals urge that we reform our policies to minimize punishment, restricting it to circumstances where we have well-grounded expectations of deterrent effect, to uphold respect for the law – since no miscreant is ever *really* morally responsible.

The demand for dramatic reforms in our inhumane systems of punishment (especially in the United States) is welcome, but the reasoning behind this particular informal campaign is dubious indeed. It depends on the assumption that the kind of "free will" that is prerequisite for moral responsibility is incompatible with determinism. Science has shown that all human actions, however deliberated, are the outcomes of causal chains extending back ultimately before our birth. Some thinkers deny this well-attested empirical claim, but with more hope than evidence. The hope is motivated by the belief that *if* our choices are thus caused, they cannot be "free" – and this would be a calamity.

* Brief statement for Rome Parliament discussion, January 26 2015.

It seems obvious to many that we must be capable of this kind of choosing for us to be morally competent agents, but this has never been demonstrated, and has been strenuously denied by *compatibilists*, who argue that such indeterminism is not at all a prerequisite for moral responsibility. The point of contention can be focused on the claim that when a person makes a morally responsible choice, it must be the case that she "could have done otherwise" – and this is never the case in a deterministic world. But this ignores an alternative, and much more plausible, interpretation of the key phrase, which we can bring out by looking at a usefully simple parallel in sports: who – if anybody – deserves to receive a red card in a football (soccer, to us Americans) match?

When a red card is issued, there is often heated discussion about whether it was *deserved*, and the distinction between deserved and undeserved penalties, while contentious in close calls, is obvious to all. One will seek far and wide for a football player or fan who thinks that the whole practice of issuing yellow cards and red cards and calling fouls should be abandoned, because it is too "punitive"; because it deals with human beings who could never really deserve anything—because of the truth of determinism. It is quite clear – so clear that even young children accept it with minimal explanation or justification – that strict rules don't just improve a game; they make it possible. If you want to play football, you have to play by the rules, and there are penalties – punishments – for violating the rules. This is fair. Life itself, as a whole, is not fair; some people are stronger, faster, more beautiful, richer, happier, more talented than others. Some are just luckier. But rules can be designed to "even the playing field" for all, and the measure of good rules is not that they never result in punishment, but that they strike a mutually acceptable balance between dangerous anarchy and over-enforcement. And one of the chief questions raised about any particular candidate for a foul is *could the player have done otherwise*? Players are held accountable for anticipating their trajectories and those of their opponents. They cannot plead "I could not have done otherwise because at the last moment I was already airborne on a collision course" if they should have foreseen this as the most likely outcome of a lunge. This is the sense of "could have done otherwise" that matters for fair rules and fair punishment, and it has nothing at all to do with whether or not determinism reigns in the physical world, or in the brains of individual people. (In fact, if causation were capricious on the football pitch so that players could not, in general, predict the outcomes of their actions, the "could have done otherwise" provision would have no application. Responsibility depends on predictability.) This is the sense of "could have done otherwise" that imposes an obligation on all participants (players of the game, or citizens of the state) to think ahead and give due consideration to likely outcomes. Nothing in neuroscience has shown that this capacity for responsible self-control is lacking in normal people.

There are those who are demonstrably not normal in this regard, and we already deem them as having diminished moral or legal responsibility, or none at all. They may have to be institutionalized against their will if they are dangerous, and they are not granted the right to sign contracts, or make legally binding promises. They, through no fault of their own, lack the requisite competence for being allowed to pass freely in the world. It is important to recognize that neuroscience does not in any way demonstrate that the difference between these unfortunate people and the rest of us is illusory.

What neuroscience has shown, and will continue to show in the coming years, is that some people whom we had thought to be normal in this regard are in fact subtly impaired in morally significant ways, and we will have to adjust our legal systems (through legislation or legal precedent) to take account of this new knowledge, but we can be confident in advance that this will be a self-limiting process – for a quite obvious political reason: people *want* to be held responsible because it is their ticket to *social* freedom, the right to act and move as they choose, making promises, and controlling their projects. We can concentrate the forces and considerations that are at play into a simple thought experiment.

Suppose you were to learn, from well-grounded neuroscientific examination, that you are at risk of developing an impairment of judgment or self-control that will destroy your moral competence. You now have two choices:

> submit to treatment that will (probably) protect you from this incapacitation, leaving you free to act in the world *at risk of being justly punished for any misdeed you commit*, or
>
> let nature take its course, in which case you can expect to commit some destructive act sooner or later that will lead to your institutionalization.

If the treatment is easy – taking a single pill, let's imagine – the choice is also easy. If the treatment is drastic, the choice is more difficult, and in many instances, it may well prove that neuroscience can offer a terrible diagnosis with no cure in sight. Life is not fair. We are already being faced by these decisions. It has been shown that young children who fail a simple test of self-control (the famous "marshmallow test") are much more likely to get in trouble with the law in adulthood than children who exhibit early self-control. Fortunately, there are non-invasive routines of education and practice that can repair this deficit, just as eyeglasses can restore normal vision. You wouldn't deny these routines to your own children, would you?

Who has the responsibility and the right to make such decisions? *These* are the questions we will have to address as neuroscience advances our ability to anticipate and explain deficits in human cognition and self-control, and notice that they presuppose that we – we fortunate ones – are morally responsible, and can be held accountable for our decisions.

Abstract

The essay is a reflection on determinism, moral and legal responsibility and punishment from the perspective of neuroscience. The author argues that compatibilist free will gives us everything we need to be morally responsible and allows us to maintain a moderately retributivist line of thinking.

Keywords: determinism; neuroscience; legal responsibility; moral responsibility.

Daniel C. Dennett
Tufts University
daniel.dennett@tufts.edu

Responsibility and the Relevance of Alternative Future Possibilities

Felipe De Brigard

A number of philosophers have claimed that if people are asked to consider the universe as being fully deterministic – that is, as a universe in which every event is necessarily entailed by a prior event in addition to the laws of nature – their intuitive reaction would be in line with incompatibilism about moral responsibility: i.e., they would be inclined to think that moral responsibility and determinism are incompatible[1]. However, in recent years, a number of results from experimental philosophy and psychology have cast doubt upon that claim. For example, in a series of seminal studies, Nahmias and collaborators presented participants with vignettes depicting deterministic scenarios[2]. When asked whether an agent in such scenario could have acted of her own free will, be responsible and/or deserve praise or blame for her actions, the majority of participants answered affirmatively. These results led Nahmias and colleagues to suggest that contrary to the received, a priori view among philosophers, people may actually be compatibilists.

Soon after, a number of studies challenged this conclusion. First, Nichols and Knobe reported results from a series of studies in which participants were presented with vignettes depicting fully deterministic scenarios[3]. How-

[1] R. Kane, *The Significance of Free Will*, Oxford University Press, Oxford-New York 1996; D. Pereboom, *Living without Free Will*, Cambridge University Press, Cambridge 2001.

[2] E.A. Nahmias-S.G. Morris-T. Nadelhoffer-J. Turner, *Surveying freedom: Folk intuitions about free will and moral responsibility*, in «Philosophical Psychology», 18 (2005), n. 5, pp. 561-584; Idd., *Is incompatibilism intuitive?*, in «Philosophy and Phenomenological Research», 73 (2006), n. 1, pp. 28-53.

[3] S. Nichols-J. Knobe, *Moral responsibility and determinism: The cognitive science of folk intuitions*, in «Nous», 41 (2007), n. 4, pp. 663-685.

ever, half of the participants received vignettes couched in abstract and emotionally neutral terms whereas the other half received vignettes couched in concrete and emotionally salient terms. They found that participants who read the deterministic scenarios described in concrete and emotionally salient terms were more likely to align their judgments of responsibility with compatibilism. In contrast, participants who read the deterministic scenarios described in abstract and emotionally neutral terms, made judgments that aligned with incompatibilism. Importantly, a more recent study suggests that this effect is evident across many cultures[4]. A second series of experiments conducted by Roskies and Nichols also challenged Nahmias et al.'s claim that people are naturally compatibilists[5]. In their study, Roskies and Nichols asked participants to read a vignette, similar to those employed in the previous studies, depicting a fully deterministic scenario. However, half of the participants were asked to imagine the described event occurring in a possible but non-actual world while the other half were told that the described event occurs in the actual world. Their results suggest that participants are more likely to give incompatibilist answers when they read vignettes depicting deterministic scenarios in a possible yet non-actual world whereas in scenarios described as occurring in the actual world their responses align with compatibilism. Finally, results from a study by Nahmias, Coates and Kvaran suggest that when presented with deterministic scenarios described in purely reductionistic terms, participants' judgments of responsibility align with compatibilism if the terms on the vignette are concrete and emotionally salient, but this is not the case if the vignettes are abstract and emotionally neutral[6].

A number of proposals have tried to accommodate these conflicting results. According to one proposal[7], the results of these studies could be accounted for if we assume a more basic psychological distinction between two distinct cognitive systems underlying our judgments of moral responsibility. On the one hand, there is a concrete system in charge of generating

[4] H. Sarkissian-A. Chatterjee-F. De Brigard-J. Knobe-S. Nichols-S. Sirker, *Is belief in free will a cultural universal?*, in «Mind & Language», 25 (2010), n. 3, pp. 346-358.

[5] A. Roskies-S. Nichols, *Bringing responsibility down to earth*, in «Journal of Philosophy», 105 (2008), n. 7, pp. 371-388.

[6] E. Nahmias-D.J. Coates-T. Kvaran, *Free will, moral responsibility, and mechanism: Experiments on folk intuitions*, in «Midwest Studies in Philosophy», 31 (2007), pp. 214-242. See also F. De Brigard-E. Mandelbaum-D. Ripley, *Responsibility and the brain sciences*, in «Ethical Theory and Moral Practice», 12 (2009), n. 5, pp. 511-524.

[7] S. Nichols-J. Knobe, *op. cit.*

judgments of moral responsibility when facing reductionistic, mechanistic, concrete and emotionally loaded deterministic scenarios. On the other hand, there is an abstract system in charge of producing judgments of responsibility for non-reductionistic, non-mechanistic, abstract and emotionally neutral deterministic scenarios. Indeed, Sinnott-Armstrong (2008)[8] has suggested that these two systems may be underwritten by the widely accepted distinction between episodic and semantic memory systems. More recently, a different proposal has been put forth by Murray and Nahmias[9]. According to their view, people are naturally compatibilists; their apparent incompatibilist judgments occur as a result of participants misinterpreting determinism as implying that the agent's mental states are bypassed in the causal chain leading up to the action. Thus, participants' incompatibilist intuitions can be explained away as an error in judgment. Needless to say, the debate as to whether peoples' judgments of responsibility align with compatibilism or incompatibilism in deterministic scenarios is far from being settled[10].

However, results from a recent study by De Brigard and Brady may pose an unexpected problem to the ecological validity of many of the studies reported in this debate[11]. In agreement with the way philosophers talk about the problem of free will, determinism and responsibility, experimental psychologists and philosophers have focused their efforts in exploring peoples' judgments of responsibility in scenarios where the only information that is provided pertains to the causal history preceding the agent's action. Specifically, researchers have been interested in determining which sorts of considerations about actual (or counterfactual) *past* events that bring about the agent's action influence peoples' judgments of responsibility in deterministic scenarios. But the fact that traditionally philosophers have only cared about the events that precede the agent's action does not mean that ordinary folk make the same assumption. There are a number of philosophical, moral and legal reasons to dismiss the import of

[8] W. Sinnott-Armstrong, *Abstrac+ Concrete = Paradox*, in J. Knobe-S. Nichols (eds.), *Experimental philosophy*, Oxford University Press, New York 2008, pp. 209-230.

[9] D. Murray-E. Nahmias, *Explaining away incompatibilist intuitions*, in «Philosophy and Phenomenological Research», 88 (2014), n. 2, pp. 434-467.

[10] S. Nichols, *Experimental philosophy and the problem of free will*, in «Science», 331 (2012), pp. 1401-1403; E. Nahmias, *Free Will and Moral Responsibility*, in «Wiley Interdisciplinary Reviews: Cognitive Science», 3 (2012), pp. 439-449.

[11] F. De Brigard-W. Brady, *The effect of what we think may happen on our judgments of responsibility*, in «Review of Philosophy and Psychology», 4 (2013), n. 2, pp. 259-269.

the consequences that may ensue if a person is held responsible at the present time. But there is no a priori reason to believe that ordinary people share those reasons and that they do not consider possible future events when judging if a person is or not responsible – even under fully deterministic and emotionally salient scenarios. Whether or not the folk's judgments about responsibility in fully deterministic scenarios are influenced by considerations about possible future events that may ensue as a result of holding an agent responsible is an open empirical question.

This issue is precisely what De Brigard and Brady set up to explore. In three between-group experiments they presented participants with mechanistic, reductionistic, emotionally loaded, and concrete deterministic scenarios of the sort that, consistently, have led participants to generate judgments of responsibility in line with compatibilism[12]. However, they manipulated whether possible future consequences that may ensue as a result of holding the agent responsible either improve or worsen the situation of an innocent third-party. Here, for instance, is the vignette read by the participants in the first experiment:

> Mary is the single mother of two: Mark, 7, Sally, 4. Mary works most of the day, and although she is known for being fairly patient and good natured, over the last year she has exhibited some unusually aggressive behavior toward her neighbor. Last week, when she came back from work late at night, she couldn't drive into her garage because her neighbor had blocked her driveway with his new BMW. Enraged, she stepped on the gas pedal and crashed her car into her neighbor's. Unfortunately, her neighbor was still inside the car (it was too dark for anyone to see him), and both his legs were seriously broken in several places. Now he is not only suing her for several thousand dollars, but he's also pressing charges. However, a neurologist examined her brain and discovered that, in the last year, Mary has been developing a rare tumor in her frontal lobe. Since the frontal lobe is necessary for emotional suppression – that is, the capacity to control one's emotions – the neurologist claims that, unlike a healthy person, Mary was completely unable to control her rage and her desire to smash the car. "In fact", he says, "any person with this kind of tumor", facing the exact same situation, would have done exactly what Mary did. She couldn't have done otherwise. "If Mary is found responsible for her actions, she may be sent to a federal medical facility for the next 6 months". There she could receive medical treatment, but she won't be able to see her children[13].

[12] *Ibidem.*
[13] *Ivi*, p. 262.

Half of the participants were randomly assigned to the *Better* condition, in which the vignette concluded with the following sentence:

> Fortunately, during that time, they would be living with Aunt Elizabeth, in what might be a much better environment for them.

The other half were assigned to the *Worse* condition, in which the vignette concluded with the following sentence:

> Unfortunately, during that time, they would be living with Social Services, in what might be a much worse environment for them.

Immediately after participants were asked to rate, on a 1-7 Lickert scale, whether or not they agreed or disagreed with the statement "Mary is morally responsible for crashing her car into her neighbor's". The results indicate that participants were significantly more likely to say that Mary was responsible in the *Better* condition (M = 5.30, SD = 1.2) than in the *Worse* condition (M = 3.15, SD = 1.7). These results suggest that, even under fully deterministic and emotionally-salient scenarios, when participants considered that the situation of an innocent third-party may worsen as a result of holding an agent responsible at the present time, their judgments are more aligned with compatibilism. However, when they considered that the condition of an innocent third-party may improve as a result of holding the agent responsible, their judgments were more in line with incompatibilism.

Since studies employing vignettes involving neural pathologies have produced conflicting results[14], De Brigard and Brady conducted two follow-up experiments in which the agent did not have a neural pathology[15]. In the first follow up, which was also a between subjects experiment, participants read a vignette similar to the one employed in the first experiment, except that this time the concrete and deterministic character of the description of the events leading up to the action was captured by assuming that Mary was wearing a brain monitoring system that recorded her brain activity. A neuroscientist then interpreted the data recorded from Mary's brain activity and concluded that the brain events leading up to Mary's action were completely determined and that she could not have done otherwise. As before, half of the participants were assigned to the *Better* condition and the other half were assigned to the *Worse* condition.

[14] F. De Brigard-E. Mandelbaum-D. Ripley, *op. cit.*

[15] F. De Brigard-W. Brady, *op. cit.*

The results of this second experiment revealed that participants were more likely to say that Mary was responsible for crashing her car into the neighbor's in the *Better* condition (M = 5.75, SD = 1.26) than in the *Worse* condition (M = 4.38, SD = 1.76) This suggests that participants were more prone to hold an agent responsibility if they considered that an innocent third-party may possibly be better off in the future as a result. Conversely, if the innocent third-party could end up worse off, participants' judgments of responsibility did not differ from the midpoint, suggesting that albeit not enough to exculpate the agent, considering this undesirable possible future consequences was sufficient to prevent participants from generating full-fledged compatibilist (or incompatibilist) judgments.

Finally, to explore whether or not the effect of considering possible consequences for innocent third-parties is a more pervasive characteristic of our judgments of responsibility, De Brigard and Brady conduced one final experiment in which the narrative about determinism was removed[16]. As before, half of the participants received a *Better* vignette, while the other half received a *Worse* vignette. Consistent with the results from their second experiment, the results of this final experiment revealed that participants were more likely to attribute responsibility to Mary if her children could be better off as a result of she going to a correctional facility (M = 6.17; SD = 1.24) than if they may be worse off (M = 4.46; SD = 2.21). Thus, taken together, the results of these three experiments strongly suggest that when assessing whether an agent is or not responsible for a particular action, people may consider possible future consequences for innocent third-parties that may be brought about as a result of holding the agent responsible at a present time. Moreover, this effect appears to be independent of whether or not the description of the conditions under which the agent acts is fully deterministic, mechanistic, reductionistic, and emotionally laden.

What may account for these results? The proposal I would like to put forth builds upon a recent and provocative paper by Phillips, Luguri and Knobe[17]. Their paper deals with the well-known phenomenon that moral judgments seem to influence non-moral assessments in a variety of domains. For instance, in a pioneer study, Knobe demonstrated that participants were more likely to say that an agent brought about a side effect he didn't care about when said side effect was morally wrong but not when it

[16] *Ibidem*.

[17] J. Phillips-J. Luguri-J. Knobe, *Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities*, in «Cognition», 145 (2015), pp. 30-42.

was morally right[18]. Relatedly, Phillips and Knobe conducted a study in which participants read a vignette depicting a scenario in which the captain of a ship saves its vessel from sinking by throwing his wife's cargo overboard (the morally neutral condition) or by throwing his wife overboard (the morally bad condition)[19]. Overall, participants were more likely to say that the captain was forced to throw something overboard in the morally neutral condition than in the morally bad condition. To explain these – and other related – results, Phillips and collaborators suggest, and offer evidence in favor of, the claim that moral considerations influence the kinds of possibilities people consider relevant when generating judgments about different notions across a number of distinct domains, such as intentional action, force, causation and doing/allowing[20]. More specifically, their suggestion is that «people show a general tendency to regard alternative possibilities as more relevant to the extent that they involve replacing morally bad things in the actual world with morally good alternatives»[21].

A similar explanation may be available for the effects uncovered by De Brigard and Brady[22]. Their results suggest that if a morally bad consequence could be brought about in the future as a result of holding an agent responsible at a present time, then participants are less likely to hold the agent responsible than if a morally good consequence were to be brought about. In agreement with Phillips and colleagues' proposal, one can hypothesize that this effect is due to a shift on attention toward relevant future possibilities that may be considered by the participants[23]. Thus, in the *Worse* condition, the morally bad effect on Mary's children renders certain possible future consequences more relevant, such as them having to live with someone they do not know, getting behind in school, or perhaps being mistreated in Social Services. Possible good consequences that may follow from this bad effect on Mary's children are not rendered relevant, thus they are not considered plausible. Because these bad consequences are ren-

[18] J. Knobe, *Intentional action and side effects in ordinary language*, in «Analysis», 63 (2003), n. 3, pp. 190-194.

[19] J. Phillips-J. Knobe, *Moral Judgments and Intuitions about Freedom*, in «Psychological Inquiry», 20 (2009), pp. 30-36. See also L. Young-J. Phillips, *The Paradox of Moral Focus*, in «Cognition», 119 (2011), pp. 166-178.

[20] J. Phillips-J. Luguri-J. Knobe, *op. cit.* See also D. Pettit-J. Knobe, *The Pervasive Impact of Moral Judgment*, in «Mind & Language», 24 (2009), n. 5, pp. 586-604.

[21] J. Phillips-J. Luguri-J. Knobe, *op. cit.*, p. 40.

[22] F. De Brigard-W. Brady, *op. cit.*

[23] J. Phillips-J. Luguri-J. Knobe, *op. cit.*

dered more plausible in the *Worse* conditions, participants may be motivated to prevent them from happening by way of judging the responsibility of the subject less harshly. Conversely, in the *Better* condition, the morally good effect on Mary's children renders other good consequences as being more relevant, like the fact that the nice aunt Elizabeth may provide a nurturing home for them, and would probably prevent them from getting in trouble or behind in school. Because good consequences are now rendered relevant – thus likely – people may be less inclined to mitigate Mary's responsibility – as there is less of an urge to prevent this outcome to occur.

Needless to say, this is merely a hypothesis. While it is inspired by Phillips and colleagues' recent proposal[24], it differs from theirs in an important respect. In their proposal, moral judgments influence the kinds of *counter-factual thoughts* participants entertain when assessing a certain situation. In the current interpretation of De Brigard and Brady's results[25], moral judgments influence *pre-factual thoughts* participants entertain when assessing Mary's moral responsibility. In other words, while their proposal states that moral judgments increase the relevance of certain thoughts about alternative ways past events could have occurred, the current proposal suggest that they can also render as relevant certain thoughts about how possible future events may unfold. Although it is so far an untested hypothesis, some extant evidence suggest that it may be promising, as it turns out that there is much in common between the neural and cognitive mechanisms underlying our capacity to entertain episodic future and counterfactual thoughts[26]. As such, the temporal dimension of the hypothetical simulation participants engage in during their judgments may not be critical[27]; what matters is the degree to which the moral character of the initially suggested possibility renders other possibilities as more or less relevant or plausible.

This need not be the whole explanation, of course. Extant evidence also suggests that our impulse to blame the perpetrator influences our attribu-

[24] *Ibidem.*

[25] F. De Brigard-W. Brady, *op. cit.*

[26] F. De Brigard-D. Addis-J.H. Ford-D.L. Schacter-K.S. Giovanello, *Remembering what could have happened: Neural correlates of episodic counterfactual thinking*, in «Neuropsychologia», 51 (2013), n. 12, pp. 2401-2414; D.L. Schacter-R. Benoit-F. De Brigard-K.K. Szpunar, *Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions*, in «Neurobiology of Learning and Memory», 117 (2015), pp. 14-21.

[27] F. De Brigard-B.S. Gessell, *Time is not of the essence: Understanding the neural correlates of mental time travel*, in S.B. Klein-K. Michaelian-K.K. Szpunar (eds.), *Seeing the Future: Theoretical Perspectives on Future-Oriented Mental Time Travel*, Oxford University Press, Oxford-New York 2016.

tions of free will and responsibility[28]. Notice, however, that this account does not conflict with the proposal put forth here, as each suggests a different process influencing our judgments of free will and responsibility. On the account put forth here, the main process is attention to relevant possibilities, whereas in the impulse-to-blame account the main process appears to be emotional. Clearly, further research is needed to fully understand the interaction between the impulse to blame and the relevance of alternative possibilities as factors influencing people's judgments of free will and responsibility.

Finally, in addition to offering a possible explanation of De Brigard and Brady's findings, it is worth mentioning at least two important methodological consequences that follow from them for both experimental philosophy and psychology of free-will and determinism[29]. First, both experimental philosophers and psychologists may want to take note of the relevance of possible future consequences when asking participants to assess the degree of responsibility of an agent in particular deterministic scenarios. The history of philosophy is full of prescriptive reasons as to why such consequences should not be taken into consideration when judging whether or not an agent is responsible for an action. However, such prescriptive considerations need not be entrenched in the psychological processes ordinary folk engage in when judging whether or not an agent is responsible. After all, our concept of responsibility presumably developed to play a social role – perhaps to curb people's behavior after a condemnable action, or to draw attention to the untrustworthiness of the agent, or who knows. But either way, it would be a mistake to assume that considerations about possible future events that we, philosophers or legal theorists, have learned to disregard on the basis of some prescriptive reason are also disregarded as a matter of course by ordinary people when judging the responsibility of an agent.

The second consequence follows, by way of generalization, from the first one: when designing vignettes to test people's intuitions about one or another notion – such as determinism, responsibility, free-will, and so forth – it is important not to mistakenly assume that our philosophical reasons for thinking that certain details are not relevant for the vignette are

[28] M.D. Alicke, *Culpable control and the psychology of blame*, in «Psychological Bulletin», 126 (2000), pp. 556-574; C.J. Clark-J.B. Luguri-P.H. Ditto-J. Knobe-A.F. Shariff-R.F. Baumeister, *Free to punish: A motivated account of free will belief*, in «Journal of Personality and Social Psychology», 106 (2014), pp. 501-513.

[29] F. De Brigard-W. Brady, *op. cit.*

also psychological reasons for thinking so. After all, one of the major downfalls of conducting research with these sorts of vignettes is that the researcher has only indirect control of the independent variable: she can manipulate what participants read, not what they think, and often what they think involves more than what they read. In experimental settings researchers work hard to keep background conditions as stable as possible in order to increase the probability that the intervention on the independent variable is predictive of the change in the dependent variable. The effects revealed by De Brigard and Brady suggest that something that was considered stable and irrelevant for the manipulation – i.e., considerations about possible future events – may actually have an effect on the dependent variable. As such, this finding constitutes an avenue for future research but also a possible worry about prior effects[30].

To conclude, let me summarize what I attempted to do in the current paper. I started off by briefly reviewing a number of recent results from experimental philosophy and psychology suggesting that, under certain conditions, people's intuitive compatibilist judgments shift toward incompatibilism even when considering fully deterministic scenarios. To account for these results, a couple of proposals have been put forth, including the suggestion that the kinds of cognitive processes involved in thinking about concrete, reductionistic, mechanistic and emotionally-laden deterministic scenarios are different from the kinds of cognitive processes involved in thinking about abstract, non-reductionistic, non-mechanistic and emotionally-neutral scenarios. However, recent findings from De Brigard and Brady put pressure on this proposal, as alternative future possibilities seem to affect participant's judgments of responsibility from compatibilist to incompatibilist even when they are presented with concrete, reductionistic, mechanistic, and emotionally-laden deterministic scenarios. As a result, building upon a recent proposal by Phillips and colleagues, a different account was put forth: that bringing attention to either morally bad or morally good outcomes renders certain related possibilities as more or less likely, thus as more or less relevant for considering whether or not the agent is responsible. Finally, I drew a couple of methodological sugges-

[30] *Ibidem.* It is worth noting that others have expressed skepticism as to whether responses to moral dilemmas in experimental settings actually reflect responses to similar situations in real-life settings (cf. G. Kahane-J.A.C. Everett-B.D. Earp-M. Farias-J. Savulescu, *"Utilitarian" judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good*, in «Cognition», 124 (2015), pp. 193-209).

tions from these results, in the hopes that bringing attention to potential confounds in extant experimental designs can help to motivate more ecologically valid studies moving forward[31].

Abstract

In the past decade, philosophical and psychological research on people's beliefs about free will and responsibility has skyrocketed. For the most part, these vignette-based studies have exclusively focused on participants' judgments of the causal history of the events leading up to an agent's action and considerations about what the agent could have done differently in the past. However, recent evidence suggests that, when judging whether or not an individual is responsible for a certain action – even in concrete, emotionally laden and fully deterministic scenarios – considerations about alternative future possibilities may become relevant. This paper reviews this evidence and suggests a way of interpreting the nature of these effects as well as some consequences for experimental philosophy and psychology of free will and responsibility going forward.

Keywords: experimental philosophy; experimental psychology; free will; responsibility.

Felipe De Brigard
Department of Philosophy
Duke University, Durham, North Carolina
felipe.debrigard@duke.edu

[31] Thanks to Paul Henne and two anonymous reviewers for helpful comments.

Ethics, Law, and Cognitive Science

T

Category Matters: The Interlocking Epistemic and Moral Costs of Implicit Bias

Lacey J. Davidson

1. *Introduction*

On October 26, 2016 the Young Conservatives of Texas (YCT) at The University of Texas at Austin held an anti-affirmative action[1] bake sale with prices based on race and sex, charging more to individuals with social identities that YCT argued benefit the most from affirmative action. The Chairman of the organization, Vidal Castañeda, stated,

> Our protest was designed to highlight the insanity of assigning our lives value based on our race and ethnicity, rather than our talents, work ethic, and intelligence [...]. It is insane that institutional racism, such as affirmative action, continues to allow for universities to judge me by the color of my skin rather than my actions[2].

Put another way, the justification offered for the bake sale is that social category membership should be irrelevant to judgments and decisions about individuals.

This is a particular kind of argument, made for a particular political purpose, in hopes of a particular outcome. Notably, philosophers writing on implicit bias, philosophers with far different goals than the YCT, also utilize a version of this same claim: social categories *should* be irrelevant to our assessments of individuals. I will call this claim The Irrelevance

[1] It is not the purpose of this paper to take a stance on affirmative action. My point will be to show that the same problematic premise can be used as the foundation for many different types of arguments.

[2] For more information see <www.cnn.com/2016/10/28/us/university-bake-sale-trnd/>.

Assumption[3]. Philosophers then add the following premises to make arguments about pernicious epistemic[4] mistakes:

P1 (The Irrelevance Assumption): Social categories should be irrelevant to our assessments of individuals.
P2: When a person harbors implicit biases and those implicit biases influence an assessment, then a social category has influenced that assessment.
P3: Including irrelevant information in assessments is an epistemic mistake.
Conclusion: Implicit bias leads to epistemic mistakes.

In this paper I argue that The Irrelevance Assumption is mistaken, or that there are at least some cases in which social categories are relevant to our assessments and should be taken into account. Following Charles Mills, who, in detailing the metaphysics of racial categories, claims «that race *should* be irrelevant is certainly an attractive ideal, but when it has *not* been irrelevant, it is absurd to proceed as if it had been»[5], I will argue that rather than aspiring to ignore, disregard, or overcome these influences, we must take social categories into account if we are to be in the best moral and epistemic positions. Though my aim here is primarily to give a negative account urging for the rejection of The Irrelevance Assumption, it is my hope that this paper clears the ground for a positive view about the right ways social categories might influence our judgments about a person or group of people. In other words, I want to show it is not *that* a social category influences our judgment that is important, but rather *how*[6].

[3] The claim is formulated and used differently by various theorists, including the two philosophers I will focus on, Saul and Gendler. These differences do not, however, change the content I wish to target.

[4] To shift the focus of the argument to the moral costs of implicit bias, one need only to replace "epistemic" with "moral" in the argument. Both versions are common in the literature. My purpose here is to demonstrate a type of argument in which The Irrelevance Assumption may be used, rather than a particular view.

[5] C. Mills, *But what are you really? The metaphysics of race*, in A. Light-N. Mechthild (eds.), *Race, Class, and Community Identity: Radical Philosophy Today*, Humanity Books, Amherst (NY) 2000, pp. 23-51, p. 41. Though there are certainly differences between types of social categories, particularly with respect to historical legacy, many of his arguments can be applied to other categories such as gender or sexual orientation.

[6] Alex Madva puts forth a similar view and explores the possibility of «*regulating* the accessibility of our social knowledge in order to have that information available when and only when we need it» (*Virtue, social knowledge, and implicit bias*, in M. Brownstein-J. Saul (eds.), *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology*, Oxford UP, Oxford 2016, pp. 191-215, p. 201.

To analyze this assumption, I will focus on two papers: *On the Epistemic Costs of Implicit Bias*, written by Tamar Szabo Gendler[7], and *Scepticism and Implicit Bias*, written by Jennifer Saul[8]. The motivation for these selections is twofold: one, both constitute novel and major contributions to the philosophical work on implicit bias, particularly regarding the special connection between the epistemic and the ethical in the identification and mitigation of implicit bias; two, the assumption that social categories should not influence our judgments of individuals is important for each argument, yet left undefended. In section 3, I clarify the work this assumption does for each argument in turn. I want to note that rejecting The Irrelevance Assumption does not entail rejecting the conclusion of either paper (though the conclusions will require different arguments). Rejection of The Irrelevance Assumption forces a reconceptualization of the guiding questions for research on the potential mitigation of implicit bias. It is not my hope to give a full account of moral responsibility for implicit bias[9]. Rather I will challenge a claim that has not been defended, yet plays a key role in important, agenda-setting arguments, but which itself is, I argue, questionable.

2. *Implicit Bias*

In this section I will give a brief descriptive account of the contemporary literature on implicit bias. The term implicit bias refers to the unre-

[7] T.S. Gendler, *On the epistemic costs of implicit bias*, in «Philosophical Studies», 156 (2011), n. 1, pp. 33-63.

[8] J. Saul, *Scepticism and implicit bias*, in «Disputatio», 5 (2012), n. 37, pp. 243-263.

[9] Those interested in these accounts should see: S. Brennan, *The moral status of micro-inequities: In favor of institutional solutions*, in M. Brownstein-J. Saul (eds.), *Implicit Bias and Philosophy, Vol. 2: Moral Responsibility, Structural Injustice, and Ethics*, Oxford UP, Oxford 2016, pp. 191-214; M. Brownstein, *The implicit mind*, manuscript; A. Rees, *A virtue ethics response to implicit bias*, in M. Brownstein-J. Saul (eds.), *op. cit.*, pp. 191-214; M. Fricker, *Fault and no-fault responsibility for implicit prejudice. A space for epistemic "agent-regret"*, in M. Fricker-M. Brady (eds.), *The Epistemic Life of Groups: Essays in the Epistemology of Collective*, Oxford UP, Oxford 2015, pp. 33-50; J. Glasgow, *Alienation and responsibility*, in M. Brownstein-J. Saul (eds.), *op. cit.*, pp. 37-61; J. Holroyd, *Responsibility for implicit bias*, in «Journal of Social Philosophy», 43 (2012), n. 3, pp. 274-306; N. Levy, *Consciousness, implicit attitudes, and moral responsibility*, in «Noûs», 48 (2012), pp. 21-40; N. Washington-D. Kelly, *Who's responsible for this? Moral responsibility, externalism, and knowledge about implicit bias*, in M. Brownstein-J. Saul, *op. cit.*, pp. 11-36; and R. Zheng, *Attributability, accountability, and implicit bias*, in M. Brownstein-J. Saul, *op. cit.*, pp. 62-89.

flective and hard to introspectively access set of automatic associations that may lead to prejudiced judgment and behavior[10]. Interest in implicit bias was triggered by psychological findings about the nature of implicit association. The development of the Implicit Association Test (IAT)[11], an indirect measure of implicit associations between two target concepts, and subsequent findings pushed psychologists and philosophers to think about possible tensions between reports of belief and unconscious associations.

Over 75% of Americans who have taken the Race IAT show an automatic preference for White faces over Black faces[12]. In addition, studies that collect both implicit preferences (through the IAT) and explicit preferences (through self-report measures) show that White participants have greater implicit preferences for White over Black (d = .83) than explicit preferences for White over Black (d = .59)[13]. This demonstrates both the strength of the preference, as well as the discordance between reports of belief and implicit associations. Since the development of the IAT, several additional indirect association measures have been developed. Some examples are the Affect Misattribution Procedure (AMP)[14], the Go/No-go Association Task[15], and the Extrinsic Affective Simon Task[16].

Further, there is evidence that implicit bias is correlated with prejudiced behavior. Since 2007, many IAT studies have also included a behavioral

[10] Though an interesting and worthwhile pursuit, the exact mental nature of implicit bias will not be explored in this paper. The arguments made in this paper will be relevant whether implicit biases are: beliefs (E. Mandelbaum, *Attitude, inference, association: On the propositional structure of implicit bias*, in «Nous», 50 (2016), n. 3, pp. 629-658; aliefs (T.S. Gendler, *Alief and belief*, in «Journal of Philosophy», 105 (2008), n. 10, pp. 634-663; *FTBA* attitudes (M. Brownstein, *op. cit.*); or character traits (E. Machery, *De-Freuding Implicit Attitudes*, in M. Brownstein-J. Saul (eds.), *Implicit Bias and Philosophy: Vol. 1, Metaphysics and Epistemology*, Oxford UP, Oxford 2016, pp. 104-129).

[11] A. Greenwald-D. McGhee-J. Schwartz, *Measuring individual differences in implicit cognition: The implicit association test*, in «Journal of Personality and Social Psychology», 74 (1998), n. 6, pp. 1464-1480.

[12] M.R. Banaji-A.G. Greenwald, *Blindspot: Hidden Biases of Good People*, Delacorte Press, New York 2013.

[13] B.A. Nosek-M.R. Banaji-A.G. Greenwald, *Harvesting intergroup implicit attitudes and beliefs from a demonstration website*, in «Group Dynamics», 6 (2002), pp. 101-115.

[14] B.K. Payne-C.M. Cheng-O. Govorun-B.D. Stewart, *An inkblot for attitudes: Affect misattribution as implicit measurement*, in «Journal of Personality and Social Psychology», 89 (2005), n. 3, pp. 277-293.

[15] B.A. Nosek-M.R. Banaji, *The go/no-go association task*, in «Social Cognition», 19 (2001), n. 6, pp. 161-176.

[16] J. De Houwer, *The extrinsic affective simon task*, in «Experimental Psychology», 50 (2003), n. 2, pp. 77-85.

measure to test the IAT's predictive validity[17]. A 2009 meta-analysis showed that the IAT effectively predicts a range of prejudiced behavior (*r* = .274) and does so better than self-report measures for socially sensitive issues such as race[18]. These studies addressed challenges claiming that the IAT may not be correlated with behavior or not be correlated more strongly than explicit report measures. Taken together these studies suggest something troubling: we likely have implicit associations that affect our judgments and behavior in ways that we would explicitly disavow. In other words, individuals may have strong commitments to racial equity[19], but think and act in ways that notably work against this goal.

3. *Gendler and Saul*

In this section I will sketch the arguments given in the papers by Gendler and Saul, highlighting their uses of The Irrelevance Assumption: that social categories should not influence our judgments of individuals. For both Gendler and Saul, epistemic and moral costs are closely tied; however, the relationship between the costs is different for each. For Saul, as epistemic costs are mitigated, so are moral costs; the costs are directly related. For Gendler, on the other hand, as moral costs are mitigated, epistemic costs are incurred; the costs are inversely related[20]. This differing relationship influences the way each philosopher utilizes The Irrelevance Assumption. Though Gendler's *On the Epistemic Costs of Implicit Bias* came before, I will discuss Saul's *Scepticism and Implicit Bias* first as she explicitly states The Irrelevance Assumption.

In this insightful and influential article, Saul claims that contemporary research about implicit biases and their effects gives rise to a new kind of skepticism, a "bias-related doubt". Though this doubt doesn't lead us to question the existence of the external world or other minds like traditional forms of skepticism, we do have «very good reason to believe that we cannot properly trust our knowledge-seeking faculties», particularly when it

[17] M.R. Banaji-A.G. Greenwald, *op. cit.*

[18] A. Greenwald-D. McGhee-J. Schwartz, *op. cit.*

[19] Other formulations of this commitment, such as to egalitarianism or treating people equally, will also pose a similar problem. The inconsistency between the avowal and the implicit association is the interesting phenomenon, not the particular nature or wording of the avowal/commitment.

[20] Particularly with respect to the encoding and use of relevant base-rates.

comes to our knowledge about other people, their capacities, and intentions[21]. She claims that this type of skepticism is in some sense stronger than traditional skepticism because it "demands action" (243). Doubting the existence of our hands doesn't prompt us to radically change our epistemic or moral situations; doubting the accuracy of our everyday credibility and like judgments does. She argues for this conclusion by giving a series of troubling empirical cases in which implicit bias plays a role – CV evaluation, journal submission evaluation, and shooter bias – and highlights the moral, political, and epistemic consequences of the impact of implicit bias. She then gives a variety of possible solutions for improving our epistemic and moral situations.

I agree with Saul. I think there is something going wrong morally and epistemically when our negative and inaccurate implicit biases affect our actions and judgments. However, I find one assumption she makes throughout the paper troubling. Here I'll give several instances of the assumption:

> These judgments are very clearly being affected by something that *should* be irrelevant – the social category of the person [...][22].
>
> [...] they shouldn't be looking at the credibility of an individual at all. They should be looking just at the study, or argument. And yet when implicit bias is at work, we are likely to be affected by the social group of the person presenting evidence or an argument even when we were [sic] are trying to evaluate that evidence or argument itself[23].
>
> These mistakes are ones in which something (the social category of the individual) that we actively think *should not* affect us does[24].
>
> if you actually are basing lots of decisions on the social categories that people you encounter belong to, then you're clearly not doing as well as you can[25].

To ensure I am not being uncharitable, I want to clarify what I think she might mean: that we are making mistakes when our *inaccurate* stereotypes about social categories alter our judgments and actions. She points to this when she details some potential mistakes:

> You're making the wrong decisions epistemically speaking: taking an argument to be better than it is, perhaps; or wrongly discounting the view of someone you

[21] J. Saul, *op. cit.*, p. 243.

[22] *Ivi*, p. 244.

[23] *Ivi*, p. 249.

[24] *Ibidem*.

[25] *Ivi*, p. 256.

should be listening to. You're also making the wrong decisions practically speaking: assigning the wrong mark to an essay, or rejecting a paper that you should accept. Finally, you're making wrong decisions morally speaking: you are treating people unfairly; and you are basing your decisions on stereotypes that you find morally repugnant[26].

In these cases, unconscious social category biases influence judgments in a way that makes one less accurate and less likely to acquire desired knowledge. However, I want to emphasize that these mistakes are not *merely* a result of social categories influencing judgment or action, but rather of erroneous and pernicious social category associations altering judgments or actions. It is not *that* a social category affects judgment, but *how*. Though this final quotation leads us in this direction, the above four selections do not. In those, the assumption is more baldly stated, i.e., that no matter how the social category influences our judgment and decisions, we've made a mistake. It is this assumption, The Irrelevance Assumption, I want to reject.

I will now turn to Gendler's discussion of epistemic costs. Though similar in topic, the argument put forth leads to a vastly different conclusion. Rather than asserting that we must do something to avoid the types of epistemic and moral mistakes that arise from implicit bias (and that this may indeed be possible with the right, but not-yet empirically discovered, kinds of strategies), Gendler concludes, «living in a society structured by race appears to make it impossible to be both rational and equitable»[27]. That is, if we mitigate the moral costs of implicit bias, we increase the epistemic costs, and vice versa.

Throughout the paper, Gendler highlights epistemic costs associated with implicit bias, three that result from the phenomenon itself, as well as living in a racialized environment in general (this discussion is where she is most closely aligned with Saul), and one from attempts at mitigating the effects of implicit bias. She emphasizes throughout the discussion that these costs are incurred regardless of whether the individual avows the content of her automatic associations.

First, she identifies the cross-race recognition deficit in which individuals are more likely to remember specific facial features of own-race individuals than other-race individuals. Rather than encoding information that

[26] *Ibidem.*

[27] T.S. Gendler, *op.* cit., p. 57.

would allow future recognition, participants encode the face as "racial category" for the purposes of classification. She asserts that the tendency to encode this way is a result of automatic associations. Second, Gendler describes stereotype threat: «a well-documented phenomenon whereby activating an individual's thoughts about her membership in a group that is associated with impaired performance in a particular domain increases her tendency to perform in a stereotype-confirming manner»[28]. Negative implicit biases turned inward lead to epistemic costs for particular tasks[29]. Third, she gives an account of cognitive depletion after interracial interaction; after white participants interacted with a different-race peer, they performed more poorly on executive control tasks. Cognitive depletion is particularly high on this task for those who have avowals that are discordant with their automatic racial preferences (implicit biases), indicating that the depletion may be caused by attempts to suppress automatic behaviors stemming from these biases. In describing these costs, Gendler's account is similar to Saul's: implicit biases lead to epistemic costs and further behavior that may not be in-line with avowed anti-discriminatory commitments.

Gendler then turns to epistemic costs associated with mitigating implicit bias; this is where her argument parts ways with Saul's. However, she still makes use of the claim I wish to reject, though she gives it a different role. In this discussion Gendler cites research on what Philip Tetlock et al. call "forbidden base rates" to claim that mitigating implicit bias often requires one to ignore important social category information that may *improve* our epistemic situations, rather than degrade them[30]. For example, citing Tetlock, she gives cases in which individuals did not take race-correlated actuarial risk into account when assigning insurance premiums and «engaged in a kind of epistemic self-censorship on non-epistemic grounds»[31]. She categorizes this censorship behavior as irrational, even though it aligns with anti-racist avowals. It is here that epistemic and moral concerns are in tension with one another.

[28] *Ivi*, p. 48.

[29] When the biases are positive, performance may improve. This phenomenon is referred to as Stereotype Lift. For further research, see L. Froehlich-S.E. Martiny-K. Deaux-T. Goetz-S.Y. Mok, *Being smart or getting smarter: Implicit theory of intelligence moderates stereotype threat and stereotype lift effects*, in «British Journal of Social Psychology», 55 (2016), n. 3, pp. 564-587.

[30] P.F. Tetlock-O. Kristel-B. Elson-M. Green-J. Lerner, *The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals*, in «Journal of Personality and Social Psychology», 78 (2000), n. 5, pp. 853-870.

[31] *Ivi*, p. 55.

The Irrelevance Assumption shows up in her discussion of bias mitigation or what one *might do* to avoid epistemic and moral costs[32]. She asserts that to reduce epistemic costs we might «fail to encode the base rate information and cultural associations that give rise to these problematic aleifs [Gendler's term for implicit attitudes]»[33]. Because I think it is empirically unlikely that we can «fail to encode» associations[34], I'll rephrase: one might keep the social category of an individual from affecting judgments and subsequent decisions in order to improve moral outcomes. You might think that this rephrasing is too self-serving; however, Gendler also suggests that we might ignore social category base-rates for ethical reasons (i.e. upholding anti-racist commitments), which seems like a clear case of The Irrelevance Assumption and related argument-type given in Section 1. Because she thinks that there are times in which relying on social categories improves our epistemic situation, Gendler's use of the assumption is slightly different than Saul's; nevertheless, both use The Irrelevance Assumption as the ideal for those wishing to improve epistemic and moral situations with respect to implicit bias.

4. *Rejection of The Irrelevance Assumption*

In this section of the paper I show that The Irrelevance Assumption is false, i.e. that there are circumstances when, in considering another person, information about the social categories that person belongs to is not only relevant to, but also *should be used* in the assessment of that person. My treatment of each challenge will be short, and it is my hope that these critiques provide the ground for a continued discussion. First, I will challenge the premise on empirical grounds, asserting that it is likely not possible to make evaluations independently of social category information. Second, I will demonstrate two epistemic benefits that arise from taking social categories into account. And, third, I will discuss base-rate neglect, emphasizing that the inclusion of negative base-rates (such as crime statistics) is not the only plausible way to include social category information.

[32] Thus, it may be unfair to attribute the assumption *to* Gendler. However, it seems like Gendler would've given another option for bias mitigation, if she thought one was available.

[33] T.S Gendler, *op. cit.*, p. 54.

[34] This will be a part of my rejection of The Irrelevance Assumption. See section 4.1 for details.

An important upshot of this reframing will be that Gendler's conclusion is mistaken; it will be possible to be both rational and equitable[35].

4.1. *Empirical Possibility*

Taking into account the psychological literature on implicit encoding and associative attitudes, we may wonder whether it is possible to keep social categories from influencing evaluations of individuals. I will take a familiar stance on responsibility here: it seems odd to require something that is not possible, even if – were it possible – it would be ideal. The empirical challenge can be mounted from two fronts: the automatic encoding of stereotype information and the automatic tendency to group individuals into social categories and apply category relevant information. More than half a century ago Gordon Allport described categorization as a basic feature of effective cognitive functioning[36]. Perhaps most telling is the early age at which individuals begin to use social categories to understand the world. Children as young as three or four already use gender and race in reasoning tasks[37]. Particularly enlightening are discussions of human kinds and their development over time and space[38], which suggest that most of our social interactions are structured by social category thinking. Further, much of the research on implicit bias itself supports the automatic encoding and use of social categories and that these categories update based on new information and experiences[39]. Further empirical and theoretical research on social category cognition and its mechanisms may allow us to shift the ways in which

[35] Joshua Mugg has taken on this assertion as well; for details cf. J. Mugg, *What are the costs of racism? A reply to Gendler*, in «Philosophical Studies», 166 (2013), pp. 217-229; Id., *How to deal with the tragic dilemma: An argument against the incommensurability thesis*, manuscript.

[36] G.W. Allport, *The Nature of Prejudice*, Addison-Wesley, Cambridge (MA) 1954. For a similar argument with a direct application to implicit bias, see L. Antony, *Bias: Friend or foe? Reflections on saulish skepticism*, in M. Brownstein-J. Saul (eds.), *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology*, Oxford UP, Oxford 2016 pp. 157-190.

[37] K. Shutts-C.K. Pemberton Roben-E.S. Spelke, *Children's use of social categories in thinking about people and social relationships*, in «Journal of Cognition and Development», 14 (2013), n. 1, pp. 35-62.

[38] I. Hacking, *The looping effects of human kinds*, in D. Sperber-D. Premack-A.J. Premack (eds.), *Causal Cognition: A Multidisciplinary Debate*, Oxford UP, Oxford 1995, pp. 351-383; R. Mallon, *The Construction of Human Kinds*, Oxford UP, Oxford 2016; C. Mills, *Alternative epistemologies*, in C. Mills, *Blackness Visible: Essays on Philosophy and Race*, Cornell UP, Ithaca (NY) 2000. The references I list here are philosophers utilizing large swaths of empirical literature to make philosophic arguments, rather than original empirical research.

[39] M. Brownstein, *op. cit.*

social category information is encoded and made salient for use in unconscious processing, explicit reasoning, and judgment. My point here is not that we shouldn't worry about these processes because they are automatic and unavoidable, but rather that we ought not make the mistake of assuming it is possible to ignore social category information[40].

4.2. *Epistemic Benefits*

In this section, I will discuss two epistemic benefits that arise from including social category information, at least when it's done right: increased testimonial credibility[41] and robust social exchange. First, the social category of a speaker should increase a hearer's credibility judgment of a speaker when the social category is relevant to the content of the testimony. One of the most common critiques of reproductive policy makers in America is that they are making arguments and decisions about something they will never experience; male representatives are making laws that determine the decisions women can make about their bodies and family planning[42]. Similarly, a common critique of policy makers and political leaders from activist groups like Black Lives Matter is that leaders make claims and decisions about black lives, even when they fail to understand the experience of black women and men[43]. In light of these critiques, one clear way social categories can play an important and valuable role in testimony evaluation is in relationship to the *content* of the testimony. Intuitively, it makes more sense to avow the testimony of someone giving evidence about common experiences of members of their *own* social category than the testimony of someone speaking about experiences had by those in other social categories[44].

[40] Thanks to an anonymous reviewer for pushing me to clarify this point.

[41] This example is prompted by Miranda Fricker's account of Epistemic Injustice (*Epistemic injustice*, Oxford UP, Oxford 2007), i.e., harms done to individuals in their capacity as knowers. Though not discussed by Fricker, it is reasonable to assert that implicit biases of the type described above can be responsible for, or produce, the kind of social identity prejudice necessary to set credibility lower than it should (based on relevant factors such as expertise, experience, etc.) be set. This leads to a testimonial epistemic injustice. For more detail, see Fricker, *op. cit.*

[42] Though I don't have space to detail the history of this critique, here is an example of a recent protest from my home state, Indiana: <http://www.nytimes.com/2016/04/08/us/periods-for-pence-campaign-targets-indiana-governor-over-abortion-law.html>.

[43] Similarly, I cannot give a full account. Here is an example: <http://www.theroot.com/articles/culture/2014/08/ferguson_how_white_people_can_be_allies/>.

[44] I am not claiming that individuals are able to speak on behalf of *all* other members of a particular social category nor that they should be asked to do so.

Further, we might expect that members of oppressed social categories have privileged insight and a more developed critical lens through which to see our social and political sphere, particularly with respect to social inequities they themselves experience[45].

Second, the influence of social category information on decisions such as hiring and academic admittance ensures a robust intellectual and work community that encourages vibrant discussion and multiple perspectives[46]. The ability of diverse groups to come to superior decisions because of deliberation[47] means that, to improve the epistemic situations of groups, one ought to pay attention to the social categories of individuals that will learn or work together. Another benefit of this strategy is that it mitigates at least some worries about CV[48] and like evaluation so often cited in the philosophical literature, thereby lowering epistemic and moral costs associated with implicit bias. Admittedly, both examples are of explicit reasoning and decision-making about the inclusion of social category information, rather than of the unconscious influences of pernicious implicit biases about which Saul and Gender are worried. My point here is to give evidence against and reject The Irrelevance Assumption, which is agnostic about the reasoning process used to assess individuals.

4.3. *Base-Rate Neglect*

A further discussion of Gendler's base-rate neglect is in order. It may seem that above I simply agree with Gendler's conclusion: that social category information, such as base-rates, should be included in our judgments of others upon penalty of irrationality. A tacit assumption in these discussions of base-rates is that social category information only provides negative information about individuals and improves our epistemic situations

[45] P.H. Collins, *Black Feminist Thought*, Unwin Hyman, Boston (MA) 1990; S. Harding, *Whose science? Whose knowledge?*, Cornell UP, Ithaca (NY) 1991; Mills, *op. cit.*; D. Smith, *Women's perspective as a radical critique of sociology*, in «Sociological Inquiry», 44 (1974), pp. 7-13.

[46] For a more robust account of the epistemic value of diversity, see E. Robertson, *The epistemic value of diversity*, in «Journal of Philosophy of Education», 47 (2013), n. 2, pp. 299-310.

[47] H. Landemore, *Democratic Reason: Politics, Collective Intelligence, and the Rule of the Many*, Princeton UP, Princeton (NJ) 2012. For a critique of this position, cf. S. Stich, *When democracy meets pluralism: Landemore's epistemic argument for democracy and the problem of value diversity*, in «Critical Review», 26 (2014), nn. 1-2, pp. 170-183.

[48] C. Moss-Racusin-J. Dovidio-V. Brescoll-M. Graham-J. Hadnelsman, *Science faculty's subtle gender biases favor moral students*, in «Proc Natl Acad Sci USA», 109 (2012), n. 41, pp. 1647-1649.

by telling us whom to avoid or whom not to trust. In the previous two sections, I've given some evidence that challenges this claim; paying attention to social categories improves our epistemic situations by signaling expertise, ensuring a variety of perspectives, and limiting epistemic mistakes.

Further, I want to comment on the purported irrationality of ignoring accurate base-rate information for practical purposes (in this case anti-racist purposes), beginning with an example that Gendler uses to describe alief. To emphasize the tension between automatic "representational-affective-behavioral" aliefs and avowed commitments (Gendler calls them "endorsed beliefs") she details the fear one may feel on a skywalk above the Grand Canyon[49]. Although one desires to view the Grand Canyon in the suspended position and avows that the skywalk is safe, one may experience intense fear upon ascending into the clear case. If one is able to brave the Skywalk regardless, it is likely because one's avowals and desires (to view the Canyon) overcome the automatic reaction[50]. I think this is similar to base-rate neglect. It is not clear that the Skywalk override is rational on Gendler's view. There *is* a chance, however slight, that the bridge will break.

If you don't think the Skywalk example demonstrates this, take riders of rollercoasters or carnival rides. Most have seen media stories of twelve passengers hanging upside down in a broken down rollercoaster or, if you'd like a more gut-retching example, someone's legs smashed when a mechanism fails. Typically, we do not say that these individuals have been irrational for ignoring their fears and continuing to experience the ride. Rather, we would say that they were irrational if their fears *kept* them from riding the rides. We could also push this example further to everyday activities that are, according to base-rates, very dangerous, but in which are seemingly not irrational to engage. Take driving. Individuals who drive on a daily basis are continually putting themselves at risk. On Gendler's picture of rationality and base-rate neglect, if we have encoded base-rates correctly, then it seems the decision to ignore the base-rate risk of driving is engaging in irrational behavior; we suffer an epistemic cost for a practical reason (i.e. driving is the most convenient mode of transportation for most people). Couching base-rate neglect in objectively rational standards

[49] Brownstein's (forthcoming) expansion and discussions of the Skywalk example inform my understanding here.

[50] The Ten Meter Tower is a vivid example of individuals experiencing these tensions: <https://www.nytimes.com/2017/01/30/opinion/ten-meter-tower.html>.

not only dangerously simplifies the multiple cares and concerns of individuals, but is also sensitive to these and like counterexamples. In this section I have analyzed Gendler's conception of base-rate neglect to show that conceptualizing social category information as providing only negative information about individuals and focusing on an objectively rational application leads us astray. Rethinking the ways social categories may influence our assessments of individuals gives us further reason to reject The Irrelevance Assumption and provides evidence against Gendler's conclusion that we must choose between being rational and moral.

5. *Conclusion*

Most troubling about the suggestion that we should render social categories irrelevant to our evaluations of others is that it seems to commit one to a sort of in-principle colorblindness[51] and lack of cultural awareness. It also bars one from taking a culturally nuanced and intersectional approach to addressing systems of social inequity. Although irrationality may be the ultimate sin for philosophers, it seems this other sort of worry is far more important for those committed to building an equitable world. In this paper I have provided evidence against The Irrelevance Assumption, the claim that social categories are or should be irrelevant to our evaluations of individuals. Further, I have provided some positive reasons to demonstrate that social categories can play an epistemically and morally productive role. The rejection of The Irrelevance Assumption does not lead to a full-stop rejection of the conclusions of either paper; rather, it leads us to reframe questions about implicit bias mitigation, as well as positive ways to move forward until empirical methods are developed. When we conceptualize, develop, and test possible implicit bias mitigation strategies, we should focus on those that render salient important aspects of an individual's social identity, while limiting the effects of inaccurate stereotypes or pernicious associations[52].

[51] Although there are some scholars that think this is the ideal to which we ought aspire, I reject this view. For some empirical reasons for this rejection, see D. Kelly-E. Machery-R. Mallon, *Race and racial cognition*, in J. Doris-The Moral Psychology Research Group (eds.), *The Moral Psychology Handbook*, Oxford UP, Oxford 2010, pp. 433-472).

[52] I would like to thank Dan Kelly, Joshua Mugg, and an anonymous reviewer for comments on previous drafts of this paper. In addition, I'm grateful to the organizers (Jules Holdroyd, Alex Madva, Erin Beeghly) and participants of the Bias in Context workshop (2016) for their feedback and rich discussion.

Abstract

In this paper I reject the claim – made both by Tamar Szabo Gendler in On the Epistemic Costs of Implicit Bias and Jennifer Saul in Scepticism and Implicit Bias – that in order to be epistemically and morally responsible, social categories should not influence our evaluations of individuals or subsequent actions. I will provide evidence against the claim by denying its empirical plausibility, emphasizing the epistemic and moral benefits that may come from social categories, and reconceptualizing the inclusion of base-rate information. Throughout the paper I will emphasize the unique interlocking of epistemic and moral considerations that are relevant to implicit bias, bias mitigation, and responsibility. It is my hope that this analysis lays the groundwork for an account of the right ways social categories can affect our judgments, i.e. the ways in which such influence may improve our epistemic and moral situations rather than degrade them.

Keywords: bias; epistemically responsibility; morally responsibility; Tamar Szabo Gendler; Jennifer Saul.

Lacey J. Davidson
Department of Philosophy
Purdue University
davidsl@purdue.edu

Social Justice, Individualism, and Cooperation: Integrating Political Philosophy and Cognitive Sciences

Mario De Caro, Benedetta Giovanola

1. *Introduction*

It is an undeniable fact that, in many of their expressions, both political philosophy (modern and contemporary) and economics (think of Adam Smith, J.S. Mill, and the marginalist school) rest on individualistic anthropological underpinnings. The *homo oeconomicus* model presupposed by mainstream economic theory is a perfect illustration of that: according to the standard definition, this is a rational and self-interested agent who, when choosing, always pursues the maximization of his/her own well-being (generally understood in terms of utility): and, because of his/her calculating and self-centered qualities, the *homo œconomicus* has traditionally been intended as a very good economic agent – and, actually, as the *best* economic agent.

As to political philosophy, a clear example of the individualistic orientation is offered by the extremely influential Hobbesian metaphor of the *homo homini lupus* ("the human is a wolf to his fellow human"). Such metaphor perfectly expresses a conception of human nature that underlays many political views according to which, first, individuality is prior to sociality and, second, sociality is a cultural product generated by the necessity to live together in order to avoid a *bellum omnium contra omnes*[1]. From this perspective, even sovereignty as such rests on individualistic underpinnings, since it is the instrument that allows self-interested individuals preoccupied with their own well-being to live together. Thus, from

[1] Y. Evrigenis, *Images of Anarchy: The Rhetoric and Science in Hobbes's State of Nature*, Cambridge UP, Cambridge 2014.

this point of view humans are not naturally altruistic, civilization is established through the repression and control of instincts, and cooperation can only work at a local level, but not at a general one (for example, there will always be wars between different States).

It is important to notice that, because of the way in which they are defined, the *homo oeconomicus* and the *homo homini lupus* represent anthropological types constitutively unable to engage in authentic interpersonal relationships – individuals who, as it has been ironically noted, nobody would like a child of theirs to be married to[2]. For this reason, in recent years more than a doubt has been raised regarding the epistemological appropriateness and fecundity of these anthropological types. However, while the models based on the idea of the *homo oeconomicus* have been criticized both at the theoretical and the empirical level (by appealing to the findings of cognitive sciences)[3], the models based on the idea of the *homo homini lupus* have mainly been contrasted at the level of "pure" (i.e. theoretical) philosophical investigation, as done by the advocates of communitarianism and of recently revitalized cosmopolitanism, who characterize human nature in terms of a strong natural predisposition to pro-sociality and cooperation (which may sometimes be spoilt by society's historical and cultural needs).

Yet, since cognitive sciences have offered new significant contributions for understanding the attitudes and motivations of human action, it is very plausible that potentially they are also of use in the field of political philosophy. In particular, those sciences have significantly improved our knowledge of the psycho-biological roots of competition and cooperation in the human world, thereby offering us the opportunity to rethink the feasibility of the many political views that assume that self-assertiveness, egoism, and competition are natural human tendencies genetically and conceptually prior to pro-sociality and cooperation (which indeed are taken as merely culturally constructed attitudes).

[2] R.H. Frank, *Microeconomics and Behavior*, McGraw-Hill, New York 1991.

[3] Some of these critical investigations have underlined the cognitive biases at stake in economic choices and have pointed out the need both to abandon the "folk psychology" on which the standard notion of economic rationality relies (cf. D. Kahneman-A. Tversky, *Prospect Theory: an Analysis of Decision Under Risk*, in «Econometrica», 47 (1979), pp. 263-291; D. Kahneman-A. Tversky (eds.), *Choices, Values and Frames*, Cambridge UP, Cambridge 2000), and to highlight how the one-sidedness of the *homo œconomicus* model is not true to the psychological complexity of human choices. See P. Slovic *et al.*, *The Affect Heuristic*, in T. Gilovich-D. Griffin-D. Kahneman (eds.), *Heuristic and Biases: The Psychology of Intuitive Thought*, Cambridge UP, Cambridge 2002.

In order to illustrate this point, let's consider the discussion on social justice. In this field liberal theories are generally taken to presuppose individualistic views of the person and of cooperation (namely, cooperation just for mutual advantage, as conveyed by the appeal to the social contract)[4]. As we will show, nowadays there are good empirical reasons for thinking that these views are empirically inadequate. However, there are also good reasons for thinking that equally empirically inadequate are the communitarian and cosmopolitan views that, vice versa, give absolute priority to pro-sociality, altruism and cooperation (taken as natural tendencies) over self-assertiveness and competition (taken as culturally generated tendencies).

In our view, in order to make the reflection on social justice more reliable and effective, it is time to develop a sounder anthropological model, more aligned with the findings of cognitive sciences.

2. *Individuality and cooperation in the theories of justice*

Most contemporary theories of justice that have developed in the framework of liberalism, particularly under the influence of John Rawls's (1971) seminal work, can be seen as attempts to reflect on how different individuals can cooperate with one another in society, so as to shape it in ways that are fair and advantageous for everyone. From the Rawlsian perspective, society is taken as a "cooperative venture for mutual advantage"[5]. Cooperation produces a better life for all; however, individuals tend to compete for larger shares of the benefits coming from cooperation. Therefore "a set

[4] This is the standard view (which will be questioned in this article) and it is usually attributed to almost all liberal theories, including contemporary or "new" liberalism and liberal theories of social justice (such as J. Rawls, *A Theory of Justice*, Harvard UP, Cambridge (MA) 1971; W. Kymlicka, *Liberalism, Community and Culture*, Clarendon Press, Oxford 1989; R. Dworkin, *Sovereign Virtue*, Harvard UP, Cambridge (MA) 2000). In our view, individualistic conceptions of the person and of cooperation should rather be looked for in classical liberalism, which establishes an intimate relation between liberty and private property (for a discussion of these issues, cf. G.F. Gaus, *Property, Rights, and Freedom*, in «Social Philosophy and Policy», 11 (1994), pp. 209-240; and H. Steiner, *An Essay on Rights*, Blackwell, Oxford 1994), as well as in contemporary liberism (F.A. Hayek, *The Constitution of Liberty*, University of Chicago Press, Chicago 1960) and libertarianism (R. Nozick, *Anarchy, State and Utopia*, Basic Books, New York 1974). In fact, in the latter cases, the centrality attributed to individual freedom has led to the vindication of a decentralized market based on private property (F.A. Hayek, *op. cit.*) and, in the case of Nozick (*op. cit.*), to a complete rejection of all redistributive demands.

[5] J. Rawls, *op. cit.*, p. 4.

of principles is required for choosing among the various social arrangements which determine the division of advantages and for underwriting an agreement on the proper distributive shares" (*ibidem*). The "original agreement", as is well-known, takes the form of an ideal social contract that makes it possible to choose principles of justice that all "free and rational persons concerned to further their own interests would accept", when put in an initial position of equality, conveyed by the original position and the veil of ignorance (*ivi*, p. 10). The original agreement is therefore conceived as a device that guarantees the fostering of social cooperation on the one hand, and the free pursuit of individual interests, provided an initial situation of equality, on the other hand.

In criticizing Rawls's and the other liberal political views, communitarians tend to focus precisely on the centrality they attribute to the individual and on their conception of it. Michael Sandel, for example, famously criticized the appeal of those views to an abstract conception of individuals as pure autonomous choosers, whose commitments, values and concerns are possessions of the self, but never constitute the self itself, and might therefore be rejected. According to Sandel[6], this is a barren and "disencumbered" conception of the self, and in order to get a more adequate one, we would need to understand the social pre-conditions of self-determination.

In the communitarian perspective, the self is the outcome of a discovery rather than of an autonomous choice – since every person discovers who they are through their belonging to a community. Therefore the self is best expressed through a narrative conception[7], as the story of one person's life is embedded in the story of the communities from which she derives her identity. At last, communities – including the obligations of membership and solidarity they bring about – are prior to individuals, and pro-sociality and cooperation for the common good are prior to the appeal to individual freedom.

Summarizing, most contemporary views of social justice are based on either of two alternative couples of anthropological presuppositions. On the one side, the liberals who advocate the theory of justice assume that (i) individuals are naturally self-interested beings and (ii) cooperation is a social construct aimed at fostering individual interests. On the other side, communitarians assume that (i) individuals are naturally cooperative, as

[6] M. Sandel, *Liberalism and The Limits of Justice*, Cambridge UP, Cambridge-New York 1982, ch. 1.

[7] A. MacIntyre, *After Virtue*, University of Notre Dame Press, Notre Dame (IN) 1981.

they jointly pursue the common good of their community, and (ii) they derive their identity from their belonging to that community[8].

That said, in our view it is time to carry out the discussion on social justice, and on the anthropological presupposition of the different views, in the context of a sounder and empirically more reliable framework. In this way, one can realize that both sets of anthropological assumptions rely on oversimplifications and have been falsified in recent years. In particular, research in cognitive psychobiological sciences has shown that human beings are complex entities that behave in ways that cannot be described as purely competitive or purely cooperative: rather, in their behavior competition and cooperation *naturally* coexist[9]. For this reason, in order to be empirically adequate, theories of social justice should account for both the pursuit of self-interest and the forms of pro-sociality and cooperation that jointly characterize human beings.

3. *Individuality and cooperation in the light of cognitive sciences*

In the last couple of decades investigations of cognitive sciences (especially, in biology, sociology, behavioral economics and psychology) have made clear that sociality does not originate only from culture; rather, it is a dimension that belongs to the definition of the human individual itself. In fact, an impressive amount of empirical data has proven beyond reason-

[8] It may be noted that the advocates of cosmopolitanism – even if they generally endorse liberal principles and consider the individual person (rather than the government) as the main unit of concern – agree, at least partially, with communitarianism in regard to the anthropological underpinning of their views: in fact, also the cosmopolitan perspective is intrinsically social rather than merely self-interested and embedded in the community. However, the community at stake in cosmopolitanism is the whole humankind (cf. T. Pogge, *World Poverty and Human Rights: Cosmopolitan Responsibilities and Reforms*, Polity Press, Cambridge 2002; and S. Benhabib, *The Claims of Culture: Equality and Diversity in the Global Era*, Princeton UP, Princeton 2002), and this lets cosmopolitans depart from the communitarian focus on local communities.

[9] S. Bowles-H. Gintis, *The evolution of Strong Reciprocity: Cooperation in Heterogeneous Populations*, in «Theoretical Population Biology», 65 (2004), n. 1, pp. 17-28; R. Boyd-H. Gintis-S. Bowles-P.J. Richerson, *The Evolution of Altruistic Punishment*, in «Proc. Natl. Acad. Sci. Usa», 100 (2003), pp. 3531-3535; J. Henrich-R. Boyd-S. Bowles-C. Camerer-E. Fehr-H. Gintis, *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, Oxford UP, Oxford 2004; M. De Caro-M. Marraffa, *Bacon against Descartes. Emotions, Rationality, Defenses*, in G. de Anna-R. Martinelli (eds.), *Moral Realism and Political Decisions. Practical Rationality in Contemporary Public Contexts*, University of Bamberg Press, Bamberg 2015, pp. 63-80.

able doubt that individuals come to the world already endowed with the tendency to sociality, cooperation and even altruism. Excellent examples in this sense have been offered by Warneken and Tomasello[10], who have carried out some groundbreaking experiments showing that, since a very early age, humans are endowed with natural predispositions to cooperative and altruistic tendencies. Moreover, and even more surprisingly, those tendencies are present also in chimpanzees, our closest evolutionary relatives. The abstract of Warneken and Tomasello's article reads:

> Human infants as young as 14 to 18 months of age help others to attain their goals, for example, by helping them to fetch out-of-reach objects or opening cabinets for them. They do this irrespective of any rewards from adults (indeed external rewards undermine the tendency), and very likely with no concern for such things as reciprocation and reputation, which serve to maintain altruism in other children and adults. Humans' nearest primate relatives, chimpanzees, also help others instrumentally without concrete rewards. These results suggest that human infants are naturally altruistic, and as ontogeny proceeds and they must deal more independently with a wider range of social contexts, socialization and feedback from social interactions with others become important mediators of these initial altruistic tendencies[11].

Many other studies have confirmed that fairness, altruism and cooperative attitudes are very common in the animal world, especially but by no means only, among the primates[12]. Another important branch of research concerns the relevance of empathy, taken as a fundamental condition of prosocial attitudes and behavior, and of moral life[13]. Not less important are the investigations on the so-called "ultimatum game", which show that individuals tend to sanction other people's behavior when this is perceived as unfair, even though these individuals pay a price in terms of personal

[10] F. Warneken-M. Tomasello, *The Roots of Human Altruism*, in «British Journal of Psychology», 100 (2008), pp. 455-471.

[11] *Ivi*, p. 455.

[12] F. De Waal, *Primates and Philosophers: How Morality Evolved*, Princeton UP, Princeton 2006; Id., *The Age of Empathy: Nature's Lessons for a Kinder Society*, Harmony Books, New York 2009; Id., *The Bonobo and the Atheist: In Search of Humanism Among the Primates*, W.W. Norton, New York 2013; J.M. Burkart *et al.*, Nature Communications 5, Article number: 4747 (2014), doi:10.1038/ncomms5747; S. Yamamoto-R. Humle-M. Tanaka, *Chimpanzee Help Each Other Upon Request*, in «PLoS One», 4 (2014), n. 10, p. e7416.

[13] A. Coplan-P. Goldie (eds.), *Empathy: Philosophical and Psychological Perspectives*, Oxford UP, Oxford 2011; K. Stueber, *Empathy*, in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, URL=<https://plato.stanford.edu/archives/spr2017/entries/empathy/>.

utility for the sanctioning action (and there is no maximization of general utility either). Moreover, convincing data suggest that genetic factors play an important role in the shaping of human sensibility to fairness[14].

There is no doubt then that humans are naturally endowed with cooperative and altruistic tendencies. It would be wrong, however, to take the extreme stance – as communitarian and cosmopolitan thinkers often do – that human nature is one-sidedly cooperative and altruistic and that the individualistic and competitive behaviors only have a cultural and social origin. As a matter of fact, many investigations confirm that we are also naturally endowed with individualistic tendencies, which potentially produce conflicts (sometimes very destructive ones) with other individuals[15].

Taken together, all these findings show that human sociality complies with very complex natural predispositions and that individuals are bearers of a very complex suite of motivations (both individualistic and altruistic)[16]. Such motivations are intrinsically relational and they give place to complex situations of compromise between two motivational systems: the first committed to self-assertiveness and competition, the second aimed to pro-sociality and cooperation[17]. The specific equilibrium between these two motivational systems at which, within a particular situation, individuals arrive depend on their personal upbringing, social interactions, environmental influences and capacity of rationally controlling their own choices and actions.

The most important moral that follows from what precedes is that – whereas most Western conceptions take competition as natural and cooperation as a culturally-built device – according to this new bio-psychologically-informed anthropological paradigm, human beings are naturally inclined both to competition (sometimes even destructivity) and to several forms of sociality, cooperation, and even altruism. Moreover, once competition and cooperation are seen in this dialectic relationship, the new paradigm parts company also from the communitarian and cosmopolitan frameworks, which build on an excessively optimistic anthropology, according to

[14] B. Wallace *et al.*, *Heritability of Ultimatum Game Responder Behavior*, in «Proceedings of the National Academy of Sciences», 104 (2007), n. 40, pp. 1561-1564.

[15] N. Augoustinos-I. Walker-N. Donaghue, *Social Cognition: An Integrated Introduction*, Sage, London 2014.

[16] J.K. Murnighan-L. Wang, *The Social World as an experimental game*, in «Organizational Behavior and Human Decision Processes», 136 (2016), pp. 80-94.

[17] See M. Di Francesco-M. Marraffa-A. Paternoster, *The Self and Its Defences*, Palgrave-Macmillan, London 2016, pp. 47-48.

which there is nothing natural in competition and conflicts, since they only derive from cultural factors. In brief, neither of the two motivational systems is prior to the other and none can definitely prevail. On the contrary, the constant concurrence of the competitive motivational system and the cooperative one plays a crucial role in the human mind[18].

In the background of this dynamic, a complex interaction between our emotional system and rational reasoning is at work, in which neither has priority over the other. And also in this regard important work has been developed at the intersection of cognitive moral psychology and philosophy of mind, which should be taken into account if one wants to develop an empirically informed and nuanced enough new anthropological perspective[19].

4. *Social justice revised: integrating individualism and cooperation*

According to the data offered by cognitive sciences, individuals are bearers of a very complex suite of motivations. More specifically they (i) are naturally inclined to both competition and cooperation, (ii) have a natural tendency to fairness, (iii) are innately endowed with aversion to inequity.

The contribution that today cognitive sciences offer to the theories of justice is very relevant. Since cognitive sciences have shown that humans have a *natural* tendency to *cooperation*, the original agreement (or social contract) should not be conceived of as a mere social construct that safeguards individuals from the possible negative outcomes of the natural tendency to competition. Instead, the original agreement is rather to be seen as the social expression of a human natural need or desire to cooperate.

Moreover, our natural tendency to *fairness* provides reasons for explaining why the members of a society ought to agree on the fundamental principles that can foster a just society. They are willing to agree on the fundamental principles of justice, not only because they seek to pursue their own interests (which they think can be best secured through an agreement on the fundamental principles), but also because the search for justice is an innate constituent of human beings as such. In other words, appealing

[18] It is worth noticing that at the epistemological level, the dialectic between cooperation and competition can only be approached by multi-level explanations, which aim at capturing the connections between innate inclinations, formal relational invariants, and cultural conventions. See G. Jervis, *Individualismo e cooperazione*, Laterza, Roma-Bari 2002, pp. 167-170.

[19] M. De Caro-M. Marraffa, *Debunking the Pyramidal Mind: A Plea for Synergy Between Reason and Emotion*, in «Journal of Comparative Neurology», 524 (2015), n. 8, pp. 1695-1698.

to the individuals' natural predispositions, features and motivations to fairness helps to tackle the problem of justifying the social contract. Thus the interaction between philosophical inquiry and cognitive sciences can produce an empirically informed, and much more reliable, anthropological framework for the reflection on justice. In this perspective, individuals are not conceived of as motivated only by the pursuit of their own interest or advantage, but also by the pursuit of justice, taken as a value in itself.

It should be clear, however, that these findings are not at odds with the empirical commitments of Rawls's theory of justice. Rather, they are consistent with it; and actually they show a way for solving the impression of a tension intrinsic to that theory. In fact, at a closer scrutiny, the anthropological underpinnings of Rawls's theory are not exhausted by the notion of self-interested individuals (as in the passage mentioned above, he writes that «free and rational persons concerned to further their own interests would accept [the social contract]»). Rawls explicitly vindicates a conception of persons as moral entities that are moved by the highest-order interests to realize the two powers of moral personality, which are indispensable for a person to flourish: «the capacity for a sense of right and justice» and «the capacity to decide upon, to revise, and rationally to pursue a conception of the good»[20]. It is evident that these two moral powers presuppose the idea that humans are endowed with the capacity of being sociable and cooperative.

Even more clearly, Rawls claims that engaging in many forms of cooperation and being member of a community are conditions of human life[21] and that only in a social union is the individual complete[22]. In this perspective, the idea of social union opposes the notion of a private society, where individuals or associations «have their own private ends which are either competing or independent, but not in any case complementary»[23]. Contrary to private society, the idea of social union conveys the importance of complementarity and interdependency, which are in turn based on the social nature of humankind[24]. In other words, Rawls recognizes that «we need one another as partners in ways of life that are engaged in for

[20] J. Rawls, *Social Unity and Primary Goods* (1985), in S. Freeman (ed.), *Collected Papers*, Harvard UP, Cambridge (MA) 1999, pp. 359-387, p. 365; see also J. Rawls, *A Theory of Justice*, cit., p. 376.

[21] *Ivi*, p. 384.

[22] *Ivi*, p. 460, footnote 459.

[23] *Ivi*, p. 457.

[24] *Ivi*, p. 458.

their own sake, and the success and enjoyment of others are necessary for and complementary to our own good» (*ibidem*). And the idea of social union leads to the notion of «the community of humankind the members of which enjoy one another's excellences and individuality», and «they recognize the good of each as an element in the complete activity the whole scheme of which is consented to and gives pleasure to all» (*ibidem*)[25].

It seems, then, that the appeal to the social nature of humankind goes beyond a merely individualistic anthropological understanding. However, at the same time the problem araises of whether, and in case how, it can be reconciled with the idea of self-interested individuals who compete and cooperate just because they want to secure their own interests. And, as we have seen, the idea of such reconciliation is extraneous to both the liberal and the communitarian paradigms, which respectively prioritize individualism and cooperative attitudes.

However, few decades after Rawls developed his theoretical proposal, we have found evidence that, far from being a suspicious philosophical construction at odds with the main traditional proposals, it is empirically well-grounded. In particular the apparent tension between its social, altruistic, and cooperative components, on the one side, and its individualistic dimension, on the other side, is confirmed by the data that come from cognitive sciences.

On the one hand, as said, overwhelming experimental data show that human beings actually display a natural inclination to fairness and cooperation. On the other hand, we also have very good empirical reasons for believing that cooperation requires a certain kind of individualism, to be understood in terms of the individuals' capacity to be autonomous, to discover and actualize their unique potentials and talents and form their own identity – that is, to realize themselves[26]. Thus, both the social and the individualistic components of Rawls's theory of justice appear to be empirically confirmed by scientific findings and its anthropological underpin-

[25] Also other advocates of liberalism, besides Rawls, have tried to complement the individualism that characterizes that view starting with its founding fathers such as Locke and Mill. Therefore, besides claiming that we are autonomous choosers who employ liberty to construct our own lives, they have insisted that we also are social creatures: cf. W. Kymlicka (*op. cit.*) for an interesting attempt to advocate a theory of the self that finds room for both cultural membership and various attachments and commitments which at least partially constitute the self. Generally, however, these kinds of proposals are only supported by theoretical arguments: in our view they could benefit from also referring to the empirical findings we mention here.

[26] Cf. R. Guerini-M. Marraffa, *this volume*.

nings are enriched and made more consistent. Moreover, in this way one can also avoid the oversimplification of the communitarian perspective, according to which the very notion of the self rests on that of community and the individuals are supposed to have a sense of justice because they share common values with the community they belong to (and discover who they are through such a belonging) (De Caro, Giovanola and Marraffa, in preparation).

To sum up, by putting the findings of cognitive sciences in a dialogue with the philosophical inquiries regarding social justice, the theory of justice can be based on an anthropological model that is much sounder and much more reliable than those presupposed by the individualistic, on the one hand, and the communitarian and cosmopolitan models, on the other hand[27].

Abstract

The authors explore the contribution that this literature can offer to the field of political philosophy. In particular, the authors argue that, in order to make the reflection on social justice more reliable and effective, political philosophers must take into account the anthropological model emerging from what cognitive sciences tell us about self-assertiveness, egoism, competition, pro-sociality, cooperation and altruism.

Keywords: cognitive sciences; competition; cooperation; individualism; political philosophy; pro-sociality; social justice; self-assertiveness.

Mario De Caro
Dipartimento di Filosofia, Comunicazione e Spettacolo
Università Roma Tre/ Tufts University
mario.decaro@gmail.com

Benedetta Giovanola
Dipartimento di Scienze Politiche, della Comunicazione
e delle Relazioni Internazionali
Università di Macerata
benedetta.giovanola@unimc.it

[27] Even if the article was conceived jointly by the two authors, Benedetta Giovannola wrote the paragraphs 2and 4 and Mario De Caro wrote the paragraphs 1 and 3. We are indebted to Massimo Marraffa for many conversations and useful suggestions on the issues discussed in this article.

Ethics, Law, and Cognitive Science

Lockean Persons, Self-Narratives, and Eudaimonia

Rossella Guerini, Massimo Marraffa

In this article we explore the ethical import of a naturalistic form of narrative constructivism that distances itself from both the non-naturalistic and antirealist strands in the theorizing on the self[1].

Our criticism builds on William James' theory of the self. On the one hand, there is the "I", which is a unifying, integrative, synthesizing process – the "selfing" or "I-ing" process. On the other hand, there is the product of this process, the "Me", which famously comes in three different forms of experiential reflexivity: the material, social, and spiritual Me.

Against this Jamesian backdrop, the claim that we constitute ourselves as morally responsible agents (as "Lockean persons") by forming and using autobiographical narratives is combined with the realist claim that the narrative self is not an idle wheel but a layer of personality that serves as a *causal* center of gravity in the history of the human psychobiological system. This alliance between narrative constructivism and self-realism takes shape in the context of a tradition of thought that views the synthesis of the various strata of personality as the highest developmental point of the selfing process – a viewpoint that aligns with an ethic that hinges on the idea of *eudaimonia*: the discovery and actualization of our unique potentials and talents.

[1] This form of narrative constructivism has been developed in M. Di Francesco-M. Marraffa-A. Paternoster, *The Self and Its Defenses*, Palgrave-Macmillan, London 2016.

1. *Lockean persons I: the consciousness criterion*

In the second edition of the *Essay* Locke famously argues that person is a “forensic” notion and that the best way to capture its normative implications is through understanding it as a psychological category whose central concept is self-consciousness[2].

In this perspective, the concept of person is not an essence but rather a psychosocial attribute that is assigned to those subjects who possess a specific set of psychological capacities. This is in agreement with the most common legal language, which suitably speaks about “natural persons” and similarly about “legal persons”, thus pointing out something precise, i.e., the presence of an agent or subject who, in virtue of his intrinsic characteristics, is fully able to perform such acts as buying real estate, making a donation or a will, or paying taxes. Here the acting subject is a person precisely to the extent that he can be held (ethically even before legally) responsible for what he does. And he is thus imputable as well: if he committed a crime, he knew very well what he was doing. The concept of person therefore rests on that of *personal responsibility*; it is easy to see, even intuitively, that the concept of responsibility rests on the concept of self-consciousness, seen precisely as awareness of one’s own acts, and hence as *critical appropriation* of one’s own projects, actions, and memories. An individual can make a will only if he is a person – and indeed a child cannot make a will, nor even an elderly person who suffers from dementia; they are not sufficiently responsible inasmuch as they are not sufficiently aware of the meaning, scope and consequences of their actions.

Thus, as just been hinted, the Lockean person is someone who possesses a set of psychological capacities. It is someone who is able to form imaginary test scenarios in order to make a planning evaluation of what can happen as a consequence of his actions. But above all it is someone who is able to grasp himself not only as a material agent in his own present, past and future acts as “public” acts, but also as an entity who has inwardness, i.e., an inner experiential space in which thoughts and affects can be situated as “private” events. Only someone with sufficient access to his own interiority (to himself as objectified in the introspective consciousness of the self) can appropriate «Actions and their Merits»[3].

[2] J. Locke, *An Essay Concerning Human Understanding*, Clarendon Press, Oxford 1975 (orig. ed. 1694).

[3] *Ivi*, p. 346.

In Locke, therefore, an individual is a person only insofar as he can reflectively appropriate his actions and their meaning – an appropriation that originates from «that consciousness which is inseparable from thinking»[4]. The philosopher also realizes that the identity of persons «is not determined by identity or diversity of substance, which it cannot be sure of, but only by identity of consciousness»[5]. The Lockean consciousness is thus a "secular" notion; it is not a substance, and it severs ties with the soul.

A question arises, however: if the identity of persons is determined by consciousness, by what is consciousness determined? Locke makes appeal to (introspective) consciousness as the most psychological and less metaphysical notion he can conceive to define the concepts of person and identity. On closer view, however, this consciousness is a "strong" stand-in for the soul; it is, actually, still a sort of secularized soul. Despite the philosopher's good intentions, it is also described as a sort of essence. For all that, Locke's consciousness is still given a priori.

A different kind of consciousness can be found in psychological sciences: something that is constructed during life, which emerges from the multifarious qualities of the body and of human existence. It is from this standpoint that Locke's notion of personal identity will be reconsidered in the Section 4.

2. *James' I/Me distinction and McAdams' personological view of narrative identity*

In his seminal chapter on the "Consciousness of Self" James takes the Lockean analysis of the self one step further[6].

According to James, the self is a *process*, «the process of reflexivity which emanates from the dialectic between the "I" and "Me"»[7] . This is well captured by the personality psychologist Dan McAdams. He opposes his interpretation of James' theory of the self to the postmodernist theorizing on identity. According to Kenneth Gergen, for example, the postmodern identity is multiple, shattered, bereft of any reality except for what is

[4] *Ivi*, p. 335.

[5] *Ivi*, p. 345.

[6] W. James, *The Principles of Psychology*, Dover, New York 1950 (orig. ed. 1890).

[7] V. Gecas, *The Self-Concept*, in «Annual Review of Sociology», 8 (1982), pp. 1-33, p. 3.

socially constructed from time to time in everyday interactions[8]. And in his view, it's all to the good: actually the multiplicity of the self (which he describes as the "multiphrenic condition") is to be accentuated in order to allow the subject to expand itself in different directions, to evolve and to create ever new opportunities of personal growth. McAdams takes issue with Gergen: the latter misses a fundamental aspect of selfhood, namely, the process of synthesizing the disparate elements that constitute the postmodern identity. This unifying activity corresponds to James' concept of the self as subject or "I"[9].

In this perspective, the I is not a thing, not even a part, a component or an aspect of the self: «[it] is really more like a verb; it might be called "selfing" or "I-ing", the fundamental process of making a self out of experience»[10]. The "Me" is instead «the primary product of the selfing process»; it is «the self that selfing makes»[11]. The Me consists in three forms of reflexive experientiality – the material, social and spiritual selves – which originate from the selfing process. It is «the making of the Me that constitutes what the I fundamentally is»[12].

James' I/Me distinction provides thus a definition of self-consciousness in terms of identity: self-consciousness is a self-describing, an identity forming, which is a unifying, integrative, synthesizing process. In this perspective, James anticipates a number of theories in developmental and personality psychology that have made appeal to a general organismic process for integrating subjective experience, – e.g., Werner's orthogenetic principle, Piaget's organization, and Jung's individuation[13]. While these various concepts differ from each other in important ways, they converge on the idea that human experience tends toward a fundamental sense of unity in that human beings apprehend experience through an integrative selfing process.

* * *

[8] K.J. Gergen, *The Saturated Self*, Basic Books, New York 1991.

[9] D.P. McAdams, *The Case for Unity in the (Post)Modern Self: a Modest Proposal*, in R.D. Ashmore-L. Jussim (eds.), *Self and Identity. Fundamental Issues*, Oxford University Press, Oxford 1997, pp. 46-78.

[10] D.P. McAdams, *Personality, Modernity, and the Storied Self: a Contemporary Framework for Studying Persons*, in «Psychological Inquiry», 7 (1996), n. 4, pp. 295-321, p. 302.

[11] *Ibidem.*

[12] D.P. McAdams-K.S. Cox, *Self and Identity Across the Life Span*, in R.M. Lerner (ed.), *The Handbook of Life-Span Development*, Wiley, New York 2010, vol. 2, pp. 158-207, p. 162.

[13] See R.M. Ryan, *Psychological Needs and the Facilitation of Integrative Processes*, in «Journal of Personality», 63 (1995), n. 3, pp. 397-427.

In McAdams' influential life-story model of identity, James' I/Me distinction is combined with Erik Erikson's theory of psychosocial development and Henry Murray's research program on the Study of Lives. Narrative identity is here defined as the internalized and evolving story of the self[14] which integrates the reconstructed past and the imagined future to provide life with some degree of unity, purpose and meaning. That is, people make sense of their own lives through narrative structures (such as characters, roles, scenes, scripts, and plots) which make the Me into «an internalized drama»[15].

Most importantly, McAdams views narrative identity as a layer of personality. Within his conceptual framework for conceptualizing the whole person across the life span[16], narrative identity hinges on two other cognitive layers. The first consists of a small set of broad *dispositional traits* implicated in social life (including the so-called "Big Five") which account for consistencies in behavioral style from one situation to the next and over time. The second layer consists of a wide range of *characteristic adaptations* (including goals, strivings, personal projects, values, interests, defense mechanisms, coping strategies, relational schemata) which capture more socially contextualized and motivational aspects of psychological individuality. During personality development, people's internalized and evolving life stories are layered over characteristic adaptations, which are, in turn, layered over dispositional traits. And this process of layering may be *integrative*: the process of selfing may succeed in bringing traits, skills, goals, values, and experiences into a meaningful life story.

Building upon Erikson's seminal approach to identity development, McAdams argued that the selfing process begins to arrange the Me into a self-defining narrative in adolescence, partly as a function of societal expectations regarding identity and the maturation of formal operational thinking[17]. Constructing and internalizing a life story provides an answer to Erikson's key identity questions – questions regarding who one is, how one came to be and where one is going in life.

[14] «[T]he broad narrative of the Me that the I[-ing] composes, edits, and continues to work on» (D.P. McAdams-K.S. Cox, *op. cit.*, p. 169).

[15] *Ibidem*.

[16] D.P. McAdams, *The Art and Science of Personality Development*, Guilford Press, New York-London 2015.

[17] D.P. McAdams, *Power, Intimacy, and the Life Story: Personological Inquiries into Identity*, Dorsey Press, Homewood (IL) 1985.

The earliest drafts of narrative identity may take the form of what has been called "the personal fable", i.e., the adolescent's grandiose fantasies about accomplishment, fame, or notoriety in the future[18]. But later drafts become more realistic and tempered, as reality testing improves and narrative skills become further refined. Habermas and Bluck (2000) have shown how adolescents gradually master the social-cognitive skills required for constructing a coherent narrative of the self[19]. By the end of their teenaged years, they regularly engage in sophisticated forms of *autobiographical reasoning*.

Autobiographical reasoning is a constructive and interpretative activity that relies on the life story format for drawing connections between remembered events and enduring and current characteristics of the self. This activity is based on four social-cognitive capabilities: (i) the ability to put past events in temporal order (temporal coherence); (ii) the ability to think about the self in abstract terms (i.e., as embodying certain personality traits) and account for changes or developments in the self over time (causal-motivational coherence); (iii) the ability to summarize and interpret themes within stories and apply these to one's own life (thematic coherence); and (iv) having normative expectations, shaped as they are by both biology and culture, regarding how a typical life is structured (the "cultural concept of biography"). Although a life narrative begins to emerge in middle childhood, temporal and causal-motivational coherence increase substantially across adolescence up to early adulthood, as does thematic coherence, which continues to develop throughout middle adulthood[20].

It is to be observed that autobiographical reasoning is *constitutive* of narrative identity. Embedding personal memories in a culturally, temporally, causally and thematically coherent life story, the life story format establishes and re-establishes the diachronic continuity of the self[21].

[18] See D. lkind, *Egocentrism in Adolescence*, in «Child Development», 38 (1967), n. 4, pp. 1025-1034.

[19] T. Habermas-S. Bluck, *Getting a Life: The Emergence of the Life Story in Adolescence*, in «Psychological Bulletin», 126 (2000), n. 5, pp. 748-769.

[20] See C. Köber-F. Schmiedek-T. Habermas, *Characterizing Lifespan Development of Three Aspects of Coherence in Life Narratives: A Cohort-sequential Study*, in «Developmental Psychology», 51 (2015), n. 2, pp. 260-275.

[21] T. Habermas-C. Köber, *Autobiographical Reasoning Is Constitutive for Narrative Identity: the Role of the Life Story for Personal Continuity*, in K.C. McLean-M. Syed (eds.), *The Oxford Handbook of Identity Development*, Oxford UP, Oxford 2015, pp. 149-165; Idd., *Autobiographical Reasoning in Life Narratives Buffers the Effect of Biographical disruptions on the sense of self-continuity*, in «Memory», 23 (2015), n. 5, pp. 664-674.

3. *Lockean persons II: the self-narrative criterion*

The claim that the type of continuity that connects psychological states across time in an identity-constituting way is specifically narrative in character is typically associated with concerns about *practical identity* (i.e., personal identity considered in its connection to ethical concerns, as Locke's theory of person does). The claim is that we constitute ourselves as Lockean persons by forming and using autobiographical narratives. The unity of a person is the unity of an autobiographical narrative.

In some cases, narrative accounts of personal identity are characterized in opposition to what has been, at least until quite recently, the most popular view of personal identity: a significantly amended version of Locke's relational memory criterion[22]. Here the question is one of "reidentification": on what basis should we reidentify a person as numerically the same despite qualitative differences over time or under different descriptions? Answering such a question calls for a criterion of diachronic numerical identity, a criterion of what makes something one and the same thing as itself at different times. But when the focus shifts from solely metaphysical puzzles about the persistence of complex objects (such as the ship of Theseus) to the relation between identity and practical and evaluative concerns, the question becomes one of "characterization": which characteristics (character traits, motivations, values, mental and bodily capacities and dispositions, emotional attachments, commitments, memories, and so on) make a person the particular person that she is? Such a question concerns «identity in the sense of what is generally called, following Erikson, an "identity crisis"»[23].

According to the proponents of the narrative view, an answer to the question of characterization may proceed in two steps. First, those activities of *self-interpretation* and *self-creation* that are central to our experience of being persons – so central that to many continuation without them (say, in a severely demented or vegetative state) is as bad as death – are built into the kind of continuity that connects person A and person B across time in an identity-constituting way[24]. Second, what enables persons

[22] See D. Shoemaker, *Personal Identity and Ethics*, in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, URL = <https://plato.stanford.edu/archives/win2016/entries/identity-ethics/>.

[23] M. Schechtman, *The Constitution of Selves*, Cornell UP, Ithaca 1996, p. 2.

[24] See, e.g., C. Korsgaard, *Personal Identity and the Unity of Agency: A Response to Parfit*, in «Philosophy and Public Affairs» 18 (1989), n. 2, pp. 109-123; D. De Grazia, *Human Identity and Bioethics*, Cambridge UP, Cambridge 2005.

to be actively self-interpreting and self-creating agents is identified with the construction of *self-narratives*[25].

This narrative thesis, however, comes in different forms. Authors such as Alisdair Macintyre and Charles Taylor view the person as a self-interpreting being in a sense inspired by the hermeneutical tradition, namely a tradition that is largely foreign to naturalistic commitments[26]. A psychologically plausible narrativist account of personal identity requires a view of self-interpretation as an activity of narrative reappropriation of the products of the unconscious processing – an activity implemented by apparatuses such as Dennett's Joycean machine or Gazzaniga's interpreter module or Carruthers' mindreading system[27]. In this perspective, persons are self-interpreting beings in a psychological sense that is congenial to Locke's forensic view of personal identity, but fundamentally foreign to the hermeneutical tradition. From our point of view the problem is that a hermeneutical notion of self-interpretation, insofar as it puts exclusive emphasis upon meaning (i.e., the intentional directing of consciousness) at the expense of the psychobiological theme of the unconscious, surreptitiously reintroduces the pre-psychoanalytic, pre-cognitivist, idealistic conception of the conscious subject as primary subject[28].

Things are similarly problematic in the case of the most rigorous psychoanalytic hermeneutics. Ricœur made a significant attempt to conciliate between Freud's metapsychology and hermeneutics[29]. For this philosopher investigated how psychoanalysis allows for both the hermeneutical theme of meaning and intentionality and the objective and biological theme of drive causality. Within this framework, Ricœur rejects the versions of psychoanalytic interpretation which are unilaterally aimed at the subjective or intersubjective reconstruction of meaning, in keeping with the standards of interpretive conventionalism. According to the latter, interpretation is ultimately committed to the freedom of deciding the meaning of the text on

[25] See J.W. Schroer-R. Schroer, *Getting the Story Right: a Reductionist Narrative Account of Personal Identity*, in «Philosophical Studies», 171 (2014), pp. 445-469.

[26] A. Macintyre, *After Virtue*, University of Notre Dame Press, Notre Dame 1984; C. Taylor, *Sources of the Self*, Harvard UP, Cambridge (MA) 1989.

[27] D.C. Dennett, *Consciousness Explained*, Little Brown, Boston 1991; M. Roser-M.S. Gazzaniga, *Automatic Brains. Interpretive Minds*, in «Current Directions in Psychological Science», 13 (2004), n. 2, pp. 56-59; P. Carruthers, *The Opacity of Mind*, Oxford UP, Oxford 2011.

[28] See G. Jervis, *La psicoanalisi come esercizio critico*, Garzanti, Milan 1989.

[29] P. Ricœur, *Freud and Philosophy: An Essay on Interpretation*, Yale UP, New Haven 1970 (orig. ed. 1965).

the strength of the agreement reached by the participants to the interpretive operation. But in this way the problems of truth and reality, of adequacy and verification, tend to disappear, being replaced by a freely creative narrativism of postmodernist type[30].

Ricœur's attempt at synthesis, however, remains within a conception of the unconscious that must be rejected. He coins the term "anti-phenomenology" to define Freud's methodological approach. According to Ricœur, Freud's establishment of the unconscious is «an *epoch* in reverse» because «what is initially best known, the conscious, is suspended and becomes the least known»[31]. Consequently, whereas the phenomenological tradition pursues a reduction of phenomena *to* consciousness, capturing them as its objects, Freud's methodological approach aims at a reduction *of* consciousness: the latter loses the Cartesian character of first and last certainty, which stops the chain of methodical doubts on the real, and becomes itself an object of doubt. Psychoanalysis becomes thus a *demystifying hermeneutics*. This project of *demystification* – the systematic search for self-deception and the uncovering of underlying truth – is at the core of the critical tradition to which Freud belongs: the "unmasking trend" that has been part of European thought from La Rochefoucauld through Enlightenment philosophers, Marx, Nietzsche, and Ibsen[32].

There is a problem, however. Freud's inquiry into the unconscious actually starts from consciousness taken as *given*, and this makes psychoanalysis a dialectical variant of phenomenology. In contrast, a dynamic psychology informed by the cognitive sciences is not vulnerable to this objection: it aims to pick up the critical content of psychoanalysis – its being a demystifying project – but within a framework where consciousness is at issue and the unconscious is understood in terms of an conception of the relationship between the subpersonal and personal levels of analysis in which the former is always in a dialectical relationship with the latter[33].

Certainly, even if we define self-interpretation as a re-appropriation of the products of the human information-processing machinery, self-narratives are *not* merely the result of the workings of a psychobiological apparatus. Socio-cultural variables may significantly modulate the construction

[30] See M.N. Eagle, *The Postmodern Turn in Psychoanalysis: A Critique*, in «Psychoanalytic Psychology», 20 (2003), n. 3, pp. 411-424.

[31] P. Ricoeur, *Freud and Philosophy*, Yale University Press, New Haven 1970 (orig. ed. 1965), p. 118.

[32] H.F. Ellenberger, *The Discovery of the Unconscious*, Basic Books, New York 1970, p. 537.

[33] See M. Di Francesco-M. Marraffa-A. Paternoster, *op. cit.*, ch. 1.

of psychological self-consciousness. Data from cultural psychology and ethnopsychiatry show that people living in small-scale societies possess a self-consciousness that is primarily physical and social rather than psychological. The construction of psychological self-consciousness requires a repertoire of conceptual and (indissolubly) lexical tools of an abstract kind. As hinted above, the capacity to think in a hypothetico-deductive manner enables to grasp one's inner life in terms of autobiography. By contrast, the intelligence of adult illiterates living in small-scale societies is entirely focused on immediate practical experience, and therefore lacks the necessary resources to make the complete shift from a physical to a psychological form of self-consciousness[34].

Yet, whereas the narrative theorists of personal identity tend to make the socially and historically situated narrative self *the* foundational aspect of human selfhood, we think that the narrative self is only one of the three dimensions of the Jamesian Me, which evolves from the bodily subjectivity. This point emerges very clearly from Mark Howe and Mary Courage's account of the genesis of autobiographical memory[35].

Most of the theories of autobiographical memory development have been cast in terms of understanding why infantile amnesia ends (and presumably true autobiographical memory begins) at the beginning of the preschool period. According to Howe and Courage, children lack a critical cognitive or social-cognitive framework before that period that would enable them to encode memories in such a way that they could later be retrieved as self relevant. This framework is self-consciousness, as commonly measured in the mirror task of self-recognition. Before children pass the task at about 18 months to 2 years, they are not capable of encoding and storing memories as self relevant. As a consequence, there is no *auto* in autobiographical. Later, when trying to retrieve these memories from the perspective of things that happened to "me", they are unsuccessful because they did not yet have a 'me' to which to attach the memory.

Now, we agree with Howe and Courage that the most important factor in the emergence of autobiographical memory is self-consciousness as measured in the mirror self-recognition task. However, we take issue with the

[34] See M. Marraffa-C. Meini, *From Piaget to Bowlby – and Back Again*, in «Paradigmi», 35 (2017), n. 3, in press.

[35] M.L. Howe-M.L. Courage, *On Resolving the Enigma of Infantile Amnesia*, in «Psychological Bulletin», 113 (1993), pp. 305-326; M.L. Howe, *The Co-emergence of the Self and Autobiographical Memory*, in P.J. Bauer-R. Fivush (eds.), *The Wiley Handbook on the Development of Children's Memory*, Wiley-Blackwell, Hoboken (NJ) 2014, pp. 545-567.

authors' construal of the fixed referent as a "cognitive self-concept" because we agree with those researchers who take mirror recognition as a marker of *bodily* self-consciousness[36], and hence reject the claim according to which children's mark-directed behavior is evidential of an introspective form of self-consciousness and a self-concept inherently linked to understanding the mental states of other people[37]. Our sense of ourselves in time is rooted in the onset of a *physical* form of self-describability: the nonverbal, analogic representation of the bodily self constructed in the second year of life acts as a fixed referent around which personally experienced event memories begin to be organized.

The Me to which the subject begins to attach episodic memories is thus the Jamesian material self. With the permission of the postmodernist reflection on identity, self-narratives do not create selves. The autobiographical self as a continuity across time and space, interpreted reflectively by the agent, would not arise without the bodily subjectivity. Bodily self and narrative self are two different kinds of experiential unity produced by the dialectic between the I and the Me.

4. *Realism about the self: autonomy and individuation*

In this process of narrative self-construction there is an essential psychodynamic ingredient.

During very early childhood, and especially from the third year of life, self-consciousness may go beyond the bodily subjectivity to become psychological self-description, and later, narrative self-description. This description of the self that the young child feverishly pursues is an "accepting description", i.e., a description that is indissolubly cognitive (as a *definition* of self) and emotional-affectional (as an *acceptance* of self). In practice, therefore, affective growth and the construction of identity cannot be separated. The child needs a clear and consistent capacity to describe herself – a capacity which is fully legitimized by caregivers, and socially valid.

On the other hand, this will continue to be the case throughout the en-

[36] See, e.g., D.J. Povinelli, *The self: Elevated in Consciousness and Extended in Time*, in C. Moore-K. Lemmon (eds.), *The Self in Time: Developmental Perspectives*, Erlbaum, Mahwah (NJ) 2001, pp. 75-95.

[37] See, e.g., J.P. Keenan-G.C. Jr. Gallup-D. Falk, *The Face in the Mirror*, Ecco, New York 2003.

tire life cycle. Adolescent crisis, and together with it the process of social autonomization in post-adolescence, is largely a problem of identity. In Erikson's theory of identity development, evoked above by Schechtman, the fundamental problem of adolescence lies in moving from a *heteronomous* identity to an *autonomous* self-definition. This requires an identity synthesis, i.e., a reworking of childhood identifications into a larger, self-determined set of self-identified ideals. The optimal outcome of such a process is a kind of dialectic balance in which the ego syntonic pole of identity synthesis is predominant over the ego dystonic pole of identity confusion (i.e., an inability to develop a workable set of ideals on which to base an adult identity).

Erikson sees identity confusion as an insufficient integration of self-images originating from a "weakness of the ego"[38]. This claim leads us into the psychopathological dimension of the inextricable link between identity self-description and self-consciousness. One cannot ascribe concreteness and solidity to one's own self-consciousness if it does not possess at its center, and as its essence, a description of identity that must be clear and, inextricably, "good", in the sense of being worthy of love[39]. If the self-description becomes uncertain, the subject soon loses the feeling of being present.

We can say then that the incessant construction and reconstruction of an acceptable and adaptively functioning identity is the process that produces our intra- and inter-personal balances, and is thus the foundation of psychological well-being and mental health. And this process is the ongoing construction of a system of defenses, the continuously renovated capacity to curb and cope with anxiety and disorder[40]. Consider, for example, the above-mentioned autobiographical reasoning. This is essentially a mechanism to compensate for threats of self-discontinuity. In circumstances of relative stability, personal sameness in time or personal stability may be established by the mechanism whereby the remembered self is systematically distorted by automatically assimilating it to the present self-concept, increasing the similarity between the present and remem-

[38] It is to be noticed that in this context Freud's *das Ich* is taken as a synthetic function, a synthesizing process, and thus coinciding with selfing. See D.P. McAdams, *The Case for Unity…*, cit., p. 57.

[39] See M. Balint, *Primary Love and Psycho-Analytic Technique*, Tavistock, London 1965, pp. 90-108 (orig. ed. 1937).

[40] See G. Jervis, *Contro il sentito dire. Psicoanalisi, psichiatria e politica*, edited by M. Marraffa, Bollati Boringhieri, Torino 2014.

bered reflected self, in order to maintain conceptual self-sameness[41]. When change is acknowledged, however, such a mechanism fails to create self-continuity. In circumstances of biographical change, the diachronic continuity of the self can be re-established by autobiographical reasoning through arguments that spell out transformations and their motives[42].

The selfing process imposes thus a teleology of self-defense on the human psychobiological system; and here is where the argument for a realist view of the self takes off. The self is the process of reflexivity which emanates from the dialectic between the Jamesian I and Me. And unlike the continuously self-rewriting autobiographies of Dennett's Joycean machine[43], the storied Me that the selfing process makes is not an epiphenomenon, but rather a layer of personality that serves as a *causal* center of gravity in the history of the system[44].

Conceiving narrative identity as a causally efficacious layer of personality pre-empts a standard antirealist objection. Narrativism, so the objection goes, is an approach that puts normative constraints on our self-narratives – constraints such as "narrative coherence". But what prevents from suspecting that «a person may possess a completely coherent self-identity that is nevertheless false»[45]? Realists are thus required to offer criteria by which they can distinguish between self-narratives that are truthful and those that are confabulated, self-deceptive, or paranoid[46]. And here is where a personological view of the narrative self comes into play.

As seen above (§2), during personality development, internalized and evolving stories of the self layer over adaptations, which layer over traits, and this process of layering may be *integrative*: «Traits capture the actor's dramaturgical present; goals and values project the agent into the future. An autobiographical author enters the developmental picture […] to inte-

[41] M.A. Conway-J.A. Singer-A. Tagini, *The Self and Autobiographical Memory: Correspondence and Coherence*, in «Social Cognition», 22 (2004), n. 5, pp. 495-537.

[42] See T. Habermas-C. Köber, *Autobiographical Reasoning is Constitutive for Narrative Identity*, cit.; Idd., *Autobiographical Reasoning in Life Narratives Buffers…*, cit.

[43] These autobiographies are only «a confabulatory byproduct of the decentralized brain activity that actually regulates behavior» (J. Ismael, *Saving the Baby: Dennett on Autobiography, Agency, and the Self*, in «Philosophical Psychology», 19 (2006), n. 3, pp. 345-360, p. 346).

[44] See O. Flanagan, *Consciousness Reconsidered*, MIT Press, Cambridge (MA) 1992, p. 195; J. Ismael, *Saving the baby*, cit., p. 353; M. Di Francesco-M. Marraffa-A. Paternoster, *op. cit.*, ch. 5.

[45] K. Kristjánsson, *The Self and its Emotions*, Cambridge UP, Cambridge 2010, p. 39.

[46] See S. Matthews-J. Kennett, *Truth, Lies, and the Narrative Self*, in «American Philosophical Quarterly», 49 (2012), n. 4, pp. 301-315.

grate the reconstructed past with the experienced present and envisioned future»[47]. The selfing process, then, takes the form of what Jung identified as individuation, namely, a search for itself that strives for a synthesis of the various strata of personality[48].

Such a process has an ethical dimension that is reminiscent of the Aristotelian ideal of *eudaimonia*. Most relevantly for our purposes, eudaimonia can be reinterpreted in terms of identity[49]. The good life can be seen, with Aristotle, as the *telos* at which the best human conduct aims but, differently than Aristotle, as a *telos* not preordained to the individual but immanent to the vicissitudes of one's mental life. To act in accordance with virtue cannot mean to perform well the task most typical of the human being *in general*, but to perform well «the task of maintaining the integrity of one's identity in the plurality of situations one encounters and of expressing the salient traits of one's identity in a unique biography»[50]. Although this task confronts every person, its content varies from individual to individual and cannot be known a priori: «The good life or *eudaimonia* [...] is then a life-course in which one is able to enrich the main plot of one's life-narrative with the largest possible amount of episodes and sub-plots compatible with the preservation of a sense of overall unity. The ability to unify one's biography into a coherent narrative is a good which plays a similar role to *eudaimonia* for Aristotle»[51].

In this personological and eudaimonic framework[52], a criterion that affords a distinction of self-knowledge from self-deception becomes available. Deceptive self-narratives are those that fail to integrate with the other layers of personality. Telling a coherent self-story is then not enough: a fully coherent but false self-narrative is a "façade" marked by bad faith, something inauthentic which tends to pass itself off as the "deep" structure of the person. Such a narrative is an idle wheel within the process of individuation.

The model of self-knowledge implied here is psychotherapeutic as well

[47] D.P. McAdams, *Tracing Three Lines of Personality Development*, in «Research in Human Development», 12 (2015), nn. 3-4, pp. 224-228, p. 226.

[48] C.G. Jung, *Collected Works, vol. 6, Psychological Types*, Routledge, London 1971 (orig. ed.).

[49] A. Ferrara, *Reflective Authenticity: Rethinking the Project of Modernity*, Routledge, London 1998.

[50] *Ivi*, p. 31.

[51] *Ibidem*.

[52] Research on *eudaimonia* and eudaimonic well-being has proliferated recently in personality psychology. For a review, see A.S. Waterman, *Eudaimonic Identity Theory: Identity as Self-discovery*, in S.J. Schwartz-K. Luyckx-V.L. Vignoles (eds.), *Handbook of Identity Theory and Research*, Springer, Berlin 2011, pp. 357-379.

as ethical. Biographies may be soliloquies, but they are also presented socially. This typically occurs in psychotherapy, and biographies serve then as vehicles for negotiations of identity[53]. In this perspective, the construction of a self-narrative characterized by the Lockean critical appropriation of one's own actions and mentations (§1) can be seen as a patient-therapist exchange of autobiographical arguments (§2) in which illusions and self-deceptions are rooted out and dispelled. This can be seen as an exercise of demystifying hermeneutics whose criterion of objectivity lies in a dynamic psychology driven by the cognitive sciences. In this psychotherapeutic context, the individual's "actual self" – what Flanagan called the "actual full identity"[54] – is the life story as told from the "ideally objective standpoint" of a subpersonal theory which is always in dialectical relationship with the personal level of analysis (§3)[55].

5. *Conclusions*

This article explored the ethical import of a naturalistic and realist version of the narrative view of the self.

First, we distanced from the non-naturalistic strands in the hermeneutical conception of narrative identity by making a case for a demystifying approach which finds its criterion of objectivity in a dynamic psychology informed by the cognitive sciences.

Second, we made a case for realism about the Jamesian duplex self since the process of self-representation originated from the I/Me dialectic is not an idle wheel but a causal center of gravity in the history of the agent. Antirealists understimate this point. Dennett, for example, affirms that the self only serves «to solve the myriad *little problems of interpersonal activity* we encounter every day, from the moment of our birth»[56]. In con-

[53] See J.M. Doris, *Talking to Our Selves*, Oxford UP, Oxford 2015.

[54] Actual full identity is «the self as seen from the point of view of a certain class of theoretical perspectives that admit the reality of the self as an emergent phenomenon and try to give an objective account of what it, in general and in particular, is like» (O. Flanagan, *Varieties of Moral Personality: Ethics and Psychological Realism*, Harvard UP, Cambridge (MA) 1991, p. 137).

[55] Thus we take very seriously Owen Flanagan's worry that theories from cognitive sciences may «couch the explanation of action in unfamiliar scientific terms, not in terms of the theory of action framed in the common sense language of ideals and commitments» (review of K. Kristjánsson, *The Self and its Emotions*, in «Notre Dame Philosophical Reviews», 2012, <http://ndpr.nd.edu/news/35356-the-self-and-its-emotions/>).

[56] D.C. Dennett, *Artifactual selves: A response to Lynn Rudder Baker*, in «Phenomenology and Cognitive Sciences», 15 (2016), n. 1, pp. 17-20, p. 16; italics ours.

trast, findings from developmental, dynamic, social and personality psychology show that our entire life takes shape in accordance with a primary need to exist solidly as *unitary* subjects.

The integrative selfing process gives rise to different kinds of unity, corresponding to the different aspects of the Me-self. The most minimal form of the Me is bodily self-awareness; the storied Me arises from such a material self. On the other hand, it is the psychological unity – and notably the unity of an autobiographical narrative – that constitutes ourselves as Lockean persons. The most fundamental unity is the integration of the personality layers, in agreement with an ethic hinged on the ideal of eudaimonia – the discovery and actualization of one's own unique potentials and talents.

Abstract

In this article we explore the ethical import of a naturalistic form of narrative constructivism that distances itself from both the non-naturalistic and antirealist strands in theorizing on the self. Our criticism builds on William James' theory of the self. Against this Jamesian backdrop, the claim that we constitute ourselves as morally responsible agents (as "Lockean persons") by forming and using autobiographical narratives is combined with the realist claim that the narrative self is not an idle wheel but a layer of personality that serves as a causal *center of gravity in the history of the human psychobiological system. This alliance between narrative constructivism and self-realism takes shape in the context of a tradition of thought that views the synthesis of the various strata of personality as the highest developmental point of the selfing process – a viewpoint that aligns with an ethic that hinges on the idea of* eudaimonia*: the discovery and actualization of our unique potentials and talents.*

Keywords: eudaimonia; demystifying hermeneutics; individuation; narrative identity; personality; realism about the self; William James' I/Me distinction.

Massimo Marraffa
Dipartimento di Filosofia, Comunicazione e Spettacolo
Università Roma Tre
massimo.marraffa@uniroma3.it

Rossella Guerini
Dipartimento di Filosofia, Comunicazione e Spettacolo
Università Roma Tre
rossella.guerini@uniroma3.it

Neurolaw and Punishment: a Naturalistic and Humanitarian View, and its Overlooked Perils

Andrea Lavazza

1. *Neurolaw as a naturalization of law*

Neurolaw is the approach that attempts to apply recent progress in neuroscience to the classic conceptions of law, often with the aim of pushing legal institutions (especially in criminal law) to be more in line with scientific knowledge[1]. This is essentially a process of naturalization *a' la* Quine applied to an area – law – that so far has been largely unaffected by naturalization. This also applies to punishment, its aims, its methods of implementation and its justification.

Two kinds of issues arise when applying neuroscientific findings to the law[2]. The first, called internal, are already being tackled by present institutions (for example, cases of imputability) and do not involve any major modifications, but only partial adjustments in some cases. A classic example is that of the legal age of majority, which can vary from system to system, and from country to country. The conventionalistic element is obviously predominant in the decision to place the age of legal responsibility at 18 rather than 16 or 21, but this choice has always been also linked to the psychological knowledge available at the time. Today, however, we know that the maturation of the prefrontal cortical areas of the brain, critical for controlling behavior and modulation of instinctive-impulse response, continues throughout adolescence and part of youth, until at least

[1] M.S. Pardo-D. Patterson, *Minds, Brains, and Law: The Conceptual Foundations of Law and Neuroscience*, Oxford University Press, New York 2013; D. Patterson-M.S. Pardo (eds.), *Philosophical Foundations of Law and Neuroscience*, Oxford University Press, New York 2016; A. Lavazza-L. Sammicheli, *Il delitto del cervello. La mente tra scienza e diritto*, Codice, Torino 2012.

[2] B.N. Waller, *Against Moral Responsibility,* MIT Press, Cambridge (MA) 2011.

age 20-22. This may have consequences for the decision whether or not to punish a young person who has committed certain types of crimes. It is no coincidence that the US Supreme Court, when deciding on the constitutionality of the death penalty for juveniles (Roper v. Simmons, 2005), also heard the opinions of neuroscientists. The decision to declare the death penalty for juveniles unconstitutional was not explicitly justified with neuroscientific findings, but many observers have expressed the belief that clinical data have had a significant role in it[3].

External issues, instead, are those involving the so-called *ius condendum*: the rewriting or radical reformulation of the main legal institutions based on the evidence provided by science, according to which such institutions and their underlying principles are no longer responsive to the known facts. Punishment belongs to this second category.

2. *The problem of free will*

A relevant line of naturalization of criminal law relies on the developments in neuroscience so as to try to prove that (if not always, at least most times) our actions are not free according to the classic definition of freedom – where the agent is capable of knowingly, voluntarily and *consciously* undertaking a course of action by choosing between alternatives. On the contrary, it is posited that our actions feature a high degree of determinism or at least of unconsciously undertaken courses of action, so that criminal conduct is regarded as deriving from the genetic asset of the subject, partly conjugated with an unfavourable environment. Other lines of research highlight that the structure and functioning of the brain strongly shape the subject's character traits (empathy in the first place) and can therefore direct or influence the behaviour of the individual in question[4].

More precisely, scepticism about free will is due to three main elements[5]. The first is the classical objection to freedom: determinism,

[3] D.L. Faigman-O.D. Jones-A.D. Wagner-M.E. Raichle, *Neuroscientists in Court*, in «Nature Reviews Neuroscience», 14 (2013), n. 10, pp. 730-736.

[4] A.R. Cashmore, *The Lucretian Swerve: The Biological Basis of Human Behavior and the Criminal Justice System*, in «Proceedings of the National Academy of Sciences», 107 (2010), n. 10, pp. 4499-4504; P.S. Churchland, *Braintrust: What Neuroscience Tells Us about Morality*, Princeton UP, Princeton (NJ) 2011.

[5] G.D. Caruso, *Introduction: Exploring the Illusion of Free Will and Moral Responsibility*, in G.D. Caruso (ed.), *Exploring the Illusion of Free Will and Moral Responsibility*, Lexington Books, Lanham (MD) 2013, pp. 1-16.

declined in several forms. The second, supported by the majority of philosophers of the mind, is the impossibility of mental causation, which is a condition for agency causation, a fundamental part of libertarian positions. The third is given by recent findings of cognitive science, indicating a progressive breakdown of the conscious self (some experiments seem to completely disconnect the latter from so-called "free" choices). In this regard, Nahmias underlines that this third strand is specifically interested in the progress of empirical psychology and cognitive neuroscience. In particular, he considers the first two strands as related to the form of mental causation, while the last is a thesis on the *content* of mental causation[6].

In cognitive science (including neuroscience) there is an ongoing process that is in line with this trend I have just described: the process of "deconstruction" of the conscious and rational unitary self – the subject of free will. Here one can distinguish two subsets. One concerns the beginning of the action: conscious intentions are preceded by subconscious cerebral processes[7]; the other concerns the conscious control of behavior, stating that consciousness is unaware of the automatic processes at work and the true reasons for our conduct[8]. The point is essentially that, under a more thorough empirical examination, more often than we would think, cognitive processing appears to be the result of subpersonal processes of which we are unaware.

These are automatic processes, triggered by the environment or the situation, bound to a repertoire that is partly innate and partly due to experience and education; such processes causes bodily responses due both to the tendency to homeostasis and to the search for what is functional to our survival and physical and mental well-being[9]. There are many examples of this decomposition of the self into cerebral modules that elaborate infor-

[6] E. Nahmias, *Is Free Will an Illusion? Confronting Challenges from the Modern Mind Sciences*, in W. Sinnott-Armstrong (ed.), *Moral Psychology. Vol. 4 Free Will and Moral Responsibility*, MIT Press, Cambridge (MA) 2014, pp. 1-25.

[7] In this respect, think of the very famous studies by Benjamin Libet: cf. B. Libet-C.A. Gleason-E.W. Wright-D.K. Pearl, *Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential): The Unconscious Initiation of a Freely Voluntary Act*, in «Brain», 106 (1983), n. 3, pp. 623-642; B. Libet, *Mind Time: The Temporal Factor in Consciousness*, Harvard UP, Cambridge (MA) 2004.

[8] Peter Carruthers is one of the most consistent supporters of this line of thought; P. Carruthers, *The Opacity of Mind: An Integrative Theory of Self-Knowledge*, Oxford UP, New York 2011; Id., *The Centered Mind: What the Science of Working Memory Shows Us About the Nature of Human Thought*, Oxford UP, New York 2015.

[9] See, for example, J.M. Doris, *Talking to Our Selves*, Oxford UP, New York 2015.

mation autonomously and subconsciously, which then emerge as a single apparent stream of consciousness. One case is that of language, where all the processes that lead us to say the words we speak are completely opaque to our consciousness[10].

Nevertheless, there is still wide consensus that neither recent experimental research through EEG and brain imaging, nor evidence coming from empirical psychology are enough to conclusively state that human beings have no free will[11]. Recent interpretations of the data collected by Libet even seem to bring back brain mechanisms of free will similar to our intuitive conception of it[12], which would also allow for a better understanding of it in terms of legal applications[13].

3. *Free will, law, and punishment*

One of the most discussed arguments regarding the notion of free will as an illusion and its consequences on the law is the one developed by Greene and Cohen[14]. According to their argument, a truly scientific description of the human being is incompatible with the attribution of *pure desert* in relation to the decisions made by the subject, on the basis of which the legitimacy (and effectiveness) of legal sanctions is determined. The proponents of this view maintain that one cannot but follow the logical

[10] T. Wilson, *Strangers to Ourselves: Discovering the Adaptive Unconscious*, MIT Press, Cambridge (MA) 2002.

[11] A.R. Mele, *Effective Intentions: The Power of Conscious Will*, Oxford UP, New York 2009; Id., *Free: Why Science Hasn't Disproved Free Will*, Oxford UP, New York 2014.

[12] A. Schurger-J.D. Sitt-S. Dehaene, *An Accumulator Model for Spontaneous Neural Activity Prior to Self-initiated Movement*, in «Proceedings of the National Academy of Sciences», 109 (2012), n. 42, pp. E2904-E2913; A. Schurger-M. Mylopoulos-D. Rosenthal, *Neural Antecedents of Spontaneous Voluntary Movement: a New Perspective*, in «Trends in Cognitive Sciences», 20 (2016), n. 2, pp. 77-79.

[13] A. Lavazza-S. Inglese, *Operationalizing and Measuring (a Kind of) Free Will (and Responsibility). Towards a New Framework for Psychology, Ethics and Law*, in «Rivista Internazionale di Filosofia e Psicologia», 6 (2015), n. 1, pp. 37-55; A. Lavazza, *Free Will and Neuroscience: From Explaining Freedom Away to New Ways of Operationalizing and Measuring It*, in «Frontiers in Human Neuroscience» 10 (2016) art. 262.

[14] J. Greene-J. Cohen, *For the Law, Neuroscience Changes Nothing and Everything*, in «Philosophical Transactions of the Royal Society of London B: Biological Sciences», 359 (2004), n. 1451, pp. 1775-1785. But see also R. Sapolsky, *The Frontal Cortex and the Criminal Justice System*, in S. Zeki-O.R. Goodenough (eds.), *Law and the Brain*, Oxford University Press, Oxford 2004, pp. 227-243; and S. Harris, *Free Will*, Simon and Schuster, New York 2012.

sequence deriving from the experimental data, which for them leads to the unavoidable pragmatic conclusion of choosing a consequentialistic kind of law and punishment.

According to Greene and Cohen, although Western penal systems claim to be compatibilist regarding free will, they actually seem to presuppose a libertarian perspective. But this view is now being threatened by the findings of neuroscience, which refers to a form of brain-related determinism. This kind of scientific data is in contrast with the widespread commonsense view of justice as well as with the retributivist conception of the law. Knowledge of the functioning of the brain points in the direction of denying the concept of free will in those who commit crimes, therefore leading to consequentialism: a view that – according to its promoters – is more in line with the scientific description of the human being.

Interestingly, the consequentialist perspective that relies on the idea of free will as an illusion also disrupts the limitations to the most undesirable consequences of the classical utilitarian perspective on punishment, which did not have arguments, for example, to exclude the use of scapegoats in some extreme cases. Among others, this "preventive" argument is supported by Hart[15]. Let's have a look at its logical path as it was retraced by De Caro and Marraffa[16].

In retributivism it is possible to identify two components, one called positive (all the guilty deserve to be punished with the required severity) and one called negative (only the guilty deserve to be punished, with no excessive severity). The second element prohibits to punish those who do not deserve it, and also has a preventive element against inhuman and disproportionate sentences. One could claim that the two components are logically independent, so that only one of them can be adopted. That is what Hart does, justifying punishment in purely utilitarian terms and using the negative component of retributivism as the "limit" to respect when attributing punishment, so as to avoid cases of blatant injustice. In other words, one can never punish an innocent, or someone who is causally but not morally responsible for a bad deed (say, because they are unfit to plead). This also holds when the punishment would have beneficial consequences for society as a whole.

[15] H.L.A. Hart, *Punishment and Responsibility: Essays in the Philosophy of Law*, Oxford UP, Oxford 1968.

[16] M. De Caro-M. Marraffa, *Mente e morale. Una piccola introduzione*, Luiss UP, Rome 2016, pp. 98-102.

For Hart, however, the justification of the sentence is not based on the merits of the offender (that is, when one punishes someone, it's not because there is an assumed balance of justice to be restored or because the act of the offender has to be punished as such). Hart refers to consequentialist model, according to which the only justification for punishment is that it is socially useful. Punishments thus serve: to create a deterrent so that, under the threat of punishment, people refrain from committing crimes; to make sure that dangerous people (because they have committed crimes) are put in a position not to further harm society; and to make criminals fit for social life through the execution of the sentence.

However, if you give up the chain that from the possibility to do otherwise – the primary condition of free will – leads to the idea of moral responsibility (understood as more than a contribution to the physical causal process of an event) then the negative retribution clause is no longer relevant. Therefore, there is no reason why classical utilitarianism shouldn't reappear at its purest, justifying the punishment of an innocent if it benefits the majority, as there are no principles against it other than merely conventional ones. A human being unable to act otherwise might be attributed some other form of dignity, but when it comes to punishment it is hard to see how the notion of "responsibility" can be replaced[17].

Indeed, as Greene and Cohen put it, given determinism in its various forms, consequentialism works in every case[18]. This is because this concept does not pose the problem of someone being truly innocent or guilty in some ultimate sense that depends on the freedom of action, but only addresses the issue of the likely effects of the punishment (although there is the problem of absolute determinism that does not seem to allow for the deterrent effect). The retributivist approach, on the contrary, seems to require the idea of free will, namely the ability to do otherwise, which is the classic condition for responsibility. If every action is the result of brain mechanisms outside of the possible conscious control of the subject, mechanisms visible through techniques capable of seeing in the "transparent bottleneck" of our nervous system, then it makes no sense to blame choices and actions on the subject who makes them. In a way, according to

[17] G. Sartori-A. Lavazza-L. Sammicheli, *Cervello, diritto e giustizia*, in A. Lavazza-G. Sartori (eds.), *Neuroetica. Scienze del cervello, filosofia e libero arbitrio*, il Mulino, Bologna 2011, pp. 135-163; A. Lavazza- L. Sammicheli, *Se non siamo liberi, possiamo essere puniti?*, in M. De Caro-A. Lavazza-G. Sartori (eds.), *Siamo davvero liberi?*, Codice, Torino 2010, pp. 147-166.

[18] J. Greene-J. Cohen, *op. cit.*

the supporters of free will as an illusion, those choices and actions are the result of an automatic response to social and environmental stimuli or internal impulses oriented to the preservation and to the physical and psychological wellbeing of the individual.

4. *The consequentialist view and its perils*

The outcome of this view is that punishment should be detached from any retributivist justification: it should not be afflictive in its main purpose, because this goes against the humanitarian principles of not harming our fellows without reason. In fact, the mere enlargement of the category of non-liability due to the discovery that many of those who commit violent crimes have serious brain abnormalities would lead to suspend or eliminate classical punishment in favour of other protective measures, such that would not be afflictive and would not have the sole purpose of punishing evil with more evil[19]. Along the same lines, the goal of rehabilitation of classical punishment would also cease to exist for people who are "mad and not bad", so to speak.

Derk Pereboom is perhaps the most important supporter of this thesis[20]. According to Pereboom, who is a hard incompatibilist[21], "living without free will" and, therefore, without responsibility, does not affect our ideas of morality, meaning and value of existence; so it's not something that produces the upheavals feared by defenders of free will. His "Spinozan" idea is that the main effects would be the end of a retributivist penal system (based on what the individual has done before) and the abolition of excessively severe punishments including, of course, the death penalty.

However, the adoption of a consequentialist perspective, inspired by crime prevention and social rehabilitation, would not exclude measures such as preventive detention. Pereboom considers the latter an instrument of social protection morally comparable to quarantine for people with highly contagious diseases. Just as the sick are not responsible for their condition,

[19] Cf. K. Kiehl, *The Psychopath Whisperer: Inside the Minds of Those Without a Conscience*, Oneworld, New York 2015.

[20] D. Pereboom, *Living without Free Will*, Cambridge UP, Cambridge 2001; Id., *Free Will, Agency, and Meaning in Life*, Oxford UP, New York 2014.

[21] Hard incompatibilists state that both determinism and indeterminism argue to the detriment of freedom, since in both cases the behavior of the subjects is caused by factors that are beyond their control.

and yet can be isolated for as long as necessary, guaranteeing them the best care and the highest personal dignity, so deemed dangerous subjects can be given the opportunity to do no harm without further afflictions, indeed, favoring their recovery. Pereboom's view is marked by the overcoming of "moral rage", which, for him, damages both individual wellbeing and interpersonal relationships due to the persistent tendency to blame and reprimand people, with the consequent creation of moral "debit" and "credit" able to poison one's life.

In a fully developed neurolaw, then, punishment would never be a substitute for a sort of social revenge, but rather the most humane tool available to control dangerous subjects and protect society, based on the medical and neuroscientific knowledge available. For sex offenders, for instance, it could be possible to act drastically with drugs that lower the hormone levels relevant to the behaviour in question (chemical castration); for impulsive and violent subject, drugs that control one's mood would be appropriate. In other cases, brain pacemakers, in the form of brain stimulators, may act by reducing certain compulsive behaviours (such as taking drugs that lead to other crimes), and so forth.

An approach of this kind would be welcomed both by society and by the very individuals subjected to it, because it would be selective and would not completely deprive them of their physical freedom (or of their life, where death penalty is in force). Nevertheless, this approach has the characteristic of potentially slipping into (1) the invasive violation of privacy and bodily integrity (the right that protects against intentional interference with one's body) on the basis of available technology; (2) preventive treatment or detention; (3) treatment or detention without a specific goal. This could happen if punishment were no longer anchored to the classical mechanism of personal responsibility in the retributive sense, for which one is punished for what one has done in accordance with a law that pre-established the punishment according to the seriousness of the crime as such. But the consequentialist perspective tends to radically break away from that model.

Let's see in more detail the three forms of punishment implied by a consequentialism that denies any retributivist element. These are forms of punishment that conflict with strong moral intuitions and violate ethical principles that seem to be a shared heritage for the defense of the individual and her autonomy. Preventive treatment or detention could be put in place for those that, on the basis of neuroscientific markers and other behavioral data, are expected to have high chances of committing violent crimes. To this end, mass screening would become mandatory from birth, and this could open the door to discrimination and strong personal autonomy limitations. If the subject shows predictive markers of serious antisocial behavior, according to the consequentialist perspective, he should be put in a condition to do no harm, as the focus is entirely on the protection of the community at the expense of the single potential criminal. The subject is indeed denied many rights in the implicit assumption that he is a "sick" person, who should be treated as for her well-being but deprived of physical freedom and self-determination, in order to protect society.

From chemical castration to genetic engineering, all systems of care to improve deviant behavior would become lawful. Furthermore, such care or indefinite periods of detention would not have a clear goal, since at least for some individuals the point would be to prevent the general threat to society that they could potentially represent – contrary to the retributivist system, there would be no need to wait for the threat to be actualized. Ultimately, the availability and justifiability of this new kind of punishment might lead to apply it to *all* those individuals who have been identified as highly likely to commit certain crimes: one might want to coercively treat with drugs both an exhibitionist and a rapist. Secondly, whenever a technique promised to be efficient in detecting or preventing criminal conduct, it would be justified to introduce it and enforce it on potentially interested parties. Thirdly, in some cases it is unclear how to assess the decreased dangerousness of subjects under coercive treatment, so that the treatment could be extended indefinitely.

It is useful to recall here the position expressed by Thomas Douglas. He has persuasively argued for criminal rehabilitation through medical interventions (such as medications that replace the drug of addiction for drug-addicted offenders, and injections of testosterone-lowering drugs for sex offenders) claiming that committing a crime can render one morally liable to certain forms of medical intervention[22]. Douglas challenges the shared assumption that medical interventions may only permissibly be administered to criminal offenders with their consent. The argument starts from the fact that it is commonly accepted that the State can impose a punishment without consent to those who commit crimes, typically a period of detention. For Douglas, if one accepts that offenders are morally liable, imposing limitations on freedom of movement and association (with all

[22] T. Douglas, *Criminal Rehabilitation Through Medical Intervention: Moral Liability and the Right to Bodily Integrity*, in «The Journal of Ethics», 18 (2014), n. 2, pp. 101-122.

that this implies) does not produce more harm than a violation of bodily integrity, when it is oriented to the rehabilitation of the offender.

Nevertheless, there is still the problem of direct brain interventions: contrary to the lowering of testosterone to temporarily reduce sexual desire, such interventions interfere with the very basis of agency and the self. As Jared Craig rightly put it, there is a "more fundamental right to 'mental integrity'" that defends an alleged inner sphere of liberty and protects critical capacities necessary for the exercise of autonomous human agency – without which a vast majority of moral rights could not exist[23]. In this sense, the State should not be entitled to administer direct brain interventions to criminal offenders without a valid consent.

Here's an example of consequentialist scenario sketched by Adrian Raine:

> Under LOMBROSO [program – Legal Offensive on Murder: Brain Research Operation for the Screening of the Offenders], all males in society aged eighteen and over have to register at their local hospital for a quick brain scan and DNA testing. [...] The result is not a perfect prediction, but it is pretty darn good [...] Those classified as LP-S (Lombroso Positive-Sex) have an 82 percent chance of committing either rape or pedophilic offenses. [...] The program works like this: those who test positive – the LP-S – are held in indefinite detention[24].

But the program can be expanded.

> Poor parenting has undeniably been linked to later violence. Genetic studies documented not just that antisocial parents transmit their bad genes to their children, but that negative social experience of having a bad parent is also a causal factor for antisocial behaviour. [...] Cars can be killers, and so you need a licence before you can drive. Kids can be killers too. So the logic goes that you should also have a licence before you can have a child[25].

Then, even something with a scientific justification and a related humanitarian goal could dangerously turn into an instrument of tyranny and discrimination, because the scientific knowledge in this area is not yet well

[23] J.N. Craig, *Incarceration, Direct Brain Intervention, and the Right to Mental Integrity. A Reply to Thomas Douglas*, in «Neuroethics», 9 (2016), n. 2, pp. 107-118; cf. also J.C. Bublitz-R. Merkel, *Crimes Against Minds: On Mental Manipulations, Harms and a Human Right to Mental Self-determination*, in «Criminal Law and Philosophy», 8 (2014), n. 1, pp. 51-77.

[24] A. Raine, *The Anatomy of Violence: The Biological Roots of Crime*, Pantheon, New York 2013, pp. 342-343.

[25] Raine, *op. cit.*, p. 349.

substantiated. Also, the laws would be adopted by political decision-makers without all the necessary technical knowledge, and judgments would be made not only by experts but also by judges guided by considerations other than the simple medical history of the defendant. Indeed, the very fact that the new type of punishment has an allegedly more humane character would end up lowering the public attention to potential miscarriages of justice.

5. *The Strawsonian view*

Another set of (more philosophical) considerations appeals to a perspective proposed by Peter Strawson, according to whom the naturalistic-consequentialistic approach treats the human beings subject to the new type of punishment as broken machines rather than as agents to be respected and considered worthy of dignity[26]. In *Freedom and Resentment*, Strawson considers "the non-detached attitudes and reactions of people directly involved in transactions with each other", or else "the attitudes and reactions of offended parties and beneficiaries; of such things as gratitude, resentment, forgiveness, love, and hurt feelings"[27]. In our interactions with our fellow human beings, we all have reactive attitudes and feelings, which we ourselves are subject to. They have an extraordinary importance for us and depend on what we think about the feelings and attitudes of others.

Resentment towards people who deliberately harm us is not philosophically problematic; however, there are two factors that could affect that feeling if those who harmed us did so under particular circumstances. The first one is related to unintentionality: "He didn't mean to", "He hadn't realized", "He was pushed". In such cases we might curb our resentment but still feel that it's appropriate to have a reactive response. The second one is related to cases when the person responsible "wasn't himself", "has been under very great strain recently", or even "is a hopeless schizophrenic", "his mind has been systematically perverted". For Strawson, such cases lead us to restrain from having our normal reaction towards the agent. Hence a contraposition between participation/involvement in a human relationship and what could be called an "objective attitude" towards other humans.

[26] P.F. Strawson, *Freedom and Resentment*, in «Proceedings of the British Academy», 48 (1962), pp. 1-25.

[27] *Ivi*, pp. 82-83. I am here taking up an exposition found in A. Lavazza-L. Sammicheli, *op. cit.*, cap. 8.

> To adopt the objective attitude to another human being is to see him, perhaps, as an object of social policy; as a subject for what, in a wide range of sense, might be called treatment; as something certainly to be taken account, perhaps precautionary account, of; to be managed or handled or cured or trained [...] The objective attitude may be emotionally toned in many ways, but not in all ways: it may include repulsion or fear, it may include pity or even love, though not all kinds of love. But it cannot include the range of reactive feelings and attitudes which belong to involvement or participation with others in inter-personal human relationships[28].

That is, if one adopts the objective attitude towards someone, feelings such as resentment, gratitude, forgiveness, anger or romantic love cannot arise. One can talk with that person, but not argue. In other words, we would say, one does not perceive them as able to respond to reason. If one accepts determinism (incompatibilism), then, should one give up the reactive attitudes? The answer is that this would be impossible, because of the very way we are made: the involvement with which human beings participate in common interpersonal relationships is too intense and runs too deep to seriously believe that a general theoretical conviction might genuinely change our world – including interpersonal relationships as we normally understand them[29].

But one may ask: what would be the rational choice, if freedom were really illusory? According to Strawson, firstly, we are *naturally* led to reactive attitudes, we cannot choose whether or not to adopt them in the way that we can, for example, accommodate or not some preferences; secondly, and most importantly, even if we had a choice, the rational option would be to evaluate gains and losses for human life, considering what enriches or impoverishes it; the truth or falsity of a general thesis related to determinism would not have any relation to the rationality of this choice. Personal reactive attitudes are based on an expectation and a need: that human beings around us show a certain degree of goodwill or regard towards us; or, at least, that they show no active manifestation of malice or indifference. It follows that it is simply *useless* to ask whether or not it would be rational for us to actualize something that by virtue of our own nature we cannot

[28] Strawson, *op. cit.*, pp. 89-90.

[29] B. Vilhauer (*The People Problem*, in G.D. Caruso (ed.), *Exploring the Illusion of Free Will and Moral Responsibility*, Lexington Books, Lanham (MD) 2013, pp. 141-160) believes that one can overcome Strawson's argument on the depersonalization of human beings by referring to the Kantian principle that prescribes to treat all our fellow beings always as ends and never as means. This principle can be declined without the use of reactive attitudes and attributions of freedom and moral responsibility.

(be able to) do. The general network of personal reactive attitudes was in fact created along with human society and, considered as a whole, does not need any external *rational* justification.

As is obvious, even regardless of the facts in favor of the deterministic thesis, common sense and scientific optimism given by the illusionary character of free will clash with the fact that one wants to choose on the basis of practical consequences of her decision. Finally, from their point of view, the skeptics who are optimistic on free will are those who confirm that it is possible to have a humanitarian "objective attitude" oriented to the welfare of others. But from the point of view of the "optimistic skeptics" this is contradictory to what previously stated on consequentialism, unless one introduces the purpose of respecting certain values that are themselves disjointed from consequentialism itself.

6. *A defence of moderate retributivism*

Given the perils coming from a purely consequentialistic perspective denying free will and responsibility, I briefly propose three arguments to defend a moderate conventionalist thesis on the classical responsibility of retributivist law. I think it is appropriate to maintain (by stipulation, according to the liberal-democratic processes that form and gradually change the legal system) a system that recognizes – mostly and with encoded exception types – freedom, rationality and the ability to answer for one's actions. Such a system also includes punishments directly related to the voluntary transgression of the norms, although also aimed at the recovery of the offender and the protection of society.

The first argument is related to naturalism, the very frame in which refoundational prospects are inscribed. The evolutionary processes of the species, driven by selection and adaptation, have endowed us with very strong intuitions – generally retributive – that cause people to be ready to bear a personal cost, with no other gain than the restoration of a sense of justice, to punish offenders who deserve it[30]. It does not seem easy to reverse this intuition with a rationalist argument, especially as one looks at the same time to found morality on the feelings of "natural sympathy" that

[30] E. Fehr-S. Gächter, *Altruistic Punishment in Humans*, in «Nature», 415 (2002), n. 6868, pp. 137-140; E. Fehr-U. Fischbacher, *Third Party Punishment and Social Norms*, in «Evolution and Human Behavior», 25 (2004), n. 2, pp. 63-87; cf. also A. Lavazza-L. Sammicheli, *op. cit.*, cap. 7.

are probably the result of evolution[31], thus replacing ethical systems *a la* Kant[32].

The second (indirect) argument is related to experiments aimed to falsify common sense and naive psychology, which provide the basis to naturalistically falsify retributivism and give arguments in favor of consequentialism denying the intuitions of freedom and responsibility. The impression is that these experiments are, so to speak, "below threshold" with respect to relevant social macro-interactions for the dynamics of allocation of responsibility and the functioning of relevant interpersonal relations. In this sense, the relationship between the description of the subpersonal mental mechanisms and intentionalist psychology could be seen in analogy with what happens in physics between relativistic mechanics and classical mechanics. Relativistic mechanics is certainly more appropriate to the current knowledge and allows for a "true" and finer description of reality, but the most intuitive and usual description offered by Newton's classical mechanics is perfectly adequate for most of the macroscopic applications that usually concern us. As for the description of human beings there is also a subjective element, which seems to prefer – for now, but for many reasons – the use in certain areas of folk psychology. Also, one could say that what allows empirical psychology to describe the disunity of the subject and the alleged behavioural automaticness is a "quantification" that covers a narrow area of our spectrum of social action. The significant interactions subject to the law might fall within the macroscopic range of relevant values in which behaviour tends to be free, aware and rational – that is, coherent with the assumptions of retributivism.

[31] S. Nichols, *Sentimental Rules: On the Natural Foundations of Moral Judgment*, Oxford UP, New York 2004; E. Lecaldano, *Prima lezione di filosofia morale*, Laterza, Roma-Bari 2011; Id., *Simpatia*, Cortina, Milano 2013.

[32] Robert Nozick presents a theoretically refined version of common sense: "In terms of the connection with value effected by punishment we can understand some of the metaphors that stud retributivist talk. Wrong puts thing out of joint in that acts and persons are unlinked with correct values; this is the disharmony introduced by wrongdoing. Punishment does not wipe out the wrong, the past is not changed, but the disconnection with the value is repaired (though in a second best way); nonlinkage is eradicated. Also, the penalty wipes out or attenuates the wrongdoer's link with incorrect values, so that he now regrets having followed them or at least is less pleased that he did" (R. Nozick, *Philosophical Explanations*, Harvard UP, Cambridge (MA) 1981, p. 379). Retributivism, just as the consequentialism evidenced by Hart, needs external principles to define its scope. Nozick himself recognizes this implicitly when he asks why we should not always relate to the value, even for those who do not commit crimes. The answer is that in that case what prevails is the individual's right to be left alone.

The third argument relies on a distinction that has been used to show how psychopaths could be considered exempt from the law (also ethically) by virtue of the fact that they fail to grasp the strength of genuinely moral prescriptions, therefore lacking the ability to understand the scope and consequences of their actions[33]. This concerns the partition between conventional norms (you have to sit straight at school) and moral norms (you mustn't pull your classmate's hair), which also small children are able to grasp. Now, if this distinction is based on some foundation, related to a specific ability of recognition, this seems to imply some form of moral realism. Not a realism that presupposes an autonomous existence of values that people can grasp with a special sense, but more likely a pre-reflective intuition shared by almost all human beings on the evaluation of a series of behaviors as a positive or negative (to do or not to do).

If these insights may serve as a point of reference and constitute a reason for exemption for those who, due to a "natural" defect, do not have them, then they must have a "validity" that enables them to act as a reference for "moral" behaviors. Those without this ability cannot be held responsible; conversely, those who do, though, when not respecting these rules, having the ability to understand them and to respect them, are exposed to reprimand and punishment. One could say that being able to grasp moral norms does not in itself amount to being able to respect them. However, in the psychopath argument, it must obviously be so, otherwise it would not make sense to use it so as to separate her position from that of other individuals. If there wasn't at least someone able to grasp moral norms and respect them, it would not make sense to declare that others (psychopaths) are instead exempt from them. If no one has the ability to respect the rules, then all, without distinction, should be declared exempt.

7. *Conclusion*

In a Kantian-Hegelian sense, punishment amounts to recognizing the freedom of action of the other, who more or less voluntarily has broken the law. The "bad person" is like us and can be "rehabilitated" with equal treatment. But quarantine, with the physical removal of the "bad" from

[33] E. Turiel, *The Development of Social Knowledge: Morality and Convention*, Cambridge UP, Cambridge 1983; R.J.R. Blair-D.G.V. Mitchell-K. Blair, *The Psychopath: Emotion and the Brain*, Blackwell, Oxford 2005.

society, is also a metaphor for the expulsion of the "sick". A "sick person" (socially non-integrated) that cannot be cured is a substantially different subject that can be legitimately treated as such.

In fact, Kant and Hegel have defended the retributivist principle, regardless of its roots in free will, as an instrument of protection of human dignity, which recognizes rational agency as constitutive of the person. In this view, the sign of autonomy – denied by the idea of criminals as sick, for whom the only option is extrinsic care – lies precisely in the ability of moral redemption through punishment. This does not mean that we should oppose the prospect of a neuroscientific punishment on a consequentialist basis, but rather that we should assess the risks and benefits of such an approach in the light of the full spectrum of punishments, their goals and their justifications. Mostly, we shouldn't fail to consider some principles that appear important for the dignity and autonomy of every human being as a subject endowed with intrinsic value.

Abstract

Neurolaw is the approach that attempts to apply recent progress in neuroscience to the classical conceptions of law, often with the aim of pushing legal institutions (especially in criminal law) to be more in line with scientific knowledge. It is essentially a process of naturalization of the law, which also applies to punishment, its aims, its methods of implementation and its justification.

A relevant line of naturalization of criminal law relies on developments in neuroscience so as to try to prove that (if not always, at least most times) our actions are not *free according to the classic definition of freedom – where the agent is capable of knowingly, voluntarily and* consciously *undertaking a course of action by choosing between alternatives. According to the proponents of this view, one cannot but follow the logical sequence deriving from the experimental data, which leads to the unavoidable pragmatic conclusion of choosing a consequentialistic kind of law and punishment.*

Consequentialist punishment is deemed to be more humane because it is not afflictive and is only targeted to protect society. But the fact that the charged person is regarded as more mad than bad, so to speak, turns her into a sort of "broken machine", with the risk of legitimizing preventive treatments or ones of indefinite duration. The objections to this approach are therefore related to the gaps of knowledge we still have, to the risks of "political" *abuse, and to the Strawsonian line of thought for which we cannot treat our fellow human beings as broken machines to be repaired, depriving them of their nature of free and rational agents (except in exceptional and rare cases). I suggest a more nuanced assessment of these possible developments and defend a moderate form of retributivism.*

Keywords: free will; consequentialism; retributivism; naturalism; self; P. Strawson.

Andrea Lavazza
Centro Universitario Internazionale -Arezzo
lavazza67@gmail.com

On the Automaticity and Ethics of Belief

Uwe Peters

It is widely accepted that we are able to think about or entertain propositions without believing them[1]. However, some philosophers have employed cognitive scientific findings to argue that this view is in fact false[2].

For instance, Millikan holds that there are psychological studies providing «evidence that when we hear someone speak, normally what is said goes directly into belief [...]. We do not first understand what is said and then evaluate whether to believe it»[3] but rather immediately accept[4] the information that we are presented with. Similarly, on the basis of empirical research, Mandelbaum maintains that merely «thinking» that *p* «is believing» that *p*[5].

[1] E.g., R. Descartes, *Meditations on First Philosophy* (1641), Hackett, Indianapolis 1984; J. Fodor, *The Modularity of Mind*, MIT Press, Cambridge (MA) 1983; D.C. Dennett, *Brainstorms*, MIT Press, Cambridge (MA) 1981; J. McDowell, *Having the world in view: Sellars, Kant and Intentionality. Lecture 1: Sellars on perceptual experience*, in «Journal of Philosophy», 95 (1998), pp. 431-450; L. O'Brien, *Self-Knowing Agents*, Oxford UP, Oxford 2007; C. McHugh, *Judging as a Non-Voluntary Action*, in «Philosophical Studies», 152 (2011), pp. 245-269; U. Kriegel, *Entertaining as a Propositional Attitude: A Non-Reductive Characterization*, in «American Philosophical Quarterly», 50 (2013), pp. 1-22.

[2] E.g., R. Millikan, *Varieties of Meaning*, MIT Press Cambridge (MA) 2004; E. Mandelbaum, *Thinking is believing*, in «Inquiry», 57 (2014), n. 1, pp. 55-96; N. Levy-E. Mandelbaum, *The powers that bind: Doxastic voluntarism and epistemic obligation*, in J. Matheson (ed.), *The Ethics of Belief*, Oxford UP, Oxford 2014, pp. 12-33.

[3] R. Millikan, *op. cit.*, p. 121.

[4] There are differences between accepting or affirming that *p* and believing that *p* (e.g., one may for the sake of the argument accept *p* without believing *p*). However, for the purpose of this paper, I shall ignore this point and follow the Spinozans in treating accepting or affirming that *p* as initiations of believing that *p* (e.g., E. Mandelbaum, *op. cit.*, p. 61), or at any rate as leading to a doxastic state that is belief-like (N. Levy-E. Mandelbaum, *op. cit.*, p. 26).

[5] E. Mandelbaum, *op. cit.*, p. 55.

The claim is that whenever we entertain a proposition *p*, we will automatically and prior to analyzing the truth of *p* come to believe *p* at the unconscious level. Upon subsequent reflection at the conscious level we may reject *p* or endorse[6] *p* but that is only possible after we have initially accepted it at the unconscious level. This view of belief formation is often called the *Spinozan theory*, as Spinoza is thought to be the first who defended it[7].

Some empirically oriented philosophers who advocate the Spinozan theory hold that the theory has implications for the ethics of belief. For instance, after arguing for the Spinozan theory, Levy and Mandelbaum continue that people «who know about» their «propensities» to believe propositions through merely entertaining them have epistemic «obligations to take the risk of forming unjustified» and «immoral beliefs into account» when they expose themselves to them[8].

In the following, I use theoretical considerations and data from psychological studies to cast doubts on the empirical case for the view that we automatically believe the propositions that we entertain. In addition, I maintain that even if we set these doubts aside, Levy and Mandelbaum's argument to the effect that the automaticity of believing creates epistemic obligations remains unconvincing.

1. *The Spinozan theory*

The most developed form of the Spinozan theory, which is also the version that I will focus on, has been introduced by Gilbert and his colleagues and elaborated by Mandelbaum[9]. It can be summarised in the following three claims[10].

[6] Spinozans use the terms "to endorse *p*" (e.g., E. Mandelbaum, *op. cit.*) or "to certify *p*" (D. Gilbert, *How mental systems believe*, in «American Psychologist», 46 (1991), n. 2, pp. 107-119) to distinguish conscious, subject-controlled affirmations of *p* from unconscious, automatic affirmations of *p*, for which they tend to use the term "to accept *p*".

[7] B. Spinoza, *The Ethics and Selected Letters* (1677), Hackett, Indianapolis (IN) 1982. However, Spinoza arguably did not use the conscious vs. unconscious distinction that contemporary Spinozans invoke in their account of belief formation.

[8] N. Levy-E. Mandelbaum, *op. cit.*, p. 30.

[9] D. Gilbert, *op. cit.*; D. Gilbert-D. Krull-P. Malone, *Unbelieving the unbelievable: Some problems in the rejection of false information*, in «Journal of Personality and Social Psychology», 59 (1990), pp. 601-613; Idd., *You can't not believe everything you read*, in «Journal of Personality and Social Psychology», 65 (1993), pp. 221-233; E. Mandelbaum, *op. cit.*

[10] Mandelbaum (*op. cit.*) adopts a fourth, negation-related claim that I shall not consider here.

1) People do not have the ability to contemplate propositions that arise in the mind [...] before believing them. Because of our mental architecture, it is (nomologically) impossible for one to not immediately believe propositions that one tokens.
2) Accepting a proposition is accomplished by a different system than rejecting a proposition. Because different systems are at play, the processes of accepting and rejecting should be affected by performance constraints in different ways. [...]
3) Forming a belief is a passive endeavour[11]. However, rejecting a proposition is an active and effortful mental action, which can only happen after a belief has been acquired. Consequently, one can effortlessly form new beliefs while being mentally taxed, but rejecting an already held belief will become more difficult the more mentally taxed one is[12].

Based on these claims, the Spinozan theory yields a number of predictions. For instance, when a subject is presented with a proposition *p* and prevented from rejecting *p* (e.g., by being distracted), she should not remain doxastically neutral about *p* but end up believing the proposition. Furthermore, since, according to the Spinozan theory, it is «(nomologically) impossible for one to not immediately believe propositions that one tokens»[13], subjects should have the tendency to accept *p* even when they are *before* they are presented with the proposition told that *p* is false.

Spinozans have appealed to empirical studies to argue that these predictions are borne out by the data. I'll briefly consider some central examples.

[11] It is worth noting that even though advocates of the Spinozan theory claim things such as (i) in general «[f]orming a belief is a passive endeavour» (E. Mandelbaum, *op. cit.*, p. 62) or (ii) «we, strictly speaking, do not form beliefs for reasons at all» (N. Levy-E. Mandelbaum, *op. cit.*, p. 17), these claims are – even on the Spinozan view itself – not correct. For, according to the Spinozan theory, one can, once one has automatically accepted *p,* still reject, or endorse *p* at the conscious level, and this will at least sometimes happen for epistemic reasons pertaining to whether *p* is true. Furthermore, the subsequent rejection or endorsement of *p* will result in a belief itself, namely in the belief that *p* is false or true, respectively. Since rejecting and endorsing *p* are on the Spinozan account "active" (D. Gilbert, *op. cit.*, p. 108; Gilbert *et al.*, *You can't not believe everything you read*, cit., p. 4; E. Mandelbaum, *op. cit.*, p. 61), it follows that some cases of belief-formation are, against claims (i) and (ii), even according to the Spinozan theory active in nature and based on reasons (e.g., cases of rejection-, or endorsement-based belief-formation).

[12] E. Mandelbaum, *op. cit.*, p. 61.

[13] *Ibidem.*

2. *Evidence supporting the Spinozan theory*

Among the psychological work that Spinozans heavily rely on are the following two experiments conducted by Gilbert and his colleagues[14].

Gilbert *et al.* asked subjects to learn statements about the meaning of words in a fictional language, for instance, "A *monisha* is a star"[15]. Each statement appeared briefly on a screen and was followed by a validating term, i.e., "true" or "false". On some trials, during the learning phase, subjects had to identify musical tones that rang out after the validating word appeared. This was meant to drain their mental resources. In the testing phase, subjects were then again shown the sentences and asked whether they were true, false, or not present during the learning phase.

Gilbert *et al.* found that participants who were distracted during the validation process by the tone-identification task didn't manage to remain doxastically neutral about the statements presented to them but tended to encode the sentences, including those marked as false, as true. Gilbert *et al.* and others Spinozans take this to show that we «first believe what is said and then, if we are not under too much cognitive stress, we may think it over critically and reject it»[16].

Gilbert *et al.* conducted another study that led to similar findings[17]. Subjects were asked to read two crime reports that included both true and false statements. True information was shown in black, false information in red. One report contained false sentences that increased the severity of the crime, and the other included false sentences that diminished it. Some test participants were asked to do a concurrent digit-search task as they read the false sentences in the reports. This was meant to impose cognitive load. Afterwards, participants were asked what prison sentence (0-10 years) they would give for the crimes that they had read about on the first line and how they evaluate the criminal's character, for instance, how much they liked him, how dangerous he was, and how much counseling would help him.

It turned out that when the text contained exacerbating information that was false, subjects in the load condition, but not those in the no-load condition, recommended harsher sentences than when mitigating information

[14] See, e.g., Millikan 2004; E. Mandelbaum, *op. cit.*; N. Levy-E. Mandelbaum, *op. cit.*

[15] Gilbert *et al.*, *Unbelieving the unbelievable*, cit.

[16] R. Millikan, *op. cit.*, p. 121.

[17] Gilbert *et al.*, *You can't not believe everything you read*, cit.

was false. Furthermore, these participants' ratings of the perpetrator's dislikableness, dangerousness, and likelihood of benefitting from counseling were higher than those of the no-load subjects.

Gilbert *et al.* and other Spinozans argue that since the subjects under load acted on the false information just as if they believed it, they did indeed believe it[18]. Since they seemed unable to suspend acceptance of the information, the findings suggest that subjects automatically believe the propositions they entertain, or so the Spinozans claim.

In fact, they maintain that this will be the case even if subjects are *before* encountering the propositions told that the propositions are false. To support this, they cite a study by Wegner *et al.*[19] in which participants were shown pairs of suicide notes and told that one note from each pair was real and the other fake[20]. The subjects' task was to sort the real ones from the fakes. After each decision, they were given feedback on their performance. Crucially, before the trial started, they were informed that the feedback they would receive was false. After the test, subjects were asked to estimate how often they answered correctly.

Surprisingly, their answers still matched the feedback. Levy and Mandelbaum write that the «knowledge of the feedback persists because the participants automatically affirm the feedback when they hear it, even though they know the feedback is false. Since they are engaged in a relatively fast-paced experiment, the participants lack the mental energy to override the false claims»[21].

On the basis of the data just introduced (and other findings), Spinozans claim that «when we hear someone speak [and think about what they are saying], normally what is said goes directly into belief»[22], that «thinking is believing»[23], and that «we are designed to initially affirm any propositions that we happen to think about»[24].

In the next three sections, I'll motivate some doubts about these claims. I begin with a theoretical consideration.

[18] E.g., E. Mandelbaum, *op. cit.*; N. Levy-E. Mandelbaum, *op. cit.*

[19] D. Wegner-G. Coulton-R. Wenzloff, *The Transparency of Denial: Briefing in the Debriefing Paradigm*, in «Journal of Personality and Social Psychology», 49 (1985), n. 2, pp. 338-346.

[20] See D. Gilbert, *op. cit.*, p. 114; E. Mandelbaum, *op. cit.*, p. 67; N. Levy-E. Mandelbaum, *op. cit.*, p.26

[21] N. Levy-E. Mandelbaum, *op. cit.*, p. 26.

[22] R. Millikan, *op. cit.*, p. 121.

[23] E. Mandelbaum, *op. cit.*, p. 55.

[24] N. Levy-E. Mandelbaum, *op. cit.*, p. 26.

3. *What the data doesn't show*

All the just-mentioned studies, which play a pivotal role in the Spinozan argument, involve subjects that are under cognitive load or «lack mental energy»[25]. The involvement of cognitive load or a depletion of mental energy is important for the Spinozan because the empirical case for the Spinozan theory rests on what Gilbert calls a «general principle of systems break-down: *When stressed, modular information-passing systems with multiple exit capabilities will often show a bias toward prematurely outputting the products of early modules*»[26].

Since cognitive load 'stresses' the modular information-passing systems involved in the comprehension of a proposition *p*, it should lead them to prematurely output the product of the module that processed *p* before the load occurred. The Spinozan then predicts that if one automatically accepted *p* when one is entertaining *p*, imposing cognitive load during the validation phase should induce the system to prematurely issue the product of the earlier processor, i.e., an acceptance of *p*. The findings do support the prediction.

However, strictly speaking, they are compatible with the view that before cognitive load is imposed, the module processing *p* remains doxastically neutral about *p* and only at the moment when the load crosses a certain threshold opts for an acceptance of *p*. On this view, the acceptance of *p* that subjects exhibit in the mentioned studies is not the output of the module processing *p* before the validation, which is done by a different module. Rather, there is just one module responsible for both comprehension and validation and this module operates in stress conditions differently than in no-stress conditions: if subjects are during validation of *p* put under load, they will not remain doxastically neutral but accept *p*.

This does undermine the claim that we can always remain doxastically neutral when we are considering a proposition *p*, for sometimes we are considering *p* under load. But it does not show that when subjects are presented with *p* and not put under load, they will initially automatically believe the proposition. For all that the above studies tell us, when we are not under cognitive load or do not lack mental energy, we do not believe what we are thinking about until or our mental energy is depleted. To retain the strong view that it is «(nomologically) impossible for one to not

[25] *Ibidem.*

[26] D. Gilbert, *op. cit.*, p. 109 (italics original).

immediately believe propositions that one tokens», as Mandelbaum claims, this possibility needs to be addressed and refuted[27].

4. *Automatic rejections*

There is reason to hold that even when we *are* under cognitive load, we don't always initially automatically believe what we think about. According to the Spinozan theory, we automatically believe *any* proposition we entertain, yet, it is worth noting that unlike in, for instance, Gilbert *at al.*'s studies, in everyday life, people often have some knowledge available to draw on when they are confronted with a piece of information. Given this, suppose that we have strong background beliefs about a proposition *p* and these beliefs contradict *p*. Do we still initially automatically accept *p* when we entertain it?

Richter *et al.* conducted a version of Gilbert et al.'s experiment that did not use nonsensical statements such as "A *monisha* is a star" but objectively true or false assertions about which subjects could be expected to have either strong or weak validity-related background beliefs[28]. They found that for statements with strong background beliefs (true or false),

[27] E. Mandelbaum, *op. cit.*, p. 61. Gilbert (*op. cit.*, pp. 114f) considers the proposal that one might understand a proposition without representing it as true (114). In response, he cites Wegener et al.'s above-mentioned study in which subjects didn't refrain from accepting false propositions even though they were told about their falsity beforehand. Gilbert holds that «subjects were unable to represent the statements in a truth-neutral fashion, even when directly motivated to do so» (*op. cit.*, p. 115). However, this is unconvincing, as it might be that participants simply forgot to bring the relevant information on the falsity of the experimenter's statements (about their performance in identifying suicide notes) to bear on the issue, and took the experimenter to be a reliable source. Also, perhaps the test participants failed to resist acceptance because they are in the study «engaged in a relatively fast-paced experiment», and hence «lack the mental energy to override the false claims» (N. Levy-E. Mandelbaum, *op. cit.*, p. 26). The findings then no longer undermine the proposal that when one's mental energy is *not* depleted, subjects can think about propositions without initially believing them. Gilbert (*op. cit.*) offers another point in support of the claim that comprehension and acceptance always fall together. He reports a study in which he and his colleagues asked subject to simply read out sentences on an imaginary creature without assessing the statements. Yet, when later on asked about the veracity of the statements, subjects took them to be true. However, as Gilbert writes himself, subjects were asked to read quickly, and there was a premium on fast readers. Hence, test subjects were under time pressure. Since time pressure reduces mental energy, the findings again don't undermine the proposal that if subjects are not mentally taxed, they can think about a proposition without initially accepting it.

[28] T. Richter-S. Schroeder-B. Wohrmann, *You don't have to believe everything you read: Background knowledge permits fast and efficient validation of information*, in «Journal of Personality and Social Psychology», 96 (2009), pp. 538-558.

say, “Soft soap is edible”, cognitive load during learning did not result in people’s accepting false propositions. That is, when subjects were, after the learning phase, in the test phase asked, for instance, “Is soft soap edible?”, they didn’t show evidence of an acceptance of the proposition.

Could it be that subjects simply accessed their stored strong background belief that soft soap is *in*edible to answer the question, and thus showed unimpaired accuracy even though the on-line effortful rejection process was disrupted and an initial automatic acceptance in the learning phase occurred? Richter *et al.* used two different measures to rule this out.

First, they included new assertions in the verification task in addition to those that subjects had been presented with in the learning phase. By comparing the error rate and response latency for new assertions and assertions presented in the learning phase, Richter et al. could delineate effects of validation processes in the learning phase and separate these effects from belief effects that occurred in the test phase.

Second, they asked subjects to make their verification judgments within a specified time frame that varied in length. The thought was that if background beliefs come in during resource-dependent validation processes in the verification task, the verification of assertions linked with strong background beliefs should be negatively affected by a shorter response timeframe. This did in fact happen with *new* (strong background belief-related) assertions. But crucially, if subjects verified assertions that were linked to strong background beliefs *and* shown in the learning phase, the decline from the long to the short response-time frame was only moderate. This suggests that the validation of the assertions already occurred under load in the learning phase, and that subjects were able to automatically reject what they thought about, which is at odds with the Spinozan theory.

Richter *et al.* conducted a second study that also speaks against the theory. Participants were very briefly (300-600ms; see experiments 3 and 4) presented with three words (one-by-one), which formed an assertion that was either consistent or inconsistent with their background beliefs. In the critical trials, the participants’ task was to quickly assess the correct spelling of the third word while it was presented to them.

Subjects committed fewer mistakes and needed less time to respond when words within true sentences (i.e., sentences that were in line with their background beliefs) were grammatically correct and when words within false sentences (i.e., sentences that were at odds with their background beliefs) were grammatically incorrect than in the two incongruent conditions (i.e., correct grammar/with false statements and incorrect grammar/with true statements). Subjects seemed to quickly validate and sometimes reject the sentences, and the outcome of their validation affected their spell checking.

Notice that they weren't allowed to answer whenever they wanted to but were prompted to respond quickly at a particular moment, which was the same moment at which the truth-value of the assertion was accessible to them (as the third word completed the assertion). Hence, at that moment, their mental energy for the valuation was depleted. If the rejection of assertions is, as Spinozans claim, resource-dependent, this should have disrupted subjects' rejection of them. But it didn't, as is evidenced by the fact that the validation outcome, which in some cases was a rejection, affected the latency and error rate of the spell check.

It might be proposed that, since strong background beliefs were in place, very little effort was required and invested for rejections.

However, it is hard to see why subjects should have invested *any* effort in rejecting assertions. For investing cognitive effort is generally something that a subject does deliberately in order to achieve some goal or other, yet in the study subjects were not asked to nor had the goal to understand, let alone validate, the assertions. It is thus less plausible to assume that they nonetheless effortfully rejected some of them. It is more likely that they did so automatically, which contradicts the Spinozan theory[29].

5. *Doxastic neutrality*

According to the Spinozan theory, there also shouldn't be cases where subjects remain doxastically neutral about a proposition[30]. But this claim too is arguably false.

Hasson *et al.* conducted an experiment in which they presented subjects with a person's face (e.g., of a smiling man) and a statement about the

[29] The assumption that subjects did invest effort in rejecting assertions despite not having the goal to validate them is also at odds with the well-documented finding that the human mind is a "cognitive miser" in that it tries to avoid spending cognitive resources and tends to adopt mental short-cuts whenever it can (S. Fiske-S. Taylor, *Social cognition*, Sage, London 2013; W. De Neys-S. Rossi-O. Houdé, *Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools*, in «Psychonomic Bulletin and Review», 20 (2013), pp. 269-273.

[30] See Gilbert's (*op. cit.*, p. 109), and E. Mandelbaum (*op. cit.*, p. 62) figures of the Spinozan models; there is no state of doxastic neutrality or suspended belief.

person shown (e.g., "This person thinks that things turn out for the best")[31]. They used three types of statements: true statements that were also indicated as true, false statements that were also indicated as false, and truth-unspecified statements that were not indicated as either true or false. Right after the presentation, participants were presented with a word (for 250ms) and had to quickly press a button to indicate whether it was an English word.

On the critical trials, the word presented (e.g., "optimist") was related to either the true or the false version of the sentence preceding it (e.g., "This person thinks that things turn out for the best"). Hasson *et al.* reasoned that if subjects represent any statement they entertain as true then those who are shown truth-value unspecified statements should respond equally quickly to terms connected with the truth of the sentences (henceforth 'true-related words') in the lexical decision task following *both* true and truth-value unspecified statements. If subjects don't do so, then they should respond more slowly to true-related words following truth-value unspecified statements than following true statements.

Hasson *et al.* found that lexical decisions about true-related words were faster when the statement was indicated to be true than when its veracity was unknown or when it was false. So, for instance, the word "optimist" was evaluated more quickly when the statement "This person thinks that things turn out for the best" was marked as true of a person than when the statement was truth-value unspecified or marked as false[32], suggesting that subjects don't always represent the statements that they entertain as true, but in some cases can remain doxastically neutral about them.

A different set of studies lends further support to this view. According to the Spinozan theory, as Gilbert puts it, «ideas whose truth» have been «ascertained through a rational», effortful «assessment procedure» are «represented in the mind in precisely the same way as» are ideas that have «simply been comprehended; only ideas» that are «judged to be false» are «given a special tag»[33]. True information that one automatically accepts or, upon reflection, endorses remains «untagged»[34].

[31] U. Hasson-J.P. Simmons-A. Todorov, *Believe It or Not: On the Possibility of Suspending Belief*, in «Psychological Science», 16 (2005), n. 7, pp. 566-571.

[32] There might be a priming effect of veracity-related terms on subsequent lexical decisions about statement-related word, but this isn't very plausible, as it is hard to see a semantic link between, e.g., "true" and "optimist".

[33] D. Gilbert, *op. cit.*, p. 109.

[34] *Ibidem.*

With this in mind, Nadarevic and Erdfelder conducted a source-memory study in which test subjects learned statements from three different sources, i.e., fictitious persons called 'Hans', 'Fritz', and 'Paul'[35]. They were told that each of the three persons differed in credibility, which meant that their statements had different truth-values (Hans = 100% true; Fritz = 50% true and 50% false; and Paul = 100% false statements). Half of the test subjects were told about Hans', Fritz's, and Paul's credibility, and therewith of the truth-value of these people's statements, *before* they were presented with the statements (pre-cue group). The other half was informed about it afterwards (post-cue group).

On the basis of studies that show that source memory for validity information is superior to source memory for names[36], Nadarevic and Erdfelder reasoned that pre-cue subjects should display better source memory than post-cue participants. Furthermore, if, as the Spinozan model predicts, people store only 'false' tags, then good source memory in the pre-cue condition should be limited to false statements.

Within the pre-cue group, source memory turned out to be equally good for the true and false statements and was much better than source memory for statements of uncertain validity. Unlike the Spinozan view predicts, subjects seemed to tag statements as true and could refrain from encoding statements as either true or false[37]. For if they had encoded the (uncertain) statements of the unreliable source automatically as true, then pre-cue subjects should have recalled the source of these statements as well as the sources of the true and false statements.

But that is not what Nadarevic and Erdfelder found, which suggests that subjects can remain doxastically neutral about propositions[38].

[35] L. Nadarevic-E. Erdfelder, *Spinoza's error: Memory for truth and falsity*, in «Memory & Cognition», 41 (2013), pp. 176-186.

[36] I. Begg-A. Anas-S. Farinacci, *Dissociation of processes in belief: Source recollection, statement familiarity, and the illusion of truth*, in «Journal of Experimental Psychology», 121 (1992), pp. 446-458.

[37] It might be argued that in the study, subjects equally well recalled the sources of true and false statements because they had enough time to consciously endorse (and not merely to unconsciously automatically accept) statements from a reliable source, which is in line with the Spinozan view. However, this still doesn't explain why subjects were worse at recalling the source of statements with uncertain validity. For, on the Spinozan view, these statements too should have been encoded as true, just as the statements in Gilbert *et al.*'s studies were under load encoded, and later on recalled, as true.

[38] An interesting experiment by Street and Kingstone's provides further evidence for this view. They presented participants with short video clips of people that were either lying or telling the truth. After each clip, the word "Truth" or "Lie" was shown on the screen, indicating

6. *Does the automaticity of believing confer obligations?*

So far I tried to cast doubts on the Spinozan claim that we always initially automatically believe everything we think about. I now want to take a critical look at Levy and Mandelbaum's argument that the automaticity of believing has implications for the ethics of belief. For the sake of argument, I shall set aside the counterevidence to the Spinozan theory that I've just reviewed.

On the basis of the empirical case for the Spinozan theory, Levy and Mandelbaum maintain that we «are designed to initially affirm any propositions that we happen to think about»[39]. They continue that, as a result, those of us "who know about" our "propensities" to believe propositions through merely entertaining them have "obligations to take the risk of forming unjustified" and "immoral beliefs into account" when we expose ourselves to them[40]. Levy and Mandelbaum's thought is that we often have control over what ideas we encounter, for instance, we have control over what television channel we put on (Fox News, BBC, etc.). And since we "make it likely that we will acquire beliefs by mere exposure to them", just «as we have obligations to take risks into account when we act, we have obligations to take the risk of forming unjustified and [...] immoral beliefs into account when we expose ourselves to them», Levy and Mandelbaum conclude[41].

A crucial assumption underlying Levy and Mandelbaum's argument is that subjects who "know about [their] propensities to acquire doxastic states through merely entertaining propositions"[42] will still have the tendency to automatically believe propositions. This assumption, however, isn't supported by the studies (nor arguments) that Spinozans, including

whether the person had told the truth or lied. In some cases, during the verification, participants had to press a button when they heard a tone ring out, which was meant to deplete their cognitive resources. Afterwards, subjects were again presented with the images of the person. Some subjects were asked whether s/he told the truth or lied (truth-lie forced choice condition). The other subjects could also respond that they were unsure as to whether s/he told the truth or lied. Street and Kingstone found that only subjects in the truth-lie forced choice condition automatically took the person to be telling the truth. Subjects who could respond by opting for "unsure" didn't exhibit that tendency, which suggests that subjects are able to merely entertain information. See C. Street-A. Kingstone, *Aligning Spinoza with Descartes: An informed Cartesian account of the truth bias*, in «British Journal of Psychology», 33 (2016), n. 3, pp. 227-239.

[39] N. Levy-E. Mandelbaum, *op. cit.*, p. 26.

[40] *Ivi*, pp. 28, 30.

[41] *Ivi*, p. 30.

[42] N. Levy-E. Mandelbaum, *op. cit.*, p. 28.

Levy and Mandelbaum, have mentioned. All of the studies that Spinozans typically cite involve subjects that are unaware that people automatically accept the propositions that they entertain. Thus, as it stands, Levy and Mandelbaum's argument contains a gap. It leaves open the intriguing possibility that a subject's self-awareness of the tendency to automatically accept the propositions that she entertains disables that tendency[43]. Interestingly, similar interference effects are in related cases not just possible but actual.

One relevant study comes from research on stereotype processing. Stereotype activation, just as Spinozan belief formation, is often taken to be unconscious and beyond the subject's control. To test this, Moskowitz and Li conducted an experiment in which they indirectly activated egalitarian goals in their subjects by asking them, to write down a short description of a past failure at being egalitarian toward African American men. Moskowitz and Li rationale was that

> [m]any models of goal selection [...] reveal that a goal is triggered when one contemplates failure in the goal domain; by a person detecting a discrepancy between their actual responses and a desired response. This discrepancy is said to produce a psychological tension that impels the organism to reduce the tension and approach the standard[44].

After the writing task, subjects were asked to do a lexical-decision task, which is often used to test automatic stereotypes[45]. Following a brief presentation of faces of either Black or White men, which they were told to ignore, subjects had to decide as quickly as possible whether a string of letters comprised an English-language word, which was either a stereotype-relevant term (e.g., "crime", "stupid", "lazy" etc.) or control word (e.g., "annoying", "nervous", "indifferent" etc.).

Moskowitz and Li's thought was that if stereotypes are activated by the face-primes (e.g., a Black face), subjects thus primed should be faster to

[43] Being told that the propositions that one will be presented with are false is distinct from being told that one has the tendency to automatically believe what one is thinking about. Hence, even if the former isn't sufficient for subjects to suspend automatic acceptance (as some Spinozans might argued by using Wegener *et al.*, *op. cit.*), the latter might still be sufficient. I motivate this view below.

[44] G. Moskowitz-P. Li, *Egalitarian Goals Trigger Stereotype Inhibition: A Proactive Form of Stereotype Control*, in «Journal of Experimental Social Psychology», 47 (2011), n. 1, pp. 103-116, p. 106.

[45] See, e.g., M. Banaji-C. Hardin, *Automatic stereotyping*, in «Psychological Science», 7 (1996), n. 3, pp. 136-141.

respond to stereotype-relevant words. And if stereotype control occurs, this effect should disappear and, due to inhibition, stereotype-relevant words should be reacted to more slowly after faces of Black men.

Unlike control participants, subjects with indirectly activated egalitarian goals *did* display stereotype control and inhibition in the lexical-decision task even though during targeted questioning in the debriefing, no participant expressed any conscious intent to inhibit stereotypes on the task, or saw the tasks performed during the computerized portion of the study as related to the reflection on past failures at being egalitarian. The reaction time task was not consciously seen as a way to address an egalitarian goal or as having anything to do with stereotyping[46].

Hence, subjects "can control stereotyping without knowing a stereotype or a goal exists. Consciousness is not required. One's wants, even implicit wants, can direct thoughts" (*ibid.*).

Moskowitz and Li's findings are relevant to Spinozan belief formation and Levy and Mandelbaum's argument pertaining to the ethics of belief. For suppose a subject S comes to believe that she has the tendency to automatically accept everything she is told. It is fair to say that S will take this to be at odds with the way she should form beliefs; gullibility is usually criticised as epistemically problematic[47]. Since that is so, she will detect a discrepancy between her actual responses to propositions and her desired response. As in the stereotype study, this discrepancy is likely to produce a psychological tension that impels her to reduce the discrepancy by approaching her normative standard[48]. If we use the results of Moskowitz and Li's stereotype study as a model, then it is not unreasonable to suspect that S will form the implicit goal to *not* automatically accept the propositions that she entertains, and that this goal will subsequently inhibit her tendency to form beliefs automatically, just as the im-

[46] G. Moskowitz-P. Li, *op. cit.*, p. 108.

[47] For instance, Faulkner writes that given that a «speaker's intentions in communicating need not be informative and given the relevance of these intentions to the acquisition of testimonial knowledge», it is «doxastically irresponsible to accept testimony without some background belief in the testimony's credibility or truth» (P. Faulkner, *The Social Character of Testimonial Knowledge*, in «Journal of Philosophy», 97 (2000), pp. 581-601, pp. 87-88).

[48] This provides a response to the objection that if subjects *could* refrain from accepting the propositions that they entertain at all then surely when they are told before the presentation of some propositions that the latter will be false, they should refrain from accepting them (which, as Wegener *et al.*'s (*op. cit.*) suggest, they don't do). The response to this point is that being told that the propositions will be false won't produce the psychological tension that is required for the mentioned interference with automatic processing.

plicit egalitarian goal in the stereotype study inhibited subjects' automatic stereotyping.

Whether this is in fact the case remains to be seen. My goal here was only to add some plausibility to the view that an insight into one's automatic believing can interfere with that processing. This is enough, because Levy and Mandelbaum's so far uncorroborated assumption that such an insight cannot have that effect is now in need for further support in order for their argument that the automaticity of believing implies epistemic obligations to succeed[49].

7. *Conclusion*

I argued that there is reason to doubt the Spinozan theory that we always initially automatically believe what we think about. The cognitive load studies, which are one of the main sources of support for the theory, are compatible with the view that when we are not under cognitive load, we don't initially automatically believe the propositions that we entertain. There are also studies that suggest that sometimes we automatically reject propositions, or remain doxastically neutral about them.

Furthermore, I maintained that even if we set these studies aside and take the empirical case for the Spinozan theory at face value, Levy and Mandelbaum's argument that those of us who are aware of their automatic belief acquisition have new epistemic obligations remains unconvincing. For one of the assumptions that the argument rests on (i.e., the view that subjects' awareness of their tendency to form beliefs automatically leaves that tendency unaffected) is unsupported and possibly false.

Nonetheless, Levy and Mandelbaum have rightly emphasised the importance of cognitive scientific findings on belief formation for ethical questions about how we should act when we expose ourselves to information. Because even if in subjects who believe that they tend to accept what they think about, this tendency is counteracted, the empirical findings on automatic belief formation do still confer one basic obligation onto us: to make sure that others – especially, for instance, judges and jury members in court, who ought to refrain from accepting (or rejecting) propositions un-

[49] Levy and Mandelbaum might also respond by holding that people who have no idea that they form beliefs automatically are still responsible for their automatic belief formation. But this would require a different argument than the one Levy and Mandelbaum currently propose.

less the evidence supports doing so – know about the way they form beliefs. Fot this knowledge may play a critical role in enabling them to engage in impartial judgment- and decision-making.

Abstract

Recently, philosophers have appealed to empirical studies to argue that whenever we think about a proposition p, *we automatically believe* p. *Levy and Mandelbaum have gone further and claimed that the automaticity of believing has implications for the ethics of belief in that it creates epistemic obligations for those who know about their automatic belief acquisition. I use theoretical considerations and psychological findings to raise doubts about the empirical case for the view that we automatically believe what we think. Furthermore, I contend that even if we set these doubts aside, Levy and Mandelbaum's argument to the effect that the automaticity of believing creates epistemic obligations remains unconvincing.*

Keywords: automaticity; believing; epistemic obligations; ethics of belief.

Uwe Peters
Centre for Logic and Analytic Philosophy, KU Leuven
Department of Economics, University College London, UK
uwe.peters@kuleuven.be

Biology, Ethics and Moral Reflection

Simone Pollo

1. *Metaethics, normative ethics and biology*

According to a well consolidated tradition in analytical philosophy, moral philosophers can engage in at least two different tasks. They can dedicate their work to analysis about the nature of ethics, that is to *metaethics*, or to elaborate arguments justifying specific declinations of moral goods, rights and virtues, that is to *normative ethics*. Of course, philosophers doing metaethics can do also normative ethics, but for a long time they have been intended as separate jobs (to which in the last two decades of 20th Century it has been added *applied ethics*, that is the application of normative theories to practical cases, such as the bioethical ones)[1]. According to the classic understanding of the tasks of philosophical ethics, work on metaethics must be separated from the normative task. This separation has never been understood as a non communication between the two fields. Nevertheless, metaethical analysis has been regarded as a work that could have been done without references to its normative consequences and, on the other side, normative ethics as an enterprise without too much reference to metaethics.

The distinction between metaethics and normative ethics has been often regarded as a dogma for analytical philosophical (and somehow it still is today), even if the possibility to sharply distinguish between the two fields has gradually been put under question. Among the reasons that for long

[1] A brief and useful presentation of 20th Century analytic ethics is: S. Darwall-A. Gibbard-P. Railton, *Toward Fin de siècle Ethics: Some Trends*, in «The Philosophical Review», 101 (1992), n. 1, pp. 115-189.

time allowed to maintain such a distinction there has been the fact that metaethics was almost exclusively understood as the analysis of the language of morals. As well known, pioneers of analytical philosophical ethics understood their work as almost entirely devoted to metaethics and, more precisely, to metaethics intended as an analysis of the language of morals. As a matter of fact this paradigma has been gradually put under question from different points of view and for various reasons[2]. Among the most recent causes that led to such a revision there is a shift that occurred in the field of metaethics in the last years. Metaethical analysis focused on the language of morals have been gradually paired with analysis devoted to the understanding of human moral psychology. Also in this case, there are many reasons for this fact and one of them is the increased interest of moral philosophers in science.

With regard to the contemporary debate, it seems that a strong connection between the theoretical enquiry on ethics and science has been firstly advocated from the side of science. According to E.O. Wilson, the founder of *Sociobiology*, research on ethics should have been, at least temporary, taken off from the hands of philosophers and given to scientists in order to be «biologicized»[3]. Wilson's provocative statement has been greatly criticized and sometimes violently rebutted, but its fundamental claim is the very idea founding the most important contemporary view about the role of science in understandings ethics. This idea (that has distinguished predecessors like David Hume) is that the philosophical analysis of morality cannot be seriously and effectively undertaken without a reliable empirical and naturalized knowledge of human beings and their material conditions of life. Biological science, after Darwin, is the best tool we have to know some basic facts about how human beings "work" and why they are as they actually are. Forty years later the publication of *Sociobiology. The New Synthesis* it can be said that Wilson's dissatisfaction with the traditional philosophical approach to ethics has been seriously taken into account by philosophers themselves. As a matter of fact, many moral philosophers of analytic background have committed their work to a strong bond between philosophical analysis and biological data[4]. This commitment has led to a

[2] The first and most influential critique to metaethics regarded just as linguistic analysis is that raised by G.E. Anscombe, *Modern Moral Philosophy*, in «Philosophy», 33 (1958), pp. 1-19.

[3] E.O Wilson, *Sociobiology. The New Synthesis*, Belknap Press, Cambridge (MA) 1975, p. 562.

[4] N. Levy, *Empirically Informed Moral Theory: A Sketch of the Landscape*, in «Ethical Theory and Moral Practice», 12 (2009), pp. 3-8.

change both of methods and aims in metaethics. Generally speaking, metaethics is no longer regarded solely and mainly as a conceptual and linguistic analysis, but also (or sometimes exclusively) as a biologically informed enquiry about human moral psychology. From this new perspective the enquiry about the nature of ethics is mostly an effort aimed at two goals: the reconstruction of the moral mind and of its biological genealogy (that is its evolutionary path). This kind of «empirically informed» metaethics is deeply intertwined with the researches of evolutionary biology and cognitive science and it is not amiss to speak of this new metaethics as a cognitive science of morality. This is particularly true when philosophers themselves participate to the design and execution of experiments (as so called «experimental philosophers»[5] do), but it is true also when there is no direct commitment to empirical research.

Here I will not attempt a review of various declinations of such a cognitive science of morality. Rather I will try to address a specific issue that is raised by the tight intertwinement of the philosophical understanding of ethics and evolutionary biology. In a nutshell, the aim of this paper is to address the question if the scientific understanding of ethical life can foster moral progress, that is some kind of improvement of real human moral life or if, on the contrary, the scientific comprehension of how morality really works can undermine the potential for human moral reflection and development. The notion of moral progress underlying this question must be clarified. Here "moral progress" is not defined according to its most common meaning, that is the progressive accumulation of some kind of value in the world (like, for example, happiness in an utilitarian framework)[6]. For my present purposes, moral progress must be understood as the development of capacities for moral reflection in actual individual moral agents[7]. The two conceptions of moral progress are not incompatible (as a matter of fact they can be thought as reciprocally bound), but for my present purposes I will assume that moral progress must be defined just at the individual level, that is as the development of individual moral capacities. Given this definition of moral progress, the question to be addressed is whether the scientific understanding of ethics can improve or not the capacities for human moral reflection.

[5] M. Alfano-D. Loeb, *Experimental Moral Philosophy*, in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2016 Edition), URL = <http://plato.stanford.edu/archives/sum2016/entries/experimental-moral/>.

[6] D. Jamieson, *Is there progress in morality?*, in «Utilitas», 14 (2002), n. 3, pp. 318-338.

[7] *Il progresso scientifico come progresso morale. Sentimentalismo, oggettività e scienza*, in «Rivista di filosofia», 107 (2016), n. 2, pp. 219-239.

Raising this question shows that the borders between metaethics and normative ethics can be blurry and fuzzy, since our views about what ethics is, how moral mind works and from where moral life comes from affect normative ideas and attitudes. More specifically the issue of where the biological understanding of ethics can lead human moral life can be considered part of a more general topic, that is how Darwinism changes our understanding of the world (and therefore also of our moral views and attitudes). Here the focus is on the moral meaning of Darwinism, that is how the Darwinian understanding of ethics changes the way morality itself is experienced by human moral agents. Before facing this issue it is necessary to deepen the notion at the core of moral progress, that is moral reflection.

2. *Moral reflection and self-knowledge*

The topic of moral reflexivity is an enormous one and for the purposes of this paper it can be treated from two different, but interlaced, perspectives. As a matter of fact, moral reflexivity is one of those research objects that a Darwinian cognitive science of morality treats and tries to explain both reconstructing its core mechanisms and drawing its evolutionary path. Nonetheless, moral reflection is also the theoretical object of the present analysis, that is the notion at the core of this discussion about the connection between scientific knowledge and moral progress. The notion of moral reflection I use here is deeply rooted in empirical findings about human moral mind and it is itself the outcome of the cooperation between philosophical analysis and scientific research.

Generally speaking moral reflexivity is the capacity moral agents have to critically examine their moral reactions and judgements. The nature of moral reflection depends from the more general conception of the moral mind. Into a rationalistic view of moral psychology reflection is regarded as a process of rational evaluation undertaken by the agent about her own motives and beliefs. Furthermore, if the rationalistic moral mind is also placed in a cognitivist and realist framework, moral reflection will be defined as an operation of discovery and knowledge of moral facts that are relevant for the beliefs subjects to examination. Even if rationalistic (and cognitivist) accounts of moral psychology represent a powerful tradition in the history of ethics since ancient times, there is another influential approach, that is the sentimentalist one. Rooted in the work of 18th Century philosophers as David Hume and Adam Smith, contemporary sentimentalist moral psychology seems to be the view more attuned with the data provided by empirical research on the functioning and development of ethics. Ethology, psychology and neuroscience confirm the basic tenet of sentimentalist moral psychology, that is the idea that the core of human moral capacities is made of affective states[8]. Essential part of a sentimentalist account of moral psychology is the role that sympathy plays in it. The attunement to other affective states and reaction is the drive of altruistic and cooperative behavior and this role is confirmed by empirical researches on humans and non-human animals phylogenetically close to us[9].

This is just a brief sketch to highlight the very basic ideas underlying ethical sentimentalism, but they are enough to present what moral reflection is according to this view of moral psychology. Moral reflection in a sentimentalist fashion is not a rational evaluation and examination, but a process of refinement and transformation of the affective states underlying our moral reactions and driving our motives to act. This process of transformation is driven by real life experiences and imagination and it is oriented at that «general point of view» from which morals sentiments aim at be expressed[10]. Reaching that point of view is not an isolated process (as in a rationalistic perspective could seem), but it is a somehow "social" enterprise. Moral reflection is not just a reflection of the agent on herself, but it is also (and maybe mainly) a process of social mirroring[11]. Our moral sentiments and habits must be, imaginatively or actually, defended in front of the social context into we live in. Therefore, moral reflection aims at establishing moral sentiments and reactions that can pass this kind of test. Moral reflection is also a process of finding justifications for our moral sentiments that could be shared by other moral agents.

According to the sentimentalist view moral reflection does not happen in isolation and relying just on the agent's own capacities (like the *lumen rationis*). Sentimentalist moral reflection is fed by a plurality of sources. Among these sources there are the experiences humans do in ordinary and daily life (for example being in touch with other people and their different

[8] A. Kauppinen, *Moral Sentimentalism*, in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2016 Edition), URL = <http://plato.stanford.edu/archives/fall2016/entries/moral-sentimentalism/>.

[9] F. de Waal, *The Age of Empathy*, Harmony Books, New York 2009.

[10] A presentation of Hume's "general point of view" and of its main interpretations can be found in W. Davie, *Hume's General Point of View*, in «Hume Studies», 24 (1998), pp. 275-294.

[11] J.A. Taylor, *Reflecting subjects. Passions, sympathy, & society in Hume's philosophy*, Oxford UP, Oxford 2015.

ways of living) or through imaginative experience (for example by trying to imagine how it is like to be a calf in a factory farm). The set of sources for moral reflection is broad and pluralist. Into this set also science and philosophical ethics are included. Evolutionary biology and cognitive science of morality can be part of the process of moral reflection: both the philosophical and the scientific understandings of the nature of morality can affect moral reflection and shape sentiments, reactions and judgements. The role played by scientific and theoretical analysis in human ordinary moral reflection should not necessarily be direct and straightforward. Scientific theories influence human moral life not just because people read specialistic articles and books, or attend lecture and conferences. Theories leach into popular culture and become part of our ordinary understanding of the world. Even if many aspects of Darwinism are counterintuitive for human minds (for example, the lack of a goal oriented order in nature), it is now part of the understanding of the world of many persons thanks not just to scientific divulgation (think, for example, to the iconic movie *Inherit the wind* by Stanley Kramer). Therefore, advocating for a role of philosophy and science in ordinary moral reflection does not entail an intellectualistic (and unrealistic) approach to moral reflection.

In particular, the picture of the nature of moral agents emerging from the scientific treatment of morality can meet one basic demand of ethical reflection, that is self-understanding. As a matter of fact, moral reflection is not simply a critical evaluation of one's own reactions, judgments and attitudes, but it is also an assessment of what underlies them, that is the kind of person we are. Moral reflection is also an evaluation of one's own character. The inquiry about the kind of person one is is not just about one's own personal biography but it is broadened beyond the borders of personal life. The question about the kind of person we are goes beyond our present existence in at least two senses. First, one can ask herself how our ancestors heritage shape the kind of person she is (a naive intuition confirmed by science, since our genes contribute to shaping character). The second meaning regards our identity as individuals belonging to a given species. Reflecting on the kind of person we are is also reflecting on what it does mean to be a human being.

The importance of this kind of self-inquiry is stressed by many traditions in ethics and we can track it back to the Greek exhortation "Know thyself!". A reconstruction (also a very sketchy one) of its importance and role in the history of ethical views is far beyond the scope of the present paper. Here I just want to highlight its role in the specific view endorsed here, that is sentimentalism. In fact, another key feature of the neo-Humean naturalistic sentimentalism endorsed here is the role played by the notion of character[12]. More precisely sentimentalism must be intended as a kind of perfectionist ethics, that is a view about morality that stresses not only the agents' behavior, but also the attention of the agent herself on the development and flourishing of her own character[13]. Self-knowledge is entailed not only by moral reflection oriented to evaluate the correct conduct but also (and maybe mainly) by reflection aimed at developing one's own character.

3. *Know thyself! Really?*

The picture of moral psychology (and of ethics in general) emerging from the intertwinement of philosophical analysis and science can contribute to individual moral reflection and, eventually, to moral progress (understood as the refinement of personal capacities for moral reflexivity). This idea establishes a connection between scientific data and theories (more precisely, the philosophical understanding of scientific data and theories) and moral life. This link falls under the the old and controversial topic of the relation between facts and values. I move from the premise that even if "values" (a term to label the different declinations of normativity) cannot be directly deduced from facts (a term to label the different declinations of descriptivity), the separation among the two domains is greatly blurred. The connection between facts and values I am endorsing is not an ontological one (this is not the kind of topic to be faced here), but it is a connection that inhabits moral psychology. Moral reflections and evaluations are soaked in facts. If I want to question from the moral point of view the kind of person I am, that is if I want to reflect on my character and my stable set of moral sentiments, many "facts" will be taken into account and some of them will be descriptions of myself as the particular individual I am and as a member of the human species.

Since the sources for moral reflection are various and different this does not mean neither to advocate for a substitution of moral reflection with

[12] E. Lecaldano, *The Passions, Character, and the Self in Hume*, in «Hume Studies», 24 (1998), pp. 275-294.

[13] W. Donner, *The Liberal Self: John Stuart Mill's Moral and Political Philosophy*, Cornell UP, Ithaca 1991.

scientific knowledge nor to state that scientific data and theories are directly prescriptive. On the contrary, this means that a scrupulous moral agent should consider into her processes of moral reflection also what comes from science and that can be of interest for the particular reflection she undertakes. More precisely, for the purposes of this paper I will examine the role that Darwinian biology and cognitive science of morality can have for moral reflection. I will pose the question whether data from such a science can foster or not moral progress (in the meaning above specified).

Traditionally the relation between darwinian biology and ethics is a controversial one. Darwin himself clearly foresaw the explanatory capacities of his theory for ethical and social behavior (a large part of *The Descent of Man* is devoted to the moral and social faculties) and also its revolutionary consequences on the normative level[14]. Nonetheless, the connection between Darwin's theory and ethics has been immediately misunderstood. The most striking example is represented by the one who is rightly regarded as the first and most passionate advocate of Darwin's theory, Thomas H. Huxley, the so called "Darwin's bulldog". When facing the theme of Darwinism and ethics Huxley substantially missed the potential of Darwin's work for ethical analysis and established an argument that survived long after him, deeply affecting the debate about the relation between Darwinism and ethics. In a nutshell, Huxley claimed that the laws governing evolution can produce just competition, egoism, violence. According to Huxley, it is the cultural human enterprise of ethics that can master the lack of discipline of our biological nature and produce order, just like a gardener take care of the garden and disciplinate the exuberance of life to give it a precise order[15]. Notwithstanding the sincere and passionate commitment for Darwin's theory, Huxley is responsible of having introduced one of the most serious misunderstanding about darwinism that affected its reception until today. Essentially, Huxley identified the law governing the biological evolution with the "law of the jungle" where the survival of the fittest is equivalent to the survival of the strongest. After Huxley many others made the same error and built a tradition of thought stating both that the source of our moral life must be found elsewhere than in our biological nature and that no useful hint for moral reflection could come from science.

[14] J. Rachels, *Created from Animals. The Moral Implications of Darwinism*, Oxford UP, Oxford 1990. See also D.C. Dennett, *Darwins's Dangerous Idea. Evolution and the Meaning of Life*, Simon & Schuster, New York 1995.

[15] T.H. Huxley, *Evolution and Ethics, and Other Essays*, Macmillan, London 1894.

A large amount of data and theoretical analysis has undermined and dismissed both these tenets and the biological roots of altruism and cooperation are nowadays a consolidated area of research (thanks also to the "infamous" sociobiology). Nonetheless, what cognitive science of morality has to say about ethics is not only that – like all other human features – moral life is biologically rooted and it subjected to the mechanisms of biological evolution. As said before, placing moral life under the focus of science leads to a better picture of the core of moral psychology, stressing its affective nature. Nonetheless, the empirically informed portrait of human morality can also yield "unpleasant" consequences for our reflection and self-understanding and, at a first sight, undermine the possibility of moral progress as defined before. Here I will list two topics that seem to undermine the possibility of moral flourishing because of the conclusions one can draw from them about the "nature" of human beings and of moral life.

First, the recognition of the biological and evolution-driven nature of morality can be the ultimate argument against any kind of realism and ontologically grounded claim for objectivity about moral judgements. This is the core of the so called *Evolutionary Debunking Arguments* (EDA) aiming at showing that, given the historical and contingent nature of the evolutionary process that originated morality, moral realism is untenable[16]. Of course, arguments against moral realism are not a novelty. They can be tracked back to philosophers previous to Darwin and also many contemporary declinations are not dependent from the biological understanding of morality. The novelty of EDA is represented by their strong empirical commitment. Antirealism produced by EDA is not just a metaphysical claim (or a linguistic analysis of moral statements), but it is the rigorous consequence of seriously taking into account our best explanations about how morality came into the world. Beside the theoretical differences of the various EDA, they make clear that nowadays claiming a moral realist thesis is untenable. The cost of moral realism is placing somehow morality outside the evolutionary genealogy of morality and this is a too onerous price since it disconnects the understanding of moral life from the best tool we have to grasp the key features of human beings. The role of this undermining of moral realism in moral reflection will be examined later on.

The second issue that has a controversial outcome is the "conservative"

[16] For a useful review and discussion, cf. E. Severini, *Evolutionary Debunking Arguments and the Moral Niche*, in «Philosophia», 44 (2016), pp. 865-875.

picture that seems to spring out of an evolutionary account of morality[17]. Evolutionary accounts of morality seem to produce a too restrictive depiction of human moral capacities. Since altruism and cooperation have been selected for their evolutionary advantage in the specific conditions in which our ancestors lived, it is unlikely that our present capacities sustaining moral sentiments and behavior could be stretched far beyond the boundaries of those conditions. In other terms, strong and inescapable bounds are imposed upon our moral life. If these bounds are truly ineradicable then the challenges of contemporary human life will be never satisfactorily met. How can moral agents selected for altruism and cooperation in small groups face the demands of a globalized life conditions where the outcomes of our daily actions affect people far beyond our sight (as in the case of our behaviors promoting pollution)? In general our biological constitution seems to bind us to a limited altruism and to a sympathy that can not be easily enlarged beyond our proximate circles.

4. *Moral progress and the first-person point of view*

The two topics briefly sketched above provide an uneasy material for moral reflection. When reflecting about ourselves as moral agents and members of the human species we seem to be trapped into both relativism and impotence. On one side moral life appears to be a historical and contingent product where no moral truth can be found. On the other side, our moral capacities seem to be constitutionally flawed and unable to meet the demands of great ideals such as universal benevolence and altruism. What kind of gain can be obtained from the moral lesson of Darwinism? Maybe we should embrace a Nietzschean view about genealogies[18] and condemn the recalling of the past as a burden for individual creativity and self-expression. At a first sight it could seem that it would be better for moral reflection to do without the knowledge of the evolutionary path of our moral mind and its core mechanisms and limitations. For if we take seriously what the cognitive science of morality we could be trapped in a very restrictive view about our potential to develop our character and to shape our

[17] Cf. A. Buchanan-R. Powell, *The Limits of Evolutionary Explanations of Morality and Their Implications for Moral Progress*, in «Ethics», 126 (2015), pp. 37-67.

[18] F. Nietzsche, *On the Uses and Disadvantages of History for Life*, in Id., *Untimely Meditations*, Cambridge UP, Cambridge 1983.

behavior. For sure it seems to be a very restrictive view if compared to the great ideals of conduct embedded in some of our moral traditions. Notwithstanding this fact it is still possible to give reasons to advocate for an empirically informed moral reflection.

Our moral reflectivity is structurally committed to "truth", that is it aspires to meet the world as it really is. Even if the scientific understanding of human nature undermines some moral ideals, it is the most correct depiction we can have. The commitment of the moral point of view and reflection to a reliable account of the world is an "oddity", that is a fact that we must simply recognize. Our moral reactions, sentiments and judgments aim at being attuned with the best depiction of the world we can attain. Part of the claim to objectivity of moral judgement is the claim that those judgments must be fit to the world as it really is (for example we can affirm that something is good for someone also because we also expect to know something true about how that individual is done). At a first sight, it seems that this commitment of the moral point of view to a reliable account of reality could cause a short circuit for moral reflection. In fact, on one side it states the impracticability of any realistic claim in ethics and on the other side affirms that the recognition of this impracticability is the most reliable horizon into which place morality and therefore moral reflection. As naturalized moral agents, we seem to be "trapped" into a contradiction: on one side we are bound to attune our moral reactions to the more realistic picture of reality we can get and on the other side we find that into this picture there is no solid ground for any kind of moral reality and that human moral capacities have strong biological ties.

The presumed contradiction is the outcome of the influence of traditional view on ethics claiming that with an ontologically guaranteed justification morality is lost (something like "if God is dead then everything is permitted"). Contrary to appearances, an empirically informed moral reflection can actually falsify such a claim. When we see ethics from the third person point of view (that is a theoretical and scientific perspective) we correctly see a world where no moral "truth" is ultimately guaranteed and. Nonetheless, we as human beings are used to live the moral life also (and mainly) from the first-person point of view. We make the experience of being social animals capable of empathy, moral sentiments and concern for other human and non-human beings. We make also experience of the limits of our moral capacities and the third-person perspective confirms them and helps us to correctly understand them At the same time, however, we find that these capacities can improve and develop, given some favorable conditions.

Among these conditions there is also a moral reflection that allows us to appreciate ethics from a third person point of view. Science can enlighten some of the conditions required to foster moral flourishing. Putting side by side the third and the first person allows us to better understand the very nature of our moral capacities and to promote their flourishing as one of the peculiar challenges that characterize our life as human animals.

Abstract

In recent years moral philosophers have increasingly paid attention to the development of scientific researches about the functioning of moral mind. Placed into the framework of Darwinian evolutionary theory the cognitive science of morality aims at discovering the core mechanisms of the moral faculties and the evolutionary path that produced them. The intertwinement of cognitive science and philosophical ethics has led to a new understanding of metaethics. Embedding cognitive science in such an investigation switches the focus from the more traditional analysis of the language of morals to the functioning of moral mind. Whereas the contribution of such empirical researches to metaethics is clear and considerable, the role of cognitive science with regard to normative ethics is much more difficult and obscure. Even if the fact/value separation ought to be intended in a soft and non dogmatic way, the normative "use" of empirical findings about human moral minds is a puzzling and slippery task. Rather than being a direct source of norms and values, the understanding of moral psychology carried out by cognitive science contributes to the task of moral reflection insofar as it is a form of self-understanding. Part of the practice of moral reflection – that is critically weighing up and evaluating one's own habits, attitudes and moral responses – is the understanding of one's own nature, both as a specific individual and as a member of the human species. My aim will be to discuss whether the cognitive science of morality could be regarded as a modern answer to the ancient exhortation "know thyself" and, therefore, whether advancements in such science could lead to moral progress.

Keywords: metaethics; science; moral progress; moral reflection; sentimentalism.

Simone Pollo
Facoltà di Lettere e Filosofia
Università di Roma "La Sapienza"
simone.pollo@uniroma1.it

Ethics, Law, and Cognitive Science

T

Emotions and Morality: is Cognitive Science a Recipe for Ethical Relativism?

Massimo Reichlin

1. A substantial amount of evidence, in contemporary (neuro)cognitive science, suggests that moral beliefs are inherently dependent on emotions. Several studies have shown – or purported to show – that emotive reactions not only accompany moral judgments, but also decisively influence them. For example, it has been reported that research subjects tend to provide much more negative judgments, about the moral permissibility of some action, when they are under the influence of a negative smell, or when they are seated at a filthy, rather than a clean, desk[1]. In a famous study involving posthypnotic suggestion, subjects were primed to feel disgust upon hearing a morally neutral word, such as 'often', and this considerably worsened their judgments on morally wrong actions, compared to the neutral condition; some subjects, when in the disgust condition, even blamed behaviour that was not in any sense wrong[2]. Moreover, studies on psychopaths, and on patients affected by lesions in the ventromedial section of the prefrontal cortex, show that these subjects – whose emotive system is highly impaired, and who seem incapable of empathic concern – are unable to distinguish between conventional and moral transgressions, and do not seem to make full-blown moral judgments[3]. These data concur with famous studies using fMRI, according to which the tendency of

[1] S. Schnall *et al.*, *Disgust as Embodied Moral Judgment*, in «Personality and Social Psychology Bulletin», 34 (2008), pp. 1096-1109.

[2] T. Wheatley-J. Haidt, *Hypnotic Disgust Makes Moral Judgments More Severe*, in «Psychological Science», 16 (2005), pp. 780-784.

[3] J. Blair, *A Cognitive Developmental Approach to Morality: Investigating the Psychopath*, in «Cognition», 57 (1995), pp. 1-29; M. Koenigs *et al.*, *Damage to the Prefrontal Cortex Increases Utilitarian Moral Judgements*, in «Nature», 446 (2007), pp. 908-911.

"normal" people to give deontological answers to moral dilemmas is highly influenced by neural activations in areas related to the limbic system[4]. Finally, social psychologists have reported on the phaenomenon of 'moral dumbfounding', *i.e.*, on the tendency of research subjects, asked to support their intuitive, emotionally-dictated responses, to confabulate 'rational' justifications clashing with the available evidence[5].

According to many scholars, the data collected so far provide sufficient ground to claim that emotions are both necessary and sufficient conditions of moral judgments. In its most ambitious form, the 'necessity-and-sufficiency' thesis implies, on the one hand, that one cannot make a moral judgment unless he or she feels an emotional reaction of approval or disapproval towards some action or character; on the other hand, that feeling any such reaction is all that is needed for a moral judgment to be generated. In other words, moral judgments are uniquely caused by emotions and voice our affective states. On this view, the practice of moral reasoning is a *post-hoc* rationalisation of processes whose real nature is entirely emotional, and sometimes even a sort of confabulation, *i.e.* the mere invention of arguments that never played a role in the formation of the judgment[6]. In particular, "deontological judgments" are the direct product of neural activations in the emotive areas and have nothing to do with our reflective capacities[7].

These empirical results have promoted a new wave of ethical sentimentalism, that, in the spirit of experimental philosophy, claims to ground philosophical conclusions on scientific evidence[8]. One such relevant proposal was put forward by Jesse Prinz who, basing on the empirical data concerning the role of emotions in morals, suggested an original view on the nature of ethics, according to which a) moral concepts are essentially

[4] J. Greene *et al.*, *An fMRI Investigation of Emotional Engagement in Moral Judgment*, in «Science», 293 (2001), pp. 2105-2108; Id., *From Neural "Is" to Moral "Ought": What Are the Moral Implications of Neuroscientific Moral Psychology?*, in «Nature Reviews Neuroscience», 4 (2003), pp. 847-850.

[5] J. Haidt *et al.*, *The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment*, in «Psychological Review», 108 (2001), pp. 814-834; J. Haidt, *The New Synthesis in Moral Psychology*, in «Science», 316 (2007), pp. 998-1002. For a good synthesis of the empirical research in this area, see M. De Caro-M. Marraffa, *Mente e morale. Una piccola introduzione*, Luiss UP, Roma 2016, pp. 105-147.

[6] J. Haidt *et al.*, *The Emotional Dog and Its Rational Tail*, cit.

[7] J. Greene, *The Secret Joke of Kant's Soul*, in W. Sinnott-Armstrong (ed.), *Moral Psychology. Volume 3: Emotion, Brain Disorders, and Development*, MIT Press, Cambridge (MA) 2008, pp. 35-79.

[8] S. Nichols, *Sentimental Rules. On the Natural Foundations of Moral Judgment*, Oxford UP, Oxford 2004.

related to emotions, so that the disposition to feel moral emotions is a condition to possess them and b) moral properties consist in emotional facts, that is, the property of being right or wrong consists in eliciting a sentiment of approbation or disapprobation relative to it in an observer. Prinz dubs thesis *a* "epistemic emotionism", and thesis *b* "metaphysical emotionism"[9]. This view singles out emotions and sentiments – the latter conceived of as dispositions to feel certain emotions – as the basic facts of morality. Contrary to old-fashioned emotivism, emotionism believes in the existence of moral facts and properties, but explains such properties with reference to the feelings and emotions of approbation and disapprobation caused in the observers: just as red is the property of causing a sensation of redness in a human perceiver, so rightness is the property of causing an emotion of approbation in a human observer. According to Prinz, then, morality is inherently subjective, since its concepts and properties refer to inner states of human individuals. This would not have relativistic implications, if one were to contend that human emotions display some substantial sort of uniformity (as modern thinkers such as Smith and even Hume suggested); however, Prinz insists much on the cultural relativity of human emotions and sentiments – a conclusion strongly suggested by the historical and anthropological evidence – and this, together with emotionism about moral properties, leads him to metaethical relativism, i.e., the view that «the truth conditions of a moral judgment depend on the context in which the judgment is formed»[10].

In sum, if a) moral properties, such as wrongness, consist in the fact that some observer feels an emotion of disapproval against it, or has a disposition to feel some such disapproval, and b) human emotions are essentially culture-dependent, than c) moral relativism is justified and moralities are sentimental cultural constructions. According to Prinz, if I say "cannibalism is wrong", I am saying that "cannibalism causes me an emotive reaction of disapproval"; but it would misleading for me to go on to say that "The Akamara people ought to refrain from cannibalism", because, while right and wrong relativize to the speaker's values, ought judgments and their normative authority relativize to the values of the agents[11]. Norms against cannibalism, therefore, have no authority against individuals who

[9] J.J. Prinz, *The Emotional Basis of Moral Judgments*, in «Philosophical Explorations», 9 (2006), pp. 29-43; Id., *The Emotional Construction of Morality*, Oxford UP, Oxford 2007.

[10] J.J. Prinz, *The Emotional Construction of Morality*, cit., p. 174.

[11] *Ivi*, p. 179.

did not internalize them. Thus, the scientific evidence on the role of emotions in morals, and the historical and anthropological research on past and distant cultures, give rise to a sort of naturalised genealogy of morals and suggest powerful arguments in favour of ethical relativism. In this paper, I will argue mainly against the epistemic thesis and, as a consequence, against the metaphysical one, suggesting that the scientific data provide no conclusive evidence for a negative conclusion on moral objectivism.

2. According to epistemic emotionism, having the moral concepts presupposes possessing the appropriate emotions; by acquiring the right kind of emotions, human individuals learn to manipulate the moral concepts. As hypothesized in the moral Mary argument[12], an individual lacking any education on the moral emotions could understand everything written by Kant and Mill on normative ethics, but would not have the concepts of right and wrong: she would not know that *x* is the right thing to do, even if she understood that *x* maximizes utility, or respects humanity as an end in itself, for she would lack the proper attitude to utility or humanity. One central piece of evidence for this conclusion is provided by the results of studies on psychopaths, who clearly fail to understand the distinction between moral and conventional norms, treating all norms as conventional[13]. The obvious explanation suggests that it is their emotional impairment that causes their cognitive deficit. Contrary to what others have suggested[14], according to Prinz psychopaths are an argument in favour of motivational internalism, because their lack of empathy, emotions and moral motivation causes their inability to make moral judgments in the first instance. It takes emotions to "see" the moral distinctions, which otherwise are as invisible as colours for colour-blind people.

This view, in my opinion, suffers from several problems. For one thing, it fails to demonstrate that the right order of causation is in all cases from the emotions to the judgments. Nobody can deny that we sometimes feel moral

[12] *Ivi*, pp. 38-42.

[13] J. Blair, *op. cit.* As a matter of fact, Blair found that psychopaths treat all norms as moral: but this finding is generally considered as biased, since the subjects wanted to make a good impression on the researchers. What is clear from Blair's research, is that the psychopaths fail to understand this distinction – which is notoriously vital for the acquisition of moral thought (E. Turiel, *The Development of Social Knowledge: Morality and Convention*, Cambridge UP, Cambridge 1983).

[14] A. Roskies, *Are Ethical Judgments Intrinsically Motivational? Lessons from "Acquired Sociopathy"*, in «Philosophy and Psychology», 16 (2003), pp. 51-66.

emotions popping up vehemently, and a moral judgment comes to our minds and lips without any cognitive interface. In these cases, the judgment is caused univocally by the emotions: however, these cases can hardly account for our whole moral experience. In many cases, in fact, we do not feel an emotion directly leading to a judgment, but, on the contrary, we have to collect a certain amount of information on previous facts – such as broken promises, false declarations and expected negative consequences – in order to reach the judgment that some action *x* is wrong or unjust: it is only when we have made up our minds on this complex structure of facts and reached the normative judgment, that we may (but not necessarily) feel resentment or anger. In such cases, to say that believing "*x* is wrong" expresses or involves an emotive reaction is quite unconvincing. Prinz would say that, in these cases, the reflective process leads us to categorize *x* under some rubric covered by an already internalised rule; such reference to the rule generates the negative emotion and this triggers the moral judgment. This story seems to me unpersuasive. In fact, even though some elements of our normative body may be constituted by rules directly linked to an emotive experience – perhaps summarising our previous emotional reactions – many others are simply learned through education: it is not that emotions give rise to the rules, rather that apprehending the moral rules shapes our emotional reactions. In many cases, babies are taught that something is unjust and this triggers certain patterns of emotional reaction. Of course, the emotive reactions associated to the 'perception' of injustice is a powerful means to reinforce the moral attitude, and therefore they are largely used in moral education; this is why the standard road to acquire moral competence passes through the development of a moral sensibility. However, the fact that emotive reactions are often associated to the moral judgment in this way does not entail that moral concepts such as 'right' or 'wrong' do not convey anything distinct from such reactions. What they convey is the fact that there are valid reasons to consider the respective actions as fit or unfit to be done, reasons that can be cashed out in terms of consequences caused for, or attitudes taken towards, other people; this is why these actions fall under a specific moral rule, and why certain emotive reactions to the actions or the acting people are appropriate.

Moreover, there are also cases in which we do feel emotions of approval or disapproval, but we realise that we lack sufficient reasons to ground them. For example, we feel disapproval for a certain kind of sexual behaviour, such as homosexuality, but then we reflect on our emotions and judgments, and realise that we have no good reasons to hold them. We embark on a process of reflection, weighing reasons for and against our judgment, and reach the conclusion that, contrary to what our education led us to think and feel, that behaviour is *not* wrong, and our judgment is unjustified. Once we have reached this conclusion, we may still feel a negative emotion of disapproval towards homosexual people, but we strive to bring our emotional states in line with our normative judgments[15]. This clearly shows that, in many circumstances, the emotions accompany moral judgments without causing them; moreover, offers a good reason for disbelieving that to make a judgment of rightness or wrongness *is* to feel the respective emotion.

Prinz considers this objection in the context of discussing the view of other "sensibility-theorists", such as McDowell and Wiggins, according to which, unlike colour judgments, that are merely *caused* by their objects, moral sentiments are *merited* by their objects. On this account, moral rightness does not consist in eliciting approbation, but in meriting the approval of qualified observers; and moral judgment is not itself an emotional response, but a judgment that an emotional response is appropriate. Prinz replies that, if the appropriateness that we are talking of is moral appropriateness, then we move in a circle. His view, thus, is that, unless one feels, or has a disposition to feel, the correspondent emotions, his or her judgment is not authentic: in other terms, the homophobic who judges that homosexuality is not wrong, but still has sentiments of disapproval for it, does not really believe that homosexuality is right, but makes a metacognitive judgment on the appropriateness of his homophobia, a judgment that will eventually lead him to the 'right' moral judgment. However, when he says that something meriting an emotion means that «a person who fails to have the emotion could be held accountable»[16], Prinz is in fact accepting that

[15] That this influence of reasoning processes on emotions is not contrary to the empirical evidence is shown by research on 'moral disengagement', in which subjects who perceive a cognitive dissonance between some prospected action and previously held moral intuitions operate an 'anticipatory rationalization': they modify their beliefs relative to the existence of the dissonance in order to avoid guilty feeling and facilitate their preferred behavior (A. Bandura, *Moral Disengagement in the Perpetration of Inhumanities*, in «Personality and Social Psychology Review», 3 (1999), pp. 193-209). It is important to note that moral dumbfounding was described with reference to other-regarding judgments, whereas moral disengagement occurs in self-regarding ones. Moreover, according to the findings of the Moral Identity Theory, moral reasoning need not be biased and self-serving, but can, at least sometimes, function as a disinterested judge (F. Hindriks, *How Does Reasoning (Fail to) Contribute to Moral Judgment? Dumbfounding and Disengagement*, in «Ethical Theory and Moral Practice», 18 (2015), pp. 237-250).

[16] J.J. Prinz, *The Emotional Construction of Morality*, cit., p. 114.

there are right and wrong kinds of emotions, that is, morally appropriate and inappropriate ones. This shows that the rightness and wrongness of actions do not depend on the emotions in the first instance, but on the goodness or badness of the reasons that we have for feeling certain emotions. Explaining this judgment passed on emotions by a meta-sentiment, or a second layer of moral emotions, obscures the fact that a cognitive belief concerning the appropriateness of emotions is needed, to make sense of such cases; in fact, a meta-sentiment lacks the authority that is conveyed by the words used when we say that you *should* or *ought* to change your emotions – for example, you should not have an emotion of disapproval towards homosexual behaviour: this 'should' or 'ought' cannot be an emotion.

A second observation is that Prinz's discussion does not rule out the possibility of expressing moral judgments without feeling the correspondent emotions. It is a fact that we often judge actions and characters in an abstract and detached way, perhaps simply applying some general norm or pattern of evaluation. Prinz himself acknowledges this possibility: according to him, however, when no on-line emotion accompanies the judgement a moral sentiment is nonetheless present, for to have a moral sentiment is to have a disposition to feel those emotions. To this, it may be replied that, in a moderate rationalistic approach, moral judgments are always accompanied by a disposition to feel some emotion: even in a Kantian view, the really virtuous man's affective dispositions are in line with the judgments of practical reason. This man believes that injustice is wrong, and is correspondingly disposed to feel anger towards the unjust: however, it is not this disposition that causes the belief, let alone that justifies it, but the other way around. The presence of a disposition to feel counts in favour of sentimentalism only if we believe that it is this disposition that grounds the moral judgment. But the very fact that we can dissociate the judgment from the emotion, and pronounce moral judgements on hypothetical cases, or discuss of individuals remote from us, with no emotional involvement, is evidence of the non-emotional character of the judgment. According to the sentimentalistic story, what we do in these cases is to reflect on the situation, which elicits no specific emotional reaction in us, and refer it to some other paradigmatic situation that did generate such reaction; it is thanks to this reference that we can make up our mind and formulate a judgment. But this story is uselessly complex. What we do, in such cases, is to reflect on the situation and apply some rule that we have internalised, to reach a moral judgment. And, as previously noted, the sentimentalistic story accounting for the generation of moral rules is far from convincing.

A third point that can be raised against strong emotionism is that it offers an unpersuasive explanation of the difference between conventional and moral rules. According to Prinz, moral wrongness is the property of eliciting a feeling of disapproval in an observer; however, it is clear that many actions which are definitely not morally wrong, do elicit some such feeling: these are the actions violating non-moral rules, based on convention or etiquette. Since both kinds of violation elicit negative emotional reactions, and since the emotional reaction is all that there is to the wrongness of the violation, the only consistent explanation that emotionism can offer of the difference between the two is based on the intensity of the respective emotions: according to Prinz, in fact, «When we think about hitting, it makes us feel bad, and we cannot simply turn that feeling off. Hitting seems phenomenologically wrong regardless of what authorities say. We are less emotional about conventional rules. Speaking without raising your hand is bad, but it does not elicit rage or guilt»[17]. Now, this is true in some cases, but definitely is not always so. For certain violations of the rules of etiquette are no doubt much more disapproved than some violations of the moral rules: for example, violations of conventional rules that arise emotions of disgust (e.g. those relative to behaving at the table, or to exhibiting bodily parts) may be much more resented than violations of fairness in cases of conflicts of interests, or violations of fidelity through the breaking of a promise. This shows that the distinction between moral and conventional rules cuts deeper than our sentimental reactions, having to do with the reasons grounding the two kinds of rules: universal reasons, referring to very general features of human life and relationships, in the case of moral rules, and contingent reasons, referring to historical and local features of a specific human community, in the case of conventional rules. Not by chance, in his later treatment Prinz offer a different explanation, linking the distinction to the fact that moral rules are grounding norms, that is, norms not needing any explanation, whereas conventional rules depend on an appeal to customs[18]. However, since, according to emotionism, also moral rules are based on sentiments grounded by custom or taste, and since grounding norms are conceived as preference-dependent rules on which no rational debate is possible and for which there is no need to argue (“just as I don’t have to argue for the deliciousness of chocolate”)[19],

[17] J.J. Prinz, *The Emotional Basis of Moral Judgments*, cit., p. 37.

[18] J.J. Prinz, *The Emotional Construction of Morality*, cit., pp. 126-127.

[19] *Ivi*, p. 125.

the distinction between moral and conventional rules is reduced to that between non-argued, preference-dependent principles and local principles, backed by traditions and authority. The rationalistic account, insisting on the principled reasons supporting the moral rules, compared to the contingent and historical reasons supporting conventional ones, is definitely superior in accounting for this psychologically fundamental distinction.

A fourth consideration against emotionism is the evidence provided by high-functioning autistic subjects, who notoriously show severe deficits in empathizing and simulating other people's moods and intentions, but nevertheless clearly distinguish between conventional and moral transgressions[20], and, in general, show the capacity for authentic moral judgments. Their moral capacity seems to be based on the mere acquaintance with received or observed rules; the evidence provided by empirical studies in this area seem to justify the conclusion that these individuals in fact develop a form of moral competence «by reasoning [...], on the basis of patient explicit enquiry, reliance on testimony and inference from past situations»[21]. This does not mean that their moral competence is quite "normal", for, in "normal" individuals, moral knowledge is accompanied by the moral emotions[22]; moreover, autistic individuals seem to base their judgments much more on the consequences of actions than on the intentions of agents, relative to typically developed individuals[23]. However, it shows that, although emotions are integral to the usual path through which humans acquire moral knowledge, they are not a strictly necessary condition for the development of moral competence. The two components may at least sometimes be dissociated, and, therefore, the competent manipulation of moral concepts does not presuppose moral emotions.

[20] R. James-R.J. Blair, *Brief Report: Morality in the Autistic Child*, in «Journal of Autism and Developmental Disorders», 26 (1996), pp. 571-579.

[21] J. Kennett, *Autism, Empathy and Moral Agency*, in «The Philosophical Quarterly», 52 (2002), pp. 340-357, p. 351.

[22] Which, of course, also play a part in motivating actions according to our judgments. In the present context, I am not making any claim on the debate between internalist and externalist conceptions of moral motivation.

[23] J.M. Moran *et al.*, *Impaired theory of mind for moral judgment in high-functioning autism*, in «Proceedings of the National Academy of Science», 108 (2011) n. 7, pp. 2688-2692). The case of autistic people is somehow the opposite to that of psychopaths. According to most commentators, autistic people are better candidates for the title of moral individuals, even though some claim that psychopaths make authentic moral judgments as well (A. Roskies, *Internalism and the Evidence from Pathology*, in W. Sinnott-Armstrong (ed.), *Moral Psychology, Volume 3: The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, MIT Press, Cambridge (MA) 2008, pp. 191-206.

3. Although strong emotionism asserts both epistemic and metaphysical emotionism, the two elements can be dissociated[24]. However, it is clear that the truth of epistemic emotionism is a strong reason in favour of metaphysical emotionism: if the mastery of moral concepts presupposes the experience of moral emotions, then either you simply reject the existence of moral facts, or you accept that moral facts are emotional facts. If, however, there are reasons to reject epistemic emotionism, as I hope to have shown, then the case for metaphysical emotionism is seriously weakened. To be true, the rejection of the epistemic thesis does not entail the falsity of the metaphysical one, but it refutes the best argument in its favour.

Of course, beyond arguing against the epistemic thesis, it is also possible to provide positive arguments against metaphysical emotionism. I will only mention one central argument of this kind: the fact that metaphysical emotionism does not account for the claim to objectivity that is characteristic of moral judgment. Actually, metaphysical emotionism has a peculiar difficulty with this element of the standard conception of morality. Prinz declares that emotionism has «a major advantage over expressivism»[25], namely, the fact that, on this view, moral judgments are truth-apt. However, if Prinz's view is right, the moral facts making our moral judgments true, when they are true, and false when they are false, are the speaker's emotions of approval and disapproval. As it happens with any form of strict subjectivism, this has the problematic consequence that we cannot ever be wrong in our moral beliefs, since we can hardly be wrong in referring our emotions. And this, in turn, shows that, according to this view, there is simply no point in our discussing controversial moral issues, such as abortion or just war: in fact, since each participant in the discussion is making moral statements that refer to his or her inner emotive states, his or her judgments are made (almost always) true by adequately reflecting to those states. And there is simply nothing that we can do to avoid the conclusion that two or more contrary beliefs may simultaneously be true. Prinz, of course, tries to avoid this conclusion, by having recourse to a traditional strategy used by expressivists: he adds that we should refer not to our first-impression sentiments, but to our *idealised* ones, i.e., to those

[24] As noted by Prinz himself, classical utilitarianism accepts the metaphysical thesis, for it identifies moral properties with a kind of feeling or sentiment (i.e. pleasure), but rejects the epistemic one, refusing to link the use of moral concepts to the experience of moral emotions; emotivism, on the other hand, is epistemically but not metaphysically emotionist, since it radically disqualifies the idea of moral properties or facts.

[25] J.J. Prinz, *The Emotional Basis of Moral Judgments*, cit., p. 35.

moral sentiments that we have in conditions of perfect knowledge, careful reflection and absence of emotional biases. This is equivalent to saying, as emotivists since Ayer have said, that we should correct all the non-moral facts, in order to pave the way for "adequately" feeling about them. But of course, this still has the unpalatable consequence that, when all the non-moral facts are corrected, and our moral views are still at odds, there is nothing more than we can do, but to admit that we live in different moral worlds. In short, the truth-aptness of the emotionist account is seriously compromised by the mere subjectivity of moral sentiments: Prinz's account provides no real improvement on the emotivistic explanation of moral controversies, an explanation that renders spurious or apparent most of our debates on right and wrong.

Moral disagreement can be readily accounted for, on the other hand, if we accept some forms of moral objectivism, according to which our discussions concern not only the empirical facts, but also the normative significance of such facts, that is, the relation that they bear to our reasons for doing certain acts, or accepting certain principles. Accepting a moral judgment does express the belief that some such fact *x* counts as a reason for doing *A* in circumstances *C*. Although it may prove difficult, in many cases, to reach agreement on what our best reasons in fact are, moral concepts do refer to such reasons, including the reasons for feeling such pro-attitudes as emotions and sentiments.

4. Epistemic emotionism is certainly not true; and this weakens the evidence in favour of metaphysical emotionism. Nothing here said in the attempt to shake the foundations of these two theses, however, is meant to imply that the empirical research on cognitive processes in moral decision-making is not important and worth studying. It is highly plausible to believe that such research can perform the role of excluding certain kinds of philosophical approaches – extreme rationalism being one likely candidate. However, it is also worth stressing that all the evidence grounding present proposals of simple sentimentalism is perfectly compatible with a moderate rationalistic picture, as the one here defended. A moderate rationalist, in fact, may readily accept that emotions are necessary for moral judgments, since, without emotions, our moral thought would be blind: emotions can be conceived as defeasible reasons for normative judgments – that is, as the raw materials of practical reason. The rationalist must only add their susceptibility to reasoning, that is, that there can be good or bad reasons for feeling certain emotions. According to Hanno Sauer, the

moderate rationalist can even accept the sufficiency of emotions for moral judgments, provided that they cause the judgments in a way that is normatively acceptable for the subject: that is, the way in which they cause our judgments must be reflectively endorsed by the subject under conditions of full information and rationality[26].

Whether a non-extreme rationalism should be willing to accept such a weak version of the 'necessary-and-sufficient view', or should concede less to the sentimentalist, I leave it open here. However, I do believe that the specific form of strong emotionism defended by Prinz is highly doubtful and seriously undermined by the arguments offered here (among others).

Abstract

Discussing Jesse Prinz's views on metaethics, the author argues (1) that, as far as epistemic emotionism is concerned, this account does not demonstrate that the right order of causation proceed in all cases from emotions to judgments; does not disprove the possibility of dispassionate judgments; has no persuasive explanation of the distinction between moral and conventional rules; cannot account for autistic morality; and 2) that, as far as metaphysical emotionism is concerned, this account offers a much too deflationary account of moral disagreement. The latter can be best understood within an objectivistic account of the facts (including pro-attitudes such as emotions and sentiments) that provide the best reasons for action.

Keywords: epistemic emotionism; metaphysical emotionism; Prinz; ethical relativism.

Massimo Reichlin
Facoltà di Filosofia
Università Vita-Salute San Raffaele
reichlin.massimo@unisr.it

[26] H. Sauer, *Psychopaths and Filthy Desks*, in «Ethical Theory and Moral Practice», 15 (2012), pp. 95-115. Cf. also K. Jones, *Metaethics and Emotions Research: a Response to Prinz*, in «Philosophical Explorations», 9 (2006), pp. 45-53.

The Ethical Convenience of Non-Neutrality in Medical Encounters: Argumentative Instruments for Healthcare Providers

Maria Grazia Rossi, Daniela Leone, Sarah Bigi

1. *Introduction*

Within the field of health communication, there is a wide consensus on the idea that communication is an important mediator of clinical outcomes. In this respect, it has become clear that the quality and appropriateness of care is guaranteed also by the quality of communication between patients and health providers. This idea has received a strong theoretical and empirical support[1], and Street and collaborators have described the state of the art of this literature by referring to the direct (i.e., an empathic communication could increase the emotional well-being of patients) and indirect pathways (i.e., a clear communication could increase patient knowledge and understanding) from communication to health outcomes, thus clarifying the reason why a good/bad communication may result in better/worse health outcomes[2].

The ethical value of the connection between communication and health outcomes is self-evident, especially in the light of the patient-centered paradigm that is recognized as the most desirable approach in healthcare. In a nutshell, this paradigm suggests that the emotional, psychological and experiential knowledge of patients should be considered as core in the process of healthcare; in this context, a patient centered style of communi-

[1] E.g., R.L.J. Street, *How Clinician-Patient Communication Contributes to Health Improvement: Modeling Pathways from Talk to Outcome*, in «Patient Education and Counseling», 92 (2013), n. 3, pp. 286-291; R.L.J. Street-G. Makoul-N.K. Arora-R.M. Epstein, *How Does Communication Heal? Pathways Linking Clinician-Patient Communication to Health Outcomes*, in «Patient Education and Counseling», 74 (2009), n. 3, pp. 295-301.

[2] R.L.J. Street and collaborators, *op. cit.*

cation should guarantee a respectful management of patient's preferences and opinions, not simply because it is «positively associated with patient satisfaction, adherence, and better health outcomes»[3], but because it ethically safeguards patients' freedom and autonomy[4].

Even if it is not easy to provide a single definition for the concepts of freedom and autonomy, and consequently, a single definition of patient-centered communication[5], there is broad consensus on the idea that doctor-patient mutual understanding counts as an indispensable ethical prerequisite for any patient-centered approach[6]. Therefore, patient understanding becomes a *conditio sine qua non* in a paradigm that aims at enabling patients to be active participants in their care, for example by expressing their preferences when choosing between different treatment options. The basic idea is that a better understanding would allow an adequate shared decision-making between patients and health providers, thus enabling the proper practice of patients' freedom and autonomy. This is the first good reason for focusing on communication, since understanding and then shared decision-making are achieved by means of and within the communication process.

For their part, health providers should give clear information and also take into account the preferences of patients. But again, it is not easy to handle such amount of (provided and received) information as the one that is exchanged during a consultation, at the same time putting into practice highly complex communicative tasks, as the ones foreseen by patient-centered medicine.

[3] M. Stewart, *Towards a Global Definition of Patient Centred Care*, in «British Medical journal», 322 (2001), n. 7284, pp. 444-445, p. 445. See also Id., *Effective PhysicianPatient Communication and Health Outcomes: a Review*, in «Canadian Medical Association Journal», 152 (1995), pp. 14231433.

[4] E. Moja-E. Vegni, *La visita medica centrata sul paziente*, Cortina, Milano 2000; D. Roter-J.A. Hall, *Doctors Talking with Patients/Patients Talking with Doctors: Improving Communication in Medical Visits*, Greenwood Publishing Group, Westport (CT) 2006.

[5] E.J. Emanuel-L.L. Emanuel, *Four Models of the Physician-Patient Relationship*, in «Jama», 267 (1992), n. 16, pp. 2221-2226; R.M. Epstein-R.L.J. Street, *Shared Mind: Communication, Decision Making, and Autonomy in Serious Illness*, in «Annals of Family Medicine», 9 (2011), n. 5, pp. 454-461; H. Ishikawa-H. Hashimoto-T. Kiuchi, *The Evolving Concept of "Patient-Centeredness" in Patient-Physician Communication Research*, in «Social Science & Medicine», 1982 (2013), n. 96, pp. 147-153.

[6] J. Appleyard, *Introduction to Ethical Standards for Person-Centered Health Research*, in «International Journal of Person Centered Medicine», 3 (2014) n. 4, pp. 258-262; J.E. Mezzich-J. Appleyard-M. Botbol-T. Ghebrehiwet-J. Groves-I. Salloum-S. van Dulmen, *Ethics in Person Centered Medicine: Conceptual Place and Ongoing Developments*, in «International Journal of Person Centered Medicine», 3 (2014), n. 4, pp. 255-257.

From a different perspective, still, these communicative tasks are strongly related to the ethical issue of healthcare providers' neutrality. Assuming that from an ethical point of view neutrality is desirable, it remains the case that healthcare providers may make their decisions and propose their therapeutic choices based on (often unconscious) cognitive biases, values, preferences and past experiences[7]. The ideal of neutrality is thus called into question from the non-neutrality emerging from the concrete communicative interactions between patients and healthcare providers. That is the reason why it would be appropriate for clinicians to learn to deal with their own non-neutrality in order to ensure the freedom and autonomy of patients[8]. Since this task is entirely communication-based, in this contribution we suggest that healthcare providers should be equipped with effective communicative and linguistic instruments to manage their non-neutrality. By proposing a case study analysis from the context of Assisted Reproductive Technology (ART), we argue that non-neutrality may paradoxically have – if it is properly managed – a higher degree of ethical convenience (§3). In summary, we show the relevance for the context of health communication of recent issues discussed in cognitive pragmatics and linguistics (§ 2); having in mind the idea that patients' autonomy and freedom is guaranteed by understanding within shared decision-making, we then introduce the argumentative theory of reasoning[9] and we discuss the significant role of argumentative instruments within patient-provider interactions. Finally, we propose a case study analysis of a medical consultation within ART and show how an ethical management of non-neutrality requires an appropriate use of communicative instruments and, more specifically, of argumentative instruments (§3). Finally, we discuss some preliminary results and sketch further lines of research (§4).

2. *Which communicative model for patient-provider interactions?*

While scholars within the field of health communication have produced a lot of evidence to support the idea that communication has direct and

[7] M. Jenicek, *Fallacy-Free Reasoning in Medicine*, American Medical Association, Chicago 2009; Truog *et al.*, Titolo?, casa editrice?, città? 2015).

[8] S. Bigi, *Communicating (with) Care*, IOS Press, Amsterdam 2016.

[9] H. Mercier-D. Sperber, *Why do humans reason? Arguments for an argumentative theory*, in «Behavioral and brain sciences», 34 (2011) n. 2, pp. 57-74.

indirect effects on health outcomes, details are still lacking about which communicative instruments can be considered effective and why. Part of the reason for this gap arises from the fact that also a comprehensive approach to human communication in this research field is missing. In this respect, Bigi has claimed that the adoption of a pragmatic-argumentative approach to the context of patient-provider interactions «can provide answers to the unanswered questions recurrently formulated by health communication scholars»[10]. Following this line of research, we are proposing to draw on recent, rather sophisticated, models for the analysis and description of human interaction outlined by pragmatists and linguists to analyze the specific institutional context of patient-provider interactions.

2.1. *A pragmatic-argumentative model for patient-provider interactions*

The dynamic between cooperation and egocentrism in communication exchanges represents the starting point behind our reasoning. There have been many discussions regarding the dimensions of cooperation and collaboration to define the specific nature of human communication[11]. Some scholars even identified in these dimensions the source of the evolutionary origin of the cognitive mechanism underpinning human communication. For example, Tomasello claimed:

> Human communication is thus a fundamentally cooperative enterprise, operating most naturally and smoothly within the context of (1) mutually assumed common conceptual ground, and (2) mutually assumed cooperative motives. [...] But if we are to understand the ultimate origins of human communication, both phylogenetically and ontogenetically, we must look outside of communication itself and into human cooperation more generally[12].

While the idea of common ground understood as a facilitator for the achievement of cooperation has been adequately investigated from a cognitive point of view and still has great significance in current language models[13], the uniqueness of cooperation as a motivation to explain human

[10] S. Bigi, *op. cit.*, p. 4.

[11] E.g., P.H. Grice, *Logic and Conversation* (1975), in P. Cole-J. Morgan (eds.), *Speech Acts*, Academic Press, New York 1995, pp. 41-58; H.H. Clark, *Using Language*, Cambridge UP, Cambridge 1996; M. Tomasello, *Origins of Human Communication*, MIT Press, Cambridge (MA) 2008; Id., *A Natural History of Human Thinking*, Harvard UP, Cambridge (MA) 2014.

[12] M. Tomasello, *Origins of Human Communication*, cit., p. 6.

[13] D. Sperber-D. Wilson, *Relevance: Communication and Cognition*, Blackwell, Oxford 1995^2.

communication is becoming increasingly controversial[14]. Indeed, many scholars have experimentally examined the psychological processes that guide communication, and discovered that humans exhibit an egocentric bias[15]: humans «have the tendency to take their own perspective to be automatically shared by the other»[16]; that is, speakers and listeners focus on their own knowledge, not on the mutual knowledge assumed as part of their common ground[17]. Also, egocentric motivation is being perceived as theoretically relevant[18] and integrated in a unified model of language[19]. Stressing this latter point, the Socio-Cognitive Approach (SCA)[20] fruitfully integrates the cognitive empirical evidence on egocentrism and describes how cooperation and egocentrism operate within the dynamic process of communication. As stated by Kecskes:

> Communication is the result of the interplay of intention and attention, as this interplay is motivated by the individuals' private socio-cultural backgrounds. This approach [*the SCA*] integrates the pragmatic view of cooperation and the cognitive view of egocentrism and emphasizes that both cooperation and egocentrism are manifested in all phases of communication, albeit to varying extents. While cooperation is an intention-directed practice which may be measured by relevance, egocentrism is an attention-oriented trait which is measured by salience. Intention and attention are identified as two measurable forces that affect communication in a systematic way[21].

[14] E.g., R. Giora, *On Our Mind: Salience, Context and Figurative Language*, Oxford UP, Oxford 2003; I. Kecskes, *The Paradox of Communication-Socio-Cognitive Approach to Pragmatics*, in «Pragmatics and Society», 1 (2010), n. 1, pp. 50-73; U. Peters, *Human Thinking, Shared Intentionality, and Egocentric Biases*, in «Biology & Philosophy», 31 (2016), pp. 299-312.

[15] E.g., N. Epley-B. Keysar-L. van Boven-T. Gilovich, *Perspective Taking as Egocentric Anchoring and Adjustment*, in «Journal of personality and social psychology», 87 (2004), n. 3, pp. 327-339; K. Savitsky-B. Keysar-N. Epley-T. Carter-A. Swanson, *The Closeness-Communication Bias: Increased Egocentrism Among Friends Versus Strangers*, in «Journal of Experimental Social Psychology», 47 (2011), n. 1, pp. 269-273.

[16] U. Peters, *op. cit.*, p. 307.

[17] E.g., D.J. Barr-B. Keysar, *Making Sense of How We Make Sense: the Paradox of Egocentrism in Language Use*, in H.L. Colston-A.N. Katz (eds.), *Figurative Language Comprehension: Social and Cultural Influences*, Erlbaum, Mahwaw (NJ) 2005, pp. 21-41.

[18] E.g., B. Keysar, *Communication and Miscommunication: the Role of Egocentric Processes*, in «Intercultural Pragmatics», 4 (2007), pp. 71-84.

[19] E.g., I. Kecskes-F. Zhang, *Activating, Seeking, and Creating Common Ground: A Socio-Cognitive Approach*, in «Pragmatics & Cognition», 17 (2009), n. 2, pp. 331-355; Id., *Intercultural Pragmatics*, Oxford UP, Oxford 2014.

[20] Proposed by I. Kecskes, *The Paradox of Communication-Socio-Cognitive Approach to Pragmatics*, cit.; Id., *Intercultural Pragmatics*, cit.; and I. Kecskes-F. Zhang, *op. cit.*

[21] I. Kecskes, *The Paradox of Communication-Socio-Cognitive Approach to Pragmatics*, cit., pp. 58-59.

Avoiding to consider only egocentric motivation in communicative interactions, this model seems useful to offer a solution for the potentially pervasive problem of miscommunication and misunderstanding: the existence of egocentric biases appears to give sufficient grounds for considering misunderstandings as problematic, particularly in relation to decision-making in asymmetrical communicative contexts. By focusing both on cooperation and egocentrism, however SCA avoids this problem and proposes a dynamic model of meaning, in which processes behind the co-construction of the emergent common ground – the specific and dynamic knowledge created through interaction – can explain why we manage to understand each other.

To explain how SCA is supposed to work, Kecskes reclaims the distinction between prior and situational context and makes clear how cooperation (by means of relevance) and egocentrism (by means of salience) are both involved within communicative interactions. By using this distinction, Bigi offers a detailed discussion of SCA in the medical context and states:

> following egocentric behaviors, hearers will be guided by what is salient to them in the effort to make sense of what their interlocutors are communicating. The most salient information is usually the most accessible information, i.e. the most easily recalled, the most familiar to the individual, etc. If speakers' and hearers' salience (or private contexts) does not coincide, then the parties will resort to the actual situational context to disambiguate the language and achieve understanding[22].

Bigi analyzes a few cases of alignment and misalignment during medical encounters between the private and actual situational contexts, thus illustrating meaning construction within the dynamic model proposed by SCA[23]. In a similar vein, the following exchange between a nurse (N) and a patient (P) in a diabetes clinic typifies the practical usefulness of SCA in the context of diabetes care[24]. More specifically, this exchange exhibits a case of misalignment between the patient's and nurse's private contexts.

[22] S. Bigi, *op. cit.*, p. 44.

[23] *Ibidem.*

[24] The example is taken from a corpus of videos of follow-up consultations in the context of diabetes care. See S. Bigi, *Healthy Reasoning: The Role of Effective Argumentation for Enhancing Elderly Patients' Self-Management Abilities in Chronic Care*, in «Studies in Health Technology and Informatics», 203 (2014), pp. 193-203.

Speaker	*Text*
1. N	your legs' skin is drier
2. P	dry, yes
3. N	drier than the feet's skin
4. P	they [skin marks] come out… Are they caused by the youth? These skin marks?
5. N	you know, dry skin breaks easily
6. P	oh…
7. N	and you know very well that all these cuts
8. P	but I have every possible lotion at home
9. N	but you leave them in the drawer!

The nurse is running a diabetes foot exam and observes the patient legs' skin with the communicative intention to require a greater skin hydration. The patient's misalignment is very clear (line 4). The patient doesn't understand what the nurse is saying and why it is salient; that is, she focuses on the senile lentigos which are salient in the patient's private contexts but not in the actual situational context (e.g., diabetes foot exam). Thus, the nurse needs to explain why it is important to hydrate the skin before reaching a solid common ground and a successful common understanding.

2.2. *Argumentative instruments for shared decision-making*

The tension between egocentrism and cooperation commonly found in communication exchanges and the need to build a solid common ground to enable understanding are two central aspects affecting decision-making. In the Introduction, we pointed out the ethical value of understanding between patients and healthcare providers within the patient-centered paradigm of care. Indeed, the precarious success of communication is a central issue in asymmetrical contexts such as patient-provider interactions. In these contexts, the distribution of knowledge and procedures is often not shared by speakers: on the one hand, healthcare providers have an advantage with regard to information about procedures, therapeutic regimen and clinical understanding; but on the other hand, patients have an advantage with regard to information about their subjective experience with illness, which can be particularly helpful in establishing diagnosis and plays a major role in disease monitoring. Patients also have an advantage when they are called upon to express their preferences and values on treatment options. This is why the management of appropriate linguistic instruments by providers is extremely relevant to support patients in expressing their autonomy and freedom[25]. Our contention is that argumentative strategies are one of the linguistic instruments available to healthcare providers to achieve this goal.

Abandoning the often implicit idea that argumentation is just a form of manipulation, scholars are showing an increasing interest for the role of argumentation in medical settings and are increasingly proposing argumentative discourse as an adequate instrument ensuring a transparent discussion about different opinions. A pragmatic-argumentative model of communication for patient-provider interactions allows to integrate the interplay between intention and attention, egocentrism and cooperation to account for the complex dynamics at play in deliberation sequences. The asymmetrical social and dialogical roles, the different background knowledge each participant brings to the interaction and the different individual goals of the participants all play a part during deliberation, in both its components, i.e. information sharing and argumentative exchanges. More specifically, through a description of the processes of argument production and (mis)interpretation, it is possible to reconstruct the tension between the individual and the social dimensions, also explaining cases of misunderstanding and misalignment of intentions[26].

These ideas are also consistent with recent insights developed in cognitive science and, more specifically, by theories of reasoning. Particularly relevant for our discussion is the proposal advanced by Mercier and Sperber to consider argumentation as the main function of reasoning[27]. Indeed, the so-called argumentative theory of reasoning claims that «the main function of reasoning is to exchange arguments in dialogical contexts in order to improve communications»[28]. At a theoretical level, this model

[25] S. Bigi, *Communicating (with) Care*, cit.; Id., *Communication Skills for Patient Engagement: Argumentation Competencies as Means to Prevent or Limit Reactance Arousal, with an Example from the Italian Healthcare System*, in «Frontiers in Psychology», 7 (2016); M.G. Rossi, *Metaphors for Patient Education: a Pragmatic-Argumentative Approach Applying to the Case of Diabetes Care*, in «Rivista Italiana di Filosofia del Linguaggio», 10 (2016), n. 2, <http://www.rifl.unical.it/index.php/rifl/article/view/403>.

[26] S. Bigi, *Communicating (with) Care*, cit.

[27] H. Mercier-D. Sperber, *op. cit.*

[28] H. Mercier-M. Boudry-F. Paglieri-E. Trouche, *Natural-Born Arguers: Teaching How to Make the Best of Our Reasoning Abilities*, in «Educational Psychologist», 52 (2017), pp. 1-16, p. 1.

uses the empirical evidence on individual reasoning failures in a positive way: the authors focus on the epistemic function and claim that many biases or errors of reasoning are less puzzling when analyzed by considering reasoning as an argumentation instrument in social dynamics[29]. In this context, argumentation is characterized by a cooperative and adversarial dimension at the same time: on the one hand, argumentation involves a public exchange of reasons by introducing the obligation for the participants to listen to each other; on the other hand, the ultimate goal is "egocentric": the production and evaluation of arguments have the final outcome to convince others and change their mind with respect to the object of discussion.

The implications of this theoretical model are relevant for the topic of this contribution. Indeed, Mercier and collaborators[30] underline links (and benefits) of this new theory of reasoning for the educational domain and, in particular, for improving critical thinking in the context of group discussions and collaborative learning. In contrast to individualist theories of reasoning, they use empirical evidence to point at the fact that in small groups, where subjects have different and contrasting opinions, teaching aimed at improving reasoning is possible to achieve. This may be the case, for example, of the management of shared decision-making between patients and healthcare providers.

3. *A case study analysis in the context of Assisted Reproductive Technology (ART)*

We are proposing that: (1) argumentative instruments are effective to manage the shared decision-making phases within medical interactions, and (2) improvements in the way these instruments are being taught to healthcare providers are necessary. However obvious these concepts may seem, they are actually quite controversial in the literature on doctor-patient communication. Indeed, it is easy to find that argumentation is confused with manipulation[31] and thus rejected as an appropriate means of

[29] See also F. Ervas-E. Gola-M.G. Rossi, *Metaphors and Emotions as Framing Strategies in Argumentation*, in «CEUR-WS», 1419 (2015), pp. 645-650.

[30] H. Mercier *et al.*, *op. cit.*

[31] S. Rubinelli, *Rational Versus Unreasonable Persuasion in Doctor-Patient Communication: a Normative Account*, in «Patient Education and Counseling», 92 (2013) n. 3, pp. 296-301.

communication in medical encounters. Moreover, there is sometimes confusion between what is measured through participants' satisfaction scales and the assessment of the quality of decision-making provided by the analyst. In particular, it is perhaps too strong a claim to assume that the former is a direct reflection of the latter. With regard to this point, Elwyn and Miron-Shatz have recently advocated that more theoretical and empirical efforts are required to evaluate the quality of deliberation[32], which also corresponds to the major component of the shared decision-making process. By taking a closer look at the context of Assisted Reproductive Technology (ART), in the following section we discuss these points in more detail[33].

3.1. *A controversial use of argumentation in an ethically sensitive context*

According to recent surveys, ART is a field with high levels of dissatisfaction: from a clinical point of view, the treatment success rates are still low, around 30%[34]; from a communicative point of view, previous studies have connected patient dissatisfaction with poor communication and low-quality relationships between patients and healthcare providers[35]. Moreover, research showed that ART patients want to be assertive and prefer to have an active role in medical decision and procedures[36].

[32] G. Elwyn-T. Miron-Shatz, *Deliberation Before Determination: the Definition and Evaluation of Good Decision Making*, in «Health Expectations: An International Journal of Public Participation in Health Care and Health Policy», 13 (2010), n. 2, pp. 139-147.

[33] See also G. Lamiani-S. Bigi-M.E. Mancuso-A. Coppola-E. Vegni Lamiani, *Applying a Deliberation Model in Haemophilia Consultations: Implications for Theory and Practice in Doctor-Patient Communication*, in «Patient Education and Counseling», 100 (2016), n. 4, pp. 690-695.

[34] A.P. Ferraretti-V. Goossens-M. Kupka-S. Bhattacharya-J. de Mouzon-J.A. Castilla-K. Erb-V. Korsak-A. Nyboe Andersen, European IVF-Monitoring (EIM) Consortium for the European Society of Human Reproduction and Embryology (ESHRE), *Assisted Reproductive Technology in Europe, 2009: Results Generated from European Registers by ESHRE*, in «Human Reproduction», 28 (2013), n. 9, pp. 2318-2331.

[35] S. Gameiro-J. Boivin-L. Peronace-C.M. Verhaak, *Why Do Patients Discontinue Fertility Treatment? A Systematic Review of Reasons and Predictors of Discontinuation in Fertility Treatment*, in «Human reproduction update», 18 (2012), n. 6, pp. 652-669; R.C. Leite-M.Y. Makuch-C.A. Petta-S.S. Morais, *Women's Satisfaction with Physicians' Communication Skills During an Infertility Cconsultation*, in «Patient Education and Counseling», 59 (2005), pp. 38-45; M. Malin-E. Hemmink-O. Räikkönen-S. Sihvo-M.L. Perälä, *What Do Women Want? Women's Experiences of Infertility Treatment*, in «Soc Sci Med», 53 (2001), pp. 123-133.

[36] E.A. Dancet-I.W. van Empel-P. Rober-W.L. Nelen-J.A. Kremer-T.M. D'Hooghe, *Patient-Centred Infertility Care: a Qualitative Study to Listen to the Patient's Voice*, in «Human Reproduction», 26 (2011), pp. 827-833; V.L. Peddie-E. van Teijlingen-S. Bhattacharya, *Ending*

The analysis we propose has a twofold purpose: on the one hand to show, through an example of suboptimal management of a deliberative sequence, how argumentative competence on the part of the clinician can be a means to safeguard patients' freedom of choice and autonomy in conditions of psycho-emotional fragility and lowered cognitive capacities; on the other, what it means that in many cases participants' perceptions are not a good measure of the quality of the interaction. The example we propose is an excerpt of a visit from a corpus of 85 visits videotaped in eight Italian ART Centers within a broader research project on doctor-patient communication in the ART context.

The excerpt corresponds to one of the deliberative sequences in the consultation; the participants are the doctor and a couple who is consulting her to begin treatment for assisted reproduction. In this particular phase, the woman states that she is willing to undergo only one cycle of treatment and puts forward her reasons for this decision. The clinician has reasons to consider this an ill-informed decision, thus tries to persuade her that she should go for more than one cycle of treatment. The analysis of this deliberation has been conducted using the Method for Dialogue Analysis (MeDA), which allows the description and assessment of dialogical sequences. The method codes dialogue moves according to 7 different categories[37] and is a direct development of the model of types of dialogue[38].

Before turning to the analysis, it is important to add that patients reported high satisfaction for this consultation, which also received high patient-centeredness scores, calculated with the Roter Interaction Analysis System (RIAS), one of the most recognized methodologies for the analysis of medical encounters[39]. The patient-centeredness mean score for all 85 visits was 0,526, where a score of "0" indicates low patient-centeredness, and "1.0" and above indicates high patient-centeredness; the visit from

in-Vitro Fertilization: Women's Perceptions of Decision Making, in «Human Fertility», 7 (2004), n. 1, pp. 31-37; V.L. Peddie-E. van Teijlingen-S. Bhattacharya, *A Qualitative Study of Women's Decision-Making at the End of IVF Treatment*, in «Human Reproduction», 20 (2005), n. 7, pp. 1944-1951.

[37] F. Macagno-S. Bigi, *Analyzing Dialogue Structure. From Types of Dialogue to Dialogue Moves*, in «Discourse Studies», in press.

[38] D. Walton, *Informal Logic*, Cambridge UP, Cambridge 1989; D. Walton-E. Krabbe, *Commitment in Dialogue*, State University of New York Press, Albany 1995.

[39] D. Roter-S. Larson, *The Roter Interaction Analysis System (RIAS): Utility and Flexibility for Analysis of Medical Interactions*, in «Patient education and counselling», 46 (2002), n. 4, pp. 243-251.

which the excerpt was extracted has a patient-centeredness score of 0,98, which is very good. Based on these qualitative and quantitative data, it would seem appropriate to assume also a good management of argumentation during the shared decision-making phases.

In what follows we present the excerpt and its analysis. Doctor (D) is giving information to a couple (labeled MP, for male patient, and FP, for female patient) for completing informed consent. In particular, patients have to decide with respect to the embryo-freezing and D explains why, in their case, they don't have to give consent. Using the model developed by Macagno and Bigi, we have analyzed the dialogical goals of the communicative interaction by coding the various types of dialogical moves[40].

Speaker	*Text*
1. D	since it's better to use a bigger number of egg cells, we can't freeze them, [otherwise]
2. FP	[no::: no::: no no (unint)]
3. D	so, no, we start all over again
4. FP	no, I already decided to go for one try
5. FP	and that's it, because, I think, I mean, I don't think I would be able be able to… start all over again another time. I mean, if it's God's will, otherwise it's like starting a farm…
6. D	wow, you sure sound negative, don't you?
7. FP	[I'm not being negative], I'm a little fatalist
8. FP	because, I feel that I am already forcing a bit… what is supposed to be, [I mean…]
9. D	[but why (unint)]?
10. FP	ah, I don't know, but… that's it
11. MP	well, doc, she's always been kind of negative about kids
12. FP	yeah, I mean, it's not like I've ever been head over heels about kids, I mean, it's not like I'm dying to become a mother. I realize it's something he really wishes, it's probably the age. Kids are cute, all right, but when I was in my thirties I was thinking, no way, I don't want any. Then you grow older and maybe you change your mind, maybe [the context]
13. D	[things change]

[40] F. Macagno-S. Bigi, *op. cit.*

14. FP	things change a bit. But it’s not like I’ve always thought that I wanted to be a mother. No, I wanted to be a woman, a daughter, that’s it. So, I’ve already tried, did everything that was possible, treatm- everything, ‘cause, the past four years we’ve spent always travelling around the place…
15. FP	this is the last time, I’m trying once and then [then that’s it]
16. D	[listen]
17. FP	[because I’m fata-]
18. MP	[listen to me, doc, in the end]
19. FP	[because] I’m fatalist
20. FP	because then,,,, I see people who don’t have any children, people who get children… what if you get a child… that’s not one hundred per cent… I know myself, so
21. D	yeah, well, all right, but then [in any case technology (unint)]
22. FP	[I know that but then…] yeah, sure, techn- of course, but, you know, I’m already forcing the hand…. For me this is forcing nature
23. D	we sure are funny, aren’t we? (chuckling softly)
24. D	you know why, I was thinking, we never have these thoughts [look]
25. D	for example, you get pneumonia
26. FP	it’s true
27. D	and you take antibiotics, when you get cancer- now [mind you, I’m not putting them on the same level]
28. FP	[yeah, of course not, no no no]
29. D	but it’s funny though, because then you don’t think that you’re forcing nature, and instead on this thing about children
30. D	[do you know why] I’m telling you? Because it’s something I get from so many [couples]
31. FP	[really?] eh
32. D	it’s something a lot of people feel, this thing about forcing nature because probably it really comes=
33. MP	[and then after all]-
34. D	[=it’s felt] like something that [should be natural]
35. FP	[should probably be natural] it’s all, mm… a cultural thing we carry with us, I don’t know if it’s something…
36. D	I guess so
37. FP	yeah, probably it’s all a cultural thing, not anything else

38. D	that is rooted
39. FP	that is rooted in- in-… all that catholic thing and bla bla bla you grow up with, it's probably that, but then in the end it's such a part of you that=
40. FP	= for me, that I didn't even want to become a mother, when I was… I mean, we started late for that reason, because when I was thirty the last thing I wanted was to become a mom so… now I'm forty and at this point I think, if I make it that's good, otherwise I go on too much and I feel like a grandma and I don't… I mean, I get all those thoughts, that when my child is thirty I'm seventy [all this kind of stuff, you know, so]
41. FP	one thing- one time, I try
42. MP	sure
43. FP	and then
44. D	ok, so, this decision is very [personal]
45.FP	[sure]
46.D	and I really don't want to interfere because…
47. FP	no no
48. D	although I would really like to tell you something, that will maybe make it a little easier for you

The doctor is giving detailed clinical and procedural information to justify why embryo-freezing is not necessary in this case, when the woman starts sharing her ideas and arguments to support her decision of making just one attempt (from line 4). She explains her position by sharing preferences regarding her individual well-being (e.g., line 12) and advancing arguments that are very often emotionally charged (e.g., line 20):

Justifications for FP's decision of making just one attempt:
1) She feels fatalist (lines 7 and 19);
2) By undergoing ART treatments FP thinks she is forcing nature (lines 8 and 22);
3) In any case, she never wanted to become a mom (lines 12, 14, 40);
4) FP is afraid to have an unhealthy baby (line 20).

The ethical value of these preferences and arguments is out of the question, because they all concern the patient's individual autonomy and freedom in an area such as ART that is *per se* value-laden and emotionally charged. Nevertheless, D's reply is emotionally very strong and ethically undesirable and consists of two main argumentative steps. First, she proposes an undue analogy by building correspondences between different health conditions and their related medical treatments:

a) pneumonia (line 25) and antibiotics (line 27)
b) cancer (line 27) and chemotherapy (implicit)
c) fertility problems (line 29) and Assisted Reproductive Technology (implicit).

In spite of D's *excusatio* at line 27, the analogy is completed at line 29.

The second step taken by D seems to give legitimacy to FP's doubts and ethical preoccupations by aligning FP's feelings with those of many other couples (lines 30, 32, 34, 36); at first, this step may sound as an indication of patient-centeredness. However, the dialogical effect on FP is not encouraging; in fact, she starts considering her worries merely as a byproduct of a cultural influence and consequently she dismisses them (lines 37, 39). D's arguments seem to undermine the patient's values and identity. Indeed, D's persuasion moves are quite personal and difficult to contrast, even more so for patients who are already in an emotionally complex and delicate situation. Furthermore, D's arguments do not relate to clinical or procedural issues, which should of course be shared with patients; instead, they concern personal values and choices, something that does not seem appropriate in this context. The doctor improperly discusses the patient's ethical preferences instead of clarifying why the proposal of a single attempt has a good chance to be unsuccessful from a clinical point of view.

In the final part of this excerpt, FP returns on her main worry (she never wanted to become a mom) and goes on to discuss the consequences of her past choices that are affecting her current decisions (she is feeling too old to become a mom, line 40). At this point, D stops presenting arguments and brings up the issue of neutrality. As shown in lines 44 and 46, she states that it is a personal choice and that is why she does not want to interfere. However, these declarations of neutrality come at the end of the sequence, after she has expressed very strong opinions regarding the patients' doubts and preferences.

Looking at the patient satisfaction score and patient-centeredness scores reported by patients for this consultation (both very high), it could be hypothesized that D's 'profession of neutrality' at the end of the sequence has the effect of canceling in the patients' perception the pragmatic value of her previous moves as arguments in a deliberation, instead suggesting that they have only been attempts at sharing ideas. However, the whole sequence had been triggered by D's comment that the procedure would have to be repeated, generating FP's reply that she had already decided to try it only once (lines 3-4). The interpretation of this sequence as conflict of opinion on a decision, and thus deliberation, is also confirmed by the conclusion of the issue, which comes towards the end of the consultation: the patient postpones the final choice and decides to evaluate her reactions to the first cycle, because later on in the conversation D explains to her that the chances of success are very low in any case, so trying more than once would give her more opportunities to actually get pregnant.

Assuming that the doctor is in good faith and has no hidden agenda, her management of the argumentative phases of this deliberative sequence clearly puts an unwarranted psycho-emotional pressure on the patient, causing her to dismiss her own legitimate doubts and worries, thus not fostering an ideal psycho-emotional condition for further decision-making on the issues at stake.

It is important to note here that the reconstruction of this exchange as an example of inappropriate argumentation by the doctor depends on the theoretical assumptions underlying MeDA[41]. Indeed, it could be argued that D correctly defuses an irrational worry voiced by FP (i.e., "I fear I am forcing nature"), while showing respect and even a tactful handling of a valid concern she presents (i.e., "I am not so sure I want to have children"). Namely, D should consider the first worry as patently unfounded for at least two reasons: first, if "forcing nature" is a genuine worry of FP, she should not even try once; second, if trying ART means going against nature, then the same should be true of curing whatever health problem one happens to have – which is precisely the analogy drawn by D[42]. To explain why this reconstruction should not be adopted we need to further specify and define what it means that healthcare providers should use non-neutrality in a proper way.

What "proper way" means from a pragmatic-argumentative point of view is defined in terms of "dialogical relevance", i.e. the ability of single dialogical moves to be coherent with the joint dialogical goal[43]. Especially in institutional contexts such as the medical encounter, the joint dialogical goals correspond to the institutional goals and admissibility rules may be in place in relation to the dialogical moves that can be used to realize them[44]. In the excerpt analyzed above, the medical explanation about the

[41] F. Macagno-S. Bigi, *op. cit.*

[42] We thank one anonymous referee for pointing out to us this alternative compelling reconstruction of the exchange.

[43] F. Macagno-S. Bigi, *op. cit.*

[44] S. Levinson, *Activity Types and Language*, in P. Drew-J. Heritage (eds.), *Talk at Work*, Cambridge UP, Cambridge 1992, pp. 66-100; S. Bigi, *Communicating (with) Care*, cit.

low success rates to get pregnant has a dialogical relevance for the realization of the higher order intention of explaining from a clinical point of view why the proposal of a single attempt has a good chance to be unsuccessful. On the contrary, D's analogy against the worry voiced by FP is not dialogically relevant in view of the joint clinical goal. A proper managing of non-neutrality requires at least the recognition of what is dialogically relevant in the light of a specific role in a specific context: from our perspective, D must face the doubts and worries expressed by PF clarifying how and why they may have an impact at the clinical level (non-neutrality managed in a proper way); D should not tackle the doubts and worries expressed by PF with a view to challenge her ethical preferences and opinions (non-neutrality managed in an improper way). And obviously, this assessment of the quality of deliberation may be further detailed to include the analysis of its argumentative structure[45].

4. *Conclusions*

Our analysis of the excerpt from an ART visit in the previous section shows a discrepancy between the high measures of both patient satisfaction and patient-centeredness, and the low quality of argumentation during a deliberative phase. Even if this analysis is just an illustration and further data are necessary to evaluate the reliability of this provisional result, new assessment tools seem necessary in order to evaluate understanding and shared decision-making in a more appropriate way. In this respect, argumentative models and tools might offer a better assessment of understanding and shared decision-making. A study by Lamiani and collaborators goes in this direction and constitutes a first step to systematically evaluate the quality of deliberation by using a pragmatic-argumentative model of language and communication[46].

Regarding the issues discussed in this contribution, there are two main concluding remarks:

(1)socio-cognitive models of language and reasoning such as those discussed in the previous sections, offer solid theoretical backgrounds for

[45] See G. Lamiani *et al.*, *op. cit.*; F. Macagno-S. Bigi, *op. cit.*

[46] G. Lamiani *et al.*, *op. cit.* See also S. Bigi, *Communicating (with) Care*, cit.; F. Macagno-S. Bigi, *op. cit.*

interdisciplinary research in the fields of education and health communication; we focused mainly on their importance for the education of healthcare providers, but the same applies also for patient education[47];

(2)concerning argumentative instruments, our general point is that healthcare providers must learn to properly use these instruments in order to guarantee understanding and manage the shared decision-making phases with patients. More specifically, precisely to avoid ambiguous and improper use of neutrality, above all in highly value-laden and emotionally charged argumentative contexts such as ART, healthcare providers should use non-neutrality in a proper way – from an argumentative and ethical point of view. Patients seek advice on the desirability of treatments, healthcare providers must be ready (and trained) to provide it properly.

It is the time to make a concerted and interdisciplinary effort to integrate knowledge and methodologies; this is the only way to view communication in institutional settings as the product of a range of skills that can (and must) be taught, and stop considering it merely as a personal talent, happening only in a few, fortunate cases[48].

Abstract

Many scholars have shown the relevance of communication as an instrument of care by arguing that the quality of the doctor-patient relationship – also based on the quality of verbal communication – affects the engagement and outcomes of patients. This understanding of such therapeutic role of communication paves the way to a re-consideration of ethical questions in clinical contexts: if communication is a therapeutic instrument, then healthcare providers need to be able to properly use it. Our main aim in this contribution is to argue that it is possible and desirable to adopt and manage non-neutral communication strategies to safeguard patients' freedom and autonomy in making decisions. More specifically, we use a pragmatic-argumentative model of verbal communication to deal with the topic of neutrality.

[47] M.G. Rossi, *op. cit.*

[48] The authors would like to thank the anonymous reviewers for their helpful comments that contributed to improve a previous version of the article. This article has been developed within the framework of the research project "Healthy reasoning: strategies and mechanisms of persuasion in chronic care", supported by a grant from the Italian Ministry of Education, University and Research (Program MIUR-FIR 2013, Grant n. RBFR13FQ5J, PI: Sarah Bigi).

Analyzing a case study from the context of Assisted Reproductive Technology (ART), we underline the highly ethical relevance of this medical context and stress the importance of an appropriate use of argumentative and communicative strategies to protect patients' values and decisions.

Keywords: doctor-patient communication; dialogical relevance; non-neutral communication; patient-centered medicine; Assisted Reproductive Technology.

Maria Grazia Rossi, Ph.D.
AgrLab-Institute of Philosophy of Language (IFINova)
Universidade Nova de Lisboa - Lisboa, Portugal
mgrazia.rossi@fcsh.unl.pt

Daniela Leone, MS
Dipartimento di Scienze della Salute
Università degli Studi di Milano
daniela.leone@unimi.it

Sarah Bigi, Ph.D
Dipartimento di Scienze linguistiche e Letterature straniere
Università Cattolica del Sacro Cuore
sarah.bigi@unicatt.it

Ethics, Law, and Cognitive Science

T

Responsibility and Control in a Neuroethical Perspective

Elisabetta Sirgiovanni

1. *Responsibility and conscious control in folk ethics and law*

The notion of responsibility is so pervasive in our daily lives that it needs proper understanding and stable conceptualization. Responsibility orients moral and legal theories and practices for many reasons, which reflect ideas of justice and fairness in a long spectrum from desert to social benefits.

The first thing to notice about responsibility is that it lacks a unitary meaning in contemporary usage, and this is why is so difficult to get a clear definition of it. There are, however, some shared assumptions, which orient the folks both in formal and informal contexts. I will concentrate mostly on retrospective responsibility in the negative form, although I believe that some clarifications might be useful also for a prospective and positive sense of "responsible".

When is someone accountable for her actions? According to folk ethics, responsibility attribution depends strictly on the idea of conscious control over actions[1] or *agency*, to use a philosophical term. Folk ethical theories claim that an agent can be held responsible for morally relevant outcomes of her actions *iff* her conscious intentions control her actions. Only agents whose actions can be ascribed to conscious control are commonly held to be responsible for the outcomes of their actions, even

[1] See E. Nahmias-S. Morris-T. Nadelhoffer-J. Turner, *Surveying Freedom: Folk Intuitions About Free Will and Moral Responsibility*, in «Philosophical Psychology», 18 (2005), pp. 561-584; E. Nahmias-J. Coates-T. Kvaran, *Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions*, in «Midwest Studies in Philosophy», 31 (2007), pp. 214-242.

though this is expressed in degrees so that the more conscious control an agent has the more s/he is held responsible. The commonsense idea of control, then developed in cybernetics and automata theory, is that «A *controls* B if and only if the relation between A and B is such that A can drive B into whichever of B's normal range of states A *wants* B to be in»[2]. As Dennett points out, the commonsense idea of control implies that «for something to be a *controller* its states must include desires», namely conscious attitudes[3].

In short, common sense favors the view that moral responsibility requires not only a causal relationship between the agent and her actions, given that we know s/he was author of those actions, but also control over her actions. Moreover this view implies that control should be conceived in terms of the agent's conscious intentions (beliefs, desires, etc.).

The first of two main assumptions about moral responsibility is that it requires something more than just causal responsibility[4]. Causation only provides a necessary link between the agent, the proscribed conduct and its outcomes, but moral responsibility is believed to arise from the agent's conscious intentional states. If something outside my intentions (e.g., someone else, a machine, a mental disorder) controlled my act, I am not usually held responsible for it. Moreover, we are held in control when these actions are the product of our decisions, which usually means that these decisions derive from our deliberation, or better from-reasoning processes to which we have access to by introspection.

The second assumption is that moral responsibility depends on a link between these internal criteria and external criteria of attribution. External criteria of responsibility attribution defines what outcomes of actions are held morally relevant. External criteria may vary among individuals, cultures and societies.

Prevailing moral and legal theories of responsibility seem to reflect these folk assumptions. This view has been defended in the history of philosophy[5]

[2] As reported by D.C. Dennett, *Elbow Room: The Varieties of Free Will Worth Wanting*, MIT Press, Cambridge (MA) 1984, p. 52.

[3] *Ibidem.*

[4] See J.M. Fisher-M. Ravizza, *Responsibility and Control: A Theory of Moral Responsibility*, Cambridge UP, Cambridge 1998.

[5] E.g., I. Kant, *Critique of Practical Reason* (1788), in T.K. Abbott (ed.), Prometheus Books, Amherst (NY) 1996; I. Kant, *The Metaphysics of Morals*, 1797, in M. Gregor (ed.), Titolo?, Cambridge UP, Cambridge 1991. For more recent literature, see N. Levy, *Consciousness and Moral Responsibility*, Oxford UP, Oxford 2014.

and in moral psychology[6]. In Western legal systems, degrees of responsibility (and punishment) for a crime are defined by a link between the so-called guilty mind (*mens rea*) and guilty act (*actus reus*) where the concept of the guilty mind includes both the agent's state of mind at the time of the act and the lack of mental insanity. Thus, on the one hand, the extent of conscious will in the action defines a taxonomy of both the nature of the crime and the degree of the punishment, while on the other hand a mental insanity defense may determine a verdict of diminished or lack of responsibility. The control condition is often referred to as *capacity*-responsibility[7], which is the idea that in order to be responsible the person must have certain capacities like understanding, reasoning and control of conduct. The idea of the guilty act, instead, is characterized by the idea that a reprovable act is not only a mechanically defined bodily conduct. What we need are some definitional features legally identifying standards of conducts and outcomes of the action. An example is that of a dangerous driver who causes a pedestrian's death. As Cane[8] claims, «the law doesn't ask whether the driver's bodily movement caused the dangerous driving; but it does ask whether the driver's bodily movement, under the description of 'dangerous driving', caused the death». So we need to legally describe the conduct (i.e., the limit of speed beyond which driving is held dangerous) and this conduct must have extrinsic consequences (i.e., third party damages).

Common circumstances that work as excuses in legal contexts are for example force majeure (unavoidable accident) or self-defense. This is particularly relevant because in these circumstances the agent is not held responsible even if s/he has full control of her own actions (s/he intentionally decided to act) and these actions have the worst possible consequences (like for example, causing someone's death).

2. *The Frail Responsibility Hypothesis from cognitive neuroscience*

As we have seen, conscious control on actions is a fundamental assumption both in folk ethics, moral philosophy and psychology, and in the law as

[6] J. Piaget, *Le Jugement Moral chez l'Enfant*, Alcan, Paris 1932; L. Kohlberg, *The Development of Modes of Thinking and Choices in Years 10 to 16*, PhD Dissertation, University of Chicago, Chicago 1958.

[7] H.L.A. Hart, *Postscript: Responsibility and Retribution*, in *Punishment and Responsibility*, Oxford UP, Oxford 1968.

[8] P. Cane, *Responsibility in Law and Morality*, Hart, Oxford 2002, p. 115.

concerns responsibility. However, research in cognitive neuroscience has introduced a hypothesis that goes against this common-sense assumption. The hypothesis has been called Frail Control Hypothesis (FCH)[9]. FCH claims that: «even in unexceptional conditions, humans have little control over their behavior»[10]. Suhler and Churchland mean to refer to the fact that we miss *conscious* control on our behavior while we may have it unconsciously. What is relevant for us here is that FCH implies a Frail Responsibility Hypothesis (FRH). But let's first concentrate on FCH.

What are the empirically motivated challenges to conscious control? They are a series of counter-intuitive findings, which inspired the birth of neuroethics itself as a separate area of inquiry[11] and have become classic in the debate. These findings, which go against moral intuitions that we consciously originate and regulate our actions, regard four main domains (even if other ways of grouping could be suggested): unconscious will[12], reason confabulation[13], emotional processes involved in moral judgments[14], and false self-attributions[15]. These findings go along with other evidence about the fallible character of mindreading faculty, presumably

[9] C.L. Suhler-P.S. Churchland, *Control: Conscious and Otherwise*, in «Trends in Cognitive Sciences», 13 (2009), n. 8, pp. 341-347.

[10] *Ivi*, p. 341.

[11] A.L. Roskies, *Neuroethics beyond Genethics*, in «EMBO Reports», 8 (2007), n. S1, pp. S52-S56.

[12] B. Libet, *Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action*, in «Behavioral and Brain Sciences», 8 (1985), pp. 529-566; B. Libet, *Mind Time: The Temporal Factor in Consciousness*, Harvard UP, Cambridge (MA) 2004; C.S. Soon-M. Brass-H.J. Heinze-J.-D. Haynes, *Unconscious Determinants of Free Decisions in the Human Brain*, in «Nature and Neuroscience», 11 (2008), pp. 543-545; S. Kühn-M. Brass, *Retrospective Construction of the Judgement of Free Choice*, in «Consciousness and Cognition», 18 (2009), n. 1, pp. 12-21; N. Wolpe-J.B. Rowe, *Beyond the "Urge to Move": Objective Measures for the Study of Agency in the Post-Libet Era*, in «Frontiers Human Neuroscience», 8 (2014), p. 450.

[13] R.E. Nisbett-T.D. Wilson, *Telling More than We Can Know: Verbal Reports on Mental Processes*, in «Psychological Review», 84 (1977), pp. 231-259; J. Haidt-F. Bjorklund-F.S. Murphy, *Moral Dumbfounding: When Intuition Finds No Reason*, Unpublished manuscript (2000); W. Hirstein, *Brain Fiction, Self-Deception and the Riddle of Confabulation*, MIT Press, Cambridge (MA) 2006.

[14] There is extended literature, classic works are: A.R. Damasio, *Descartes' Error: Emotion, Reason, and the Human Brain*, Vintage, London 1994; J.D. Greene-R.B. Sommerville-L.E. Nystrom-J.M. Darley-J.D. Cohen, *An fMRI Investigation of Emotional Engagement in Moral Judgment*, in «Science», 293 (2001), n. 5537, pp. 2105-2108.

[15] D.M. Wegner-T. Wheatley, *Apparent Mental Causation: Sources of the Experience of the Will*, in «American Psychologist», 54 (1999), pp. 480-491; A. Dijksterhuis-H. Aarts-P.K. Smith, *The Power of the Subliminal: On Subliminal Persuasion and Other Potential Applications*, in R. Hassin-J. Uleman-J.A. Bargh (eds.), *The New Unconscious*, Oxford UP, Oxford 2005, pp. 77-106.

devoted also to interpret our own unconscious conceptual processing[16], which is thought not to be directly broadcasted to awareness contrary to what happens for sensory information[17]. All these data undermine the idea of reliability of self-reports of one's own actions as well[18]. I will not discuss results that have received wide attention and criticism in the neuroethical debate over the years. I will only point out that even a broad interpretation of these findings is an open issue but still a concern for defenders of common-sense theories of responsibility, which require conscious control.

There are a number of advocates of various versions of FCH among psychologists and philosophers[19]. I must clarify that I am interested here in various conceptual meanings of the conscious control issue, but not with that of causation either in the free will version[20] or in that of the causal role for the conscious mind[21].

According to Suhler and Churchland, FCH inspires a Frail Responsibility Hypothesis (FRH) that may be summarized as follows:

1. The common-sense idea is «that to be responsible we must have "normative competence", meaning that we consciously weigh the evidence, effectively deliberate, and make a decision».

[16] S. Nichols-S. Stich, *How to Read Your Own Mind: A Cognitive Theory of Self-Consciousness*, in Q. Smith-A. Jokic (eds.), *Consciousness: New Philosophical Essays*, Oxford UP, Oxford 2003, pp. 157-200. P. Carruthers, *How we Know Our Own Minds: The Relationship between Mindreading and Metacognition*, in «Behavioural and Brain Sciences», 2 (2009), pp. 121-138.

[17] B.J. Baars, *A Cognitive Theory of Consciousness*, Cambridge UP, Cambridge 1988; S. Dehaene-J.P. Changeux, *Experimental and Theoretical Approaches to Conscious Processing*, in «Neuron», 70 (2011), n. 2, pp. 200-227.

[18] P. Carruthers, *The Opacity of Mind, An Integrative Theory of Self-knowledge*, Oxford UP, Oxford 2011.

[19] E.g., J.M. Doris, *Persons, Situations, and Virtue Ethics*, in «Nous», 32 (1998), pp. 504-530; G. Harman, *Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error*, in «Proceedings of the Aristotelian Society», 99 (1999), pp. 315-331; T.D. Wilson, *Strangers to Ourselves: Discovering the Adaptive Unconscious*, Harvard UP, Cambridge (MA) 2002; D.M. Wegner, *The Illusion of Conscious Will*, MIT Press, Cambridge (MA) 2002; J.A. Bargh, *Free Will Is Un-Natural*, in J. Baer *et al.* (eds.), *Are We free?: The Psychology of Free Will*, Oxford UP, Oxford 2008, pp. 128-154; K.A. Appiah, *Experimental Philosophy*, in «Proceedings and Addresses of the American Philosophical Association», 82 (2008), pp. 7-22; P. Carruthers, *The Opacity of Mind*, cit.; P.S. Churchland, *Touching a Nerve: The Self as Brain*, WW Norton & Company, New York-London 2013.

[20] A.L. Roskies, *Neuroscientific Challenges to Free Will and Responsibility*, in «Trends in Cognitive Sciences», 10 (2006), n. 9, pp. 419-423.

[21] M. King-P. Carruthers, *Moral Responsibility and Consciousness*, in «Journal of Moral Philosophy», 9 (2012), pp. 200-228; N. Levy, *A Role For Consciousness After All*, in «Journal of Moral Philosophy», 99 (2012), pp. 255-264.

2. Neurocognitive evidence shows that «the deciding and weighing is below the level of consciousness».
3. There is no effective self-knowledge or proper mirroring in our consciousness of our unconscious intentions.
4. So «normative competence is compromised» (there is no conscious normative competence, maybe only conscious one).
5. «No [conscious] normative competence, no responsibility»[22].

Note that premise 3 is not in Suhler and Churchland's original argument. However, I believe we need to add it in order to preempt the objection that conscious deliberation still counts toward responsibility insofar as it actually reflects our unconscious intentions[23]. On the contrary, we need to take into account evidence to the effect that our introspective processes are terribly fallible (to what extent is question for future research). If self-reports of conscious intentions are fallible and hardly overlap unconscious intentions, responsibility attributions based on self-reports are likely to be fabrications (again we do not know how much). But what if this comes out to be true in the worst sense, that they are completely fabricated? How should we face the question of responsibility in such a neuroscientifically informed account, given that responsibility is essential to our social relations?

My aim in this paper is to examine whether and how we can preserve a notion of moral and legal responsibility in terms of control that fits with a neurocognitive perspective. I interpret the area of neuroethics as a ground for reformulating concepts and theories in the ethical domain thanks to achievements that come from neuroscientific studies[24]. I will examine possible solutions to the neuroscientific threat to conscious control and responsibility and will discuss objections to all of them. Then, I will try to give some suggestions for building a neurocognitive account of responsibility that unifies the benefits of these hypotheses and takes their limitations into consideration. I will not consider the issue of punishment since, although related, I believe it requires a separate scrutiny.

[22] C.L. Suhler-P.S. Churchland, *op. cit.*, p. 342.

[23] See N. Levy, *A Role For Consciousness After All*, cit.; Id., *Consciousness and Moral Responsibility*, cit.

[24] According to Eric Racine, this is a *knowledge-driven* perspective on neuroethics endorsed by Patricia Churchland and Adina Roskies. See E. Racine, *Pragmatic Neuroethics, Improving Treatment and Understanding the Mind-Brain*, MIT Press, Cambridge MA, 2010; P.S. Churchland, *Braintrust, What Neuroscience Tell Us about Morality*, Princeton University Press, Princeton, 2011; A.L. Roskies, *Neuroethics for the New Millennium*, in «Neuron», 35 (2002), n. 1, pp. 21-23.

3. *Standing still and four roads ahead*

Here I will present four possible roads ahead in order to account for responsibility from a neuroethical perspective. Before considering them, I will mention the so-called "Let's Pretend Hypothesis"[25], which we can think of as a sort of "standstill", so not a proper solution. According to it, «in neuroscientific terms, no person is more or less responsible than any other for actions», but what matters is that responsibility is a «social choice»[26].

The idea is that responsibility is a social construction, which exists in the rules of society and not in the brain, with the purpose of maintaining and protecting civil society, a «legal fiction» driven by «our collective interest»[27]. Michael Gazzaniga, a defender of this view, ends up saying that if responsibility works this way we should maintain it, basically by pretending it describes how things actually are – even if this is not the case – because this notion succeeds in its purposes.

A main objection to this hypothesis is that it cannot «specify relevant criteria for distinguishing between those who could have done otherwise and those who could not have, and between those cases in which *mens rea* (literally, a guilty mind) obtains and those in which it does not», and that it «implies that there are no relevant factual differences between someone with, say, obsessive-compulsive disorder and someone who can resist impulses», so it is «not particularly compelling, nor even coherent»[28].

Another objection is that a real difference cannot be discerned between Gazzaniga's view and authors who believe that folk psychology works and should be preserved because it is true[29], except for the fact that retributivism would make sense in the folk conception and not in the Gazzaniga's view but that is, like I said, a different question.

Thirdly, the idea of preserving a fiction seems to threaten the role of neuroethics itself because in that case we must justify why we should have such a specific area of ethics and what its goals are, if neuroscientific find-

[25] This is how Patricia Churchland explains Michael Gazzaniga's view, see P.S. Churchland, *Brain-Based Values*, in «American Scientist», (2005), available online: <http://www.americanscientist.org/bookshelf/pub/brain-based-values>.

[26] M.S. Gazzaniga, *The Ethical Brain*, Dana Press, New York-Washington (DC) 2005.

[27] P.S. Churchland, *op. cit.*

[28] *Ibidem.*

[29] This view is defended for example by Stephen Morse in most of his works, see the recent S.J. Morse, *Criminal Law and Common Sense: An Essay on the Perils and Promise of Neuroscience*, in «Public Law and Legal Theory Research Paper Series», 99 (2015), pp. 38-72.

ings about moral behavior are not, in the end, ethically relevant.

There is however a benefit of Gazzaniga's hypothesis. That is, it is highlighting the core social constructive character of the notion of responsibility, which varies among cultures and societies and that functions differently in each particular social environment where it applies. This hypothesis reminds especially that it is not possible to universally decide who is responsible and who is not, because responsibility attributions depend on the specific social rules that inspire them.

3.1. *The Consequence-based Hypothesis*

A first actual hypothesis to respond to FRH consists in denying that conscious control matters and in defining responsibility only in terms of consequences (the outcomes of actions), so that one is held responsible only on the basis of the consequences of one's own actions. We may call this the "Consequence-based hypothesis"[30]. So, even if s/he did not intend to act that way, the dangerous driver is responsible for the bad consequences s/he produces or *may* produce (otherwise we should accept that the reckless driver who doesn't hit anyone is not held responsible, something which most consequentialists would not agree to). This is a sort of*forward-looking* kind of attribution[31] and expresses a sense of responsibility that goes against the «basic desert sense»[32], which is both a backward-looking and an internal criterion of responsibility that claims that «agents are blameworthy [...] when they knowingly do wrong»[33]. This view is usually merged with a consequentialist view of punishment, which works as a justification for punishment in terms of its future beneficial effects, such as protection of the public through the prevention of future crime via the deterrent effect and containment of dangerous individuals[34]. Another justification for the Consequence-based hypothesis is its abandonment of the attitudes of moral resentment and indignation that usually go with the basic desert account, in favor of other emotions such as sadness, disappointment and sorrow, which,

[30] This view usually comes out of the spectrum of views within the free will debate, but it may be applied to control issues as well.

[31] D. Pereboom, *Living without Free Will*, Cambridge UP, Cambridge 2001; D. Pereboom. *Free Will, Agency, and Meaning in Life*, Oxford UP, Oxford 2014.

[32] D. Pereboom, *op. cit.*

[33] *Ivi*, p. 81.

[34] J.D. Greene-J. Cohen, *For the Law, Neuroscience Changes Nothing and Everything*, in «Philosophical Transactions of the Royal Society B: Biological Sciences», 359 (2004), pp. 1775-1785.

according to some[35], are more effective in discouraging misbehavior[36].

There are well-known objections to a Consequence-based account. The first is that it gives good reasons to justify social attributional practices of responsibility and punishment, but gives no criteria for understanding how to distribute them[37]. Then, basing responsibility attributions exclusively on consequences is an all-inclusive strategy that leaves no room for authorship and, in sum, leads to impersonal attributions of responsibility for actions[38], opening the way for indiscriminate attributions: whoever produces bad consequences is always responsible, no matter what the conditions that led him to act were (accidents, lack of control, ignorance, etc.). Suppose I fall on a knife that is accidentally thrown at someone causing her death, we are usually not inclined to believe I was responsible for murder, even if I was originally involved in the causal chain. Nor in the case that I donate a friend a decorative knife but she uses it to kill her husband, I am held responsible for it, because what seems to intervene in this second case is her voluntary act, while mine is missing[39]. Authorship of actions seems to count for responsibility. A third related objection regards the necessity of defining how proximate the consequences should be to the action[40]. If the agent's role in the chain is very far from the consequences, we are inclined to think that an attribution of responsibility would be unfair.

3.2. *The Neurobiological Hypothesis*

The second hypothesis comes from neurobiology, so I will call it the Neurobiological Hypothesis. According to it, one is held responsible for actions in which one exercises brain control (whether consciously experienced or not) over her neurological processes producing choices and actions. An example is Suhler and Churchland's[41] account that refers to neurobiological findings about mechanisms underlying control that could help understanding whether or not a subject's control was maintained or compromised.

[35] D. Pereboom, *op. cit.*.

[36] Shaun Nichols argues against this view. See S. Nichols, *After Incompatibilism: a Naturalistic Defense of the Reactive Attitudes*, in «Philosophical Perspectives», 21 (2007), pp. 405-428.

[37] H.L.A. Hart, *op. cit.*

[38] B. Williams, *Ethics and the Limits of Philosophy*, Fontana, London 1985.

[39] See W. Sinnott-Armstrong, *Consequentialism*, in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy (Winter 2015 Edition)*, URL = <https://plato.stanford.edu/archives/win2015/entries/consequentialism/>.

[40] H.L.A. Hart-T. Honoré, *Causation in the Law*, Oxford UP, Oxford 1985.

[41] C.L. Suhler-P.S. Churchland, *op. cit.*

The capacity to exercise control, especially self-control, and to select a particular action, is strongly related to the reward system, which provides two dissociable manifestations: deterring gratification and response inhibition[42]. Damage from trauma or disease that implies an impairment in control capacities can indicate the brain structures involved in control, such as the fronto-basal-ganglia circuit[43] and the prefrontal cortex[44], which are usually referred to as the seat of executive function. Moreover, as we can learn from addiction disorders and recidivism in addicts, neurotransmitters, hormones and enzymes (e.g., serotonin, corticotrophin, glucocorticoids, catecholamines like dopamine, epinephrine, norepinephrine) – and also genes –[45] contribute to the regulation of control mechanisms. This idea is that of the neuroscientific multilevel description of control (from genes and hormones, through neurotransmitters, to brain areas and neural networks)[46].

Objections to the Neurobiological Hypothesis are the following. There are cases where responsibility seems to apply even in the absence of full control, given the consequences of the action. If the dangerous driver who kills the pedestrian was drunk, s/he still seems to be responsible for the action. The Neurobiological Hypothesis described here seems compromised by the hierarchical idea that control coincides merely with "top-level" (executive function and frontal) areas whereas other views see control as the orchestrating of multiple unconscious brain functions, including "bottom" components like emotional structures[47]. The idea of presuming a stronger role for emotions aligns with sentimentalist views according to which moral emotions are fundamental ingredients for moral behavior[48].

[42] See also P.S. Churchland, *Touching a Nerve*, cit.

[43] A.R. Aron *et al.*, *Converging Evidence for a Fronto-Basal-Ganglia Network for Inhibitory Control of Action and Cognition*, in «Journal of Neuroscience», 27 (2007), pp. 11860-11864.

[44] E.K. Miller-J.D. Cohen, *An Integrative Theory of Prefrontal Cortex Function*, in «Annual Review of Neuroscience», 24 (2001), pp. 167-202.

[45] C.J. Ferguson, K.M. Beaver, *Natural Born Killers: The Genetic Origins of Extreme Violence*, in «Aggression and Violent Behavior», 14 (2009), pp. 286-294.

[46] See P.S. Churchland, *Moral Decision-Making and the Brain*, in J. Illes (ed.), *Neuroethics: Defining the Issues in Theory, Practice and Policy*, Oxford UP, Oxford 2005, pp. 3-16.

[47] J. Moll-R. De Oliveira-Souza-P.J. Eslinger-I.E. Bramati-J. Mourao-Miranda-P.A. Andreiuolo-L. Pessoa, *The Neural Correlates of Moral Sensitivity: A Functional Magnetic Resonance Imaging Investigation of Basic and Moral Emotions*, in «Journal of Neuroscience», 22 (2002), n. 7, pp. 2730-2736; J. Moll-R. De Oliveira-Souza, *Moral Judgments, Emotions and the Utilitarian Brain*, in «Trends in Cognitive Sciences», 11 (2007), n. 8, pp. 319-321.

[48] J.J. Prinz, *The Emotional Basis of Moral Judgments*, in «Philosophical Explorations», 9 (2006), pp. 29-43.

Another objection is the following. Just think about a case in a trial where neurobiological data might suggest that the offender was in control while performing a given offense but s/he firmly claims not to be (that s/he didn't consciously intend to do that). Should we say that neurobiology would count more? First of all, there could be even the opposite case in which the agent believes s/he was responsible (consciously in control) but neuroscience could prove s/he was not. Secondly, the scenario seems less worrying if we think that there already are cases where we hold responsible, and even culpable and punishable by a court, sane people who sincerely and convincingly claim not to have acted under conscious control, on the basis of other evidence we weigh stronger than their own words.

3.3. *The Psychodynamic Hypothesis*

The third solution to FRH I wish to consider is what I will call the Psychodynamic Hypothesis. Since I am aware that there could be different accounts of moral responsibility within this approach, I will refer to a prototypical psychodynamic account of moral responsibility here[49]. This states that we are morally responsible retrospectively even for actions we did not consciously intend if the unconsciously guided actions were in fact deeply our own (they came from our deep "selves"). We simply do not know ourselves well enough to succeed in monitoring our motivations to cause harmful behavior.

According to psychoanalysis[50], actions are guided by the activity of unconscious wishes, drives and motives (Id), which are uncontrollable by the "conscious will" or Ego. From Freud on, the psychoanalytic perspective dedicated much literature to the self-deceptive and repressive character of negative emotions like the feeling of guilt[51]. Think again of the example of

[49] See H. Fingarette, *Psychoanalytic Perspectives on Moral Guilt and Responsibility: a Reevaluation*, in «Philosophy and Phenomenological Research», 16 (1955), n. 1, pp. 18-36; E. Wallwork, *Ethics in Psychoanalysis*, in G. Gabbard-B.L. Cooper-P.W.A. Cooper, *Textbook of Psychoanalysis, 2nd Edition*, American Psychiatric Publishing, Washington (DC) 2012, pp. 349-366.

[50] E.g., the classic by S. Freud, *The Origin and Development of Psychoanalysis*, trans. in «The American Journal of Psychology», 21 (1910), n. 2, pp. 181-218.

[51] J.M. Hughes, *Guilt and Its Vicissitudes: Psychoanalytic Reflections on Morality*, Routledge, London 2008. In recent cognitive literature guilt is considered beneficial to moral behavior while most of negative outcomes for morality are attributed to the presence of shame (see J.P. Tangney-J. Stuewig-D.J. Mashek, *Moral Emotions and Moral Behavior*, in «Annual Review of Psychology», 58 (2007), pp. 345-372), although there is disagreement about how to define, distinguish and measure them (see T.R. Cohen-S.T. Wolf-A.T. Panter-C.A. Insko, *Introducing the*

the dangerous driver[52]. As killing someone by driving is an extremely sad and isolated event in her life, the driver may experience it as a foreign event and may try to find excuses for her conduct, even though s/he is usually accustomed to drive dangerously. Psychoanalytic therapy basically aims at bringing one's own unconscious functioning to consciousness, and – in psychoanalytic terms – at making the Ego gain a degree of autonomy from the Id's impulses and from the conflicts with Super-Ego prescriptions (or morality).

There are serious objections to this hypothesis and they mainly concern its general approach. The first objection is related to the psychoanalytic concept of the "unconscious", compared to the neurobiological one, where the former is unlikely to be observed, measured precisely, or manipulated easily, and it is unfalsifiable, so basically ascientific. Secondly, psychoanalytic therapy, which works with free associations, dream interpretations and various other uncontrollable techniques, turns out to be an ineffective practice for disclosing unconscious processes, which are much more likely to be determined by tools from scientific psychology and neuroscience. Thirdly, consciousness (Ego) seems still to be dominant in the psychodynamic tradition, regaining role through psychoanalytic treatment, while we have to face the possibility that consciousness might be actually ineffective by being only a mere fallible monitoring system.

The Psychodynamic Hypothesis however makes us understand that we need to clarify the concept of responsibility, with respect to the moral emotions involved, as concerns to unconscious functioning as a whole.

3.4. *The Global Traits Hypothesis*

For what I call the Global Traits Hypothesis, one is responsible for one's own actions when one's global traits can answer to the (foreseeable) consequences of her actions[53]. For global traits we may intend what commonly

GASP Scale: A New Measure of Guilt and Shame Proneness, in «Journal of Personality and Social Psychology», 100 (2011) n. 5, pp. 947-966). In the psychodynamic perspective guilt is prevalently characterized as undifferentiated from shame (with some exceptions, see G. Piers-A. Singer, *Shame and Guilt*, Thomas, Springfield (IL) 1953).

[52] G. Jervis, *Colpa e responsabilità individuale*, interview available at: http://www.emsf.rai.it/grillo/trasmissioni, 1998.

[53] I will refer to N.E. Snow, *Virtue as Social Intelligence*, Routledge, London 2010. Traces of this hypothesis can be found in M. Weber, *The Profession and Vocation of Politics*, in *Political Works*, Cambridge UP, Cambridge 1919; see also G. Jervis, *Individualismo e cooperazione*, Laterza, Roma-Bari 2002.

referred to as “character”, resulting by the combination of internal neurogenetic traits with environmental influences, and including the cognitive-behavioral expression of a complex cognitive-affective neurocognitive system (i.e., the complex interaction of capacities like reasoning, motivation and affect in the social domain)[54]. This hypothesis is a version of traditional virtue ethics, which dates back to Aristotle, was defended by David Hume and more recently by Elizabeth Anscombe. The ancient idea of “virtue” corresponds nowadays to the idea of a disposition to act determined by components that constitute personality, where personality is «conceived of as temporally stable and regularly manifested in behavior across a wide array of objectively different types of situations»[55]. According to this hypothesis, agents, encountering with situational features, can activate responses even outside of their conscious awareness, resulting in some kind of behavior we may classify as moral (or legal) or immoral (or illegal). So moral behavior is a subset of traits that constitute personality, or better behavioral regularities that cross different situation types, and these responses can be activated by triggering stimuli and influence actions even without the agent’s conscious awareness (i.e., habitual moral actions). However, habitual moral actions are not reflex reactions or automatic behavior like driving or typewriting but intelligent, flexible responses that express goal-directed actions even unconsciously[56]. They reflect the agent’s commitments and values, potentially detectable by neuropsychological indirect measures testing personality traits and implicit attitudes, like for example psychometric inventories, IAT and tests performed in neuroimaging scans[57]. They may be caused by biological factors as well as by operating conditioning, or more generally induced by environmental stimuli. The agent’s reason for acting does not need to be «present at her consciousness at the time of acting but is operative in her psychological economy» so that «we can tell a coherent story justifying the agent’s habitual virtuous [or vicious] actions

[54] W. Mischel-Y. Shoda, *A Cognitive-Affective System Theory of Personality: Reconceptualizing Situations, Dispositions, Dynamics, and Invariance in Personality Structure*, in «Psychological Review», 102 (1995), n. 2, pp. 246-268.

[55] N.E. Snow, *op. cit.*, p. 3.

[56] J.A. Bargh-P.M. Gollwitzer-A. Lee-Chair-K. Barndollar-R. Trotschel, *The Automated Will: Nonconscious Activation and Pursuit of Behavioral Goals*, in «Journal of Personality and Social Psychology», 81 (2001), pp. 1014-1027.

[57] For a complete list and description of techniques testing implicit attitudes, see N. Strohminger, B. Caldwell, D. Cameron, J. Schaich Borg, W. Sinnott-Armstrong, *Implicit Morality: a Methodological Survey*, in C. Luetge, H. Rusch, M. Uhl (edited by) *Experimental Ethics, Toward an Empirical Moral Philosophy*, Palgrave Macmillan, New York, 2014.

from a third person perspective»[58]. This is a kind of objective personality profiling[59] which does not depend on the self-reflective narrative. Imagine an irritable person whose repeated encounter with certain stimuli has triggered her biological dispositions which made her prone to irritability, without her being even aware of the way she behaves. Suppose that this person's global traits are detectable through neuropsychological measures. The example may work for the dangerous driver as well.

Objections to this view come from situationists[60], who believe that personality is fragmented and that agents' responses vary from situation to situation, and even from non-situationists, who admit personality changes (deliberative or not) over time. This gets very hard to make agents' responsibility be grounded in personality. Nevertheless, this issue is however solvable by introducing a criterion that circumscribes the assessment of the agent's cognitive-affective system functioning "at the time of acting". Such a formula is usually invoked in criminal systems in the context responsibility and insanity evaluations, but it may things more difficult as it implies we possess the kind of scientific tools to reconstruct an agent's global functioning at a time in the past. No less important, we also need to exclude that those global traits at that given time are expression of any psychiatric or neurological disease. This is another difficulty complicating the picture.

A consequence of this view is that if the event is shown to be independent from the agent's global functioning, this may excuse her from responsibility for that act or omission. As noticed above, the only internal criterion seems not to be a satisfactory criterion for responsibility attributions since also the evaluation of the consequences should be included as well as responsibility should be attributed in degrees accordingly.

4. *Merging the benefits of possible solutions within neuroethics: conclusive remarks*

I outlined obstacles and directions we should consider if we wish to

[58] N.E. Snow, *op. cit.*, p. 51; *Ivi*, p. 60.

[59] While expert testimonies that aim at assessing the offender's personality outside the context of insanity evaluations are allowed in Western countries such as the U.S., France or Germany, they are forbidden in others, for example in Italy (CPP, art. 220).

[60] E.g., G. Harman, *No Character or Personality*, in «Business Ethics Quarterly», 13 (2003), pp. 87-94; J.M. Doris, *Lack of Character: Personality and Moral Behavior*, Cambridge UP, Cambridge, 2002.

build a neuroethical account of responsibility that may respond effectively to threats deriving from the neuroscientific conception of Frail Control. Yet responsibility remains an open issue.

Although all suggested solutions appear to be defective, we may draw some important cues from the discussion to stimulate future research. Firstly, what neuroscience may help to identify are conditions for defining capacity-responsibility or control at the descriptive level, but not general conditions for responsibility normatively speaking, which are socially and culturally oriented. This means that the kind of norms or rules we shall assume as standards for responsible behavior (e.g., the speed threshold for drivers) are still locally defined and prevalently matter of convention.

Moreover, it emerged that we should consider the importance of the consequences of actions, which however seem not to be a sufficient condition per se to account for the agents' responsibility, because we need some internal criteria as well. Neurobiological and psychoanalytical accounts appear to share a relevant suggestion, which is that plausible responsibility attributions should rely somehow on unconscious processing. But more importantly, if we wish to endorse such an account this needs to be scientifically reliable (so there is no much room for psychoanalysis here), it should not forget the positive contributions of affect and emotions, and it should include more global traits than the neurobiological account actually does. I have argued that unconscious functioning should be thought of as a complex whole of the functioning of the subpersonal mechanisms within the agent ("global traits" or "moral character"), and that this whole may be conceived as the actual link between the agent (an internal criterion) and the consequences of her actions (an external criterion) to attribute responsibility to the agent in degrees. Moreover, an agent's global traits are to be intended as the organization and the interaction of multiple underlying mechanisms at various levels of biological, cognitive and behavioral description[61]. Since these interacting mechanisms determine the agent's moral response, and considering that these mechanisms operate on internal and external inputs, moral response functioning is dependent upon internal components as well as upon the environment (i.e., upbringing and education, interpersonal relationships, sociocultural factors, etc.).

I am not in the position to say if and when we will come out with reliable tools from neuroscience assessing control capacity globally in these

[61] For a multilevel perspective of cognitive neuroscience, see C.F. Craver, *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*, Clarendon Press, Oxford 2007.

terms, which could be effectively employed in legal trials. Moreover, I have mentioned but skipped a fundamental philosophical question so far. That is, we do not know how much fallible consciousness is in representing brain processes. There is a wide variability and evidence of significant differences between mindreading capacities among individuals and over their lifetimes, so we may presume that self-knowledge varies and may be improved, even though it remains fallible in itself. Nevertheless, I am not sure if and how much conscious intentions are something we should still count on in formal contexts where we retrospectively attribute responsibility to agents. Probably very low and only as a starting point of inquiry towards more reliable reconstructions[62].

Abstract

Folk ethical theories presupposed by prevailing moral theories and current legal systems tend to identify a close link between responsibility and conscious control. They generally claim that we can hold an agent responsible for outcomes of actions over which s/he exercises a certain degree of conscious control. In the last few decades, however, cognitive neuroscience has offered evidence about unconscious control processes and self-deceptive attributions of control, the so-called Frail Control Hypothesis. This hypothesis threatens the common notion of responsibility itself. I will consider possible solutions to the neuroscientific threat and discuss objections to all of them. Then, I will provide some suggestions for building a neuroethical account of responsibility that unifies the benefits of the different solutions but takes their limitations into consideration.

Keywords: responsibility; control; unconscious; neuroethics; legal.

Elisabetta Sirgiovanni
Center for Bioethics
New York University
elisabetta.sirgiovanni@nyu.edu

[62] I am very grateful to Jesse Prinz, Matthew Liao, CUNY Interdisciplinary Committee for Science Studies, Mario De Caro, Pietro Pietrini, Gilberto Corbellini and Neoclis Barone, who insightfully commented on previous drafts and conference versions of this article, and to Ben Abelson, who gave helpful suggestions on the English style. This research has been conducted at the City University of New York and funded by the National Research Council of Italy, Short-Term Mobility Program. The work inspired also early stages of my ongoing project on moral responsibility and self-control within the Fulbright Scholarship Program at the New York University Center for Bioethics.

Ethics, Law, and Cognitive Science

T

"Publicity", Privacy and Social Media. The Role of Ethics Above and Beyond the Law

Veronica Neri

Premise

One area where the relationship between ethics and the law seems to be ever more important is that of social media. There are many aspects in which this relationship touches upon dynamics which are either completely new or, at the very least, highly original. Among these, one of the main areas in need of further attention is that of the meaning, which the term "publicity" takes on within the realm of the social media. This is an area for which the limits of the law (and of deontological rules) are becoming increasingly evident. Consequently, the ethical dimension has become the most central in determining where the boundary lies between that which may be considered "public" and that which, by its very nature, is "private" and, as such, must be protected.

1. *"Publicity", a polysemic notion*

The original concept of "publicity" has also, over the course of time, taken on diametrically opposite meanings from both a semantic and from a categorical view point. The meaning adopted here is the one closest to the Latin etymology of the word, and subsequently the French *publicité*. It derives from the verb *publicare*, meaning to present something, to make something known to all or, better still, «to make public», «to occur in the presence of the public»[1]. This clearly differs from the meaning that we

[1] Meaning extrapolated both from the entry «*pubblicità*» in *Il Vocabolario Treccani* (Istituto

usually attribute to the word today, to refer to messages, which are aimed at a specific market sector.

In the realm of the World Wide Web, and in particular of social media, the verb "publish" has regained its original meaning, "to make public", that is to communicate to an indeterminate public. In the case of social media, this means to the circle of established social relations, or "friends".

On this basis, the concept of publicity evokes the alternative between that which is "public" and that which is "private", the latter being a term, which in turn calls to mind the Anglo-Saxon notion of privacy, conceived by the American legal doctrine as the «right to be let alone»[2]. This notion has become increasingly associated, in parallel with the technological development of recent decades, as the right to the protection of one's personal data against the unauthorized use by third parties. This concept can be compared with what Floridi refers to as the «informational privacy» of an individual or of a small or larger group of individuals[3].

In Italian law, the legislative decree of 30 June 2003, n. 196, sets out the «Code for data protection» in Art. 3 as follows: «information systems and programmes shall be configured to minimize the use of personal data and identification data, so as to rule out their processing if the purposes sought in the individual cases can be achieved by using either anonymous data, or mechanisms that allow identification of the person concerned, only in the case of necessity». It establishes this concept, in deontic terms, stipulating the principle of necessity in the processing of personal data.

Essentially, if each individual corresponds to their own information (obviously not in the journalistic sense, but as a set of data that contributes to revealing and creating the – virtual – identity of a particular individual), the right to privacy can thus be understood as «a right to personal immunity from unknown, undesired, or unintentional changes in one's own identity as an informational entity, both *actively* and *passively*. Actively, because collecting, storing, reproducing, manipulating, etc. [...]. Passively, because

della Enciclopedia Italiana, Roma 2003, p. 1382) as well as the entry «publicity» in the online version of the *Oxford Dictionary* (https://en.oxforddictionaries.com/definition/publicity).

[2] S. Warren-L.D. Brandeis, *The Right to Privacy*, in «Harvard Law Review», 4 (1890), pp. 193-220.

[3] Floridi also identifies a further two types of privacy which are, also in my opinion, and in the context of the present essay, in some way pertinent to the concept of informational privacy: mental privacy, that refers to protection from psychological and persuasive interference and decisional privacy, that refers to protection from procedural interference in the decision making process. Cf. L. Floridi, *The Fourth Revolution: How the Infosphere is Reshaping Human Reality*, Oxford University Press, Oxford 2014, pp. 102 ff.

[...] privacy may now consist in forcing [the individual] to acquire unwanted data, thus altering [his/her] nature as an informational entity without consent»[4]. Privacy means, however, also having the right to renew one's identity, an identity understood as the sum of the personal and the social.

In its various forms, the common feature of privacy is to highlight (in particular) the potentially negative side of "making public". Hence the need to develop adequate protection of what is private, *against* such publication. Such protection, as mentioned above, takes on a rather particular role in the realm of social media, an aspect upon which the moment for reflection has now well and truly arrived.

2. *The protection of privacy and the growing inadequacy of the law: social media as an emblematic phenomenon*

The emergence of the need for the protection of privacy as a «right to be let alone» was at the origin, on different levels, of the elaboration of a system of legal rules. Thanks to various judicial decisions, these have reached a satisfactory degree of effectiveness and equity in balancing conflicting requirements, such as – to make a paradigmatic example – the freedom of the press for journalists. The case of Italy is emblematic of how the protection of privacy has evolved: in the absence of specific legislation, the case-law progressively recognized (up until the final consecration of the Supreme Cassation Court, in its judgment of 27 May 1975, No. 2129, in the *Soraya* case) the existence of a right to freedom from intrusion into one's personal sphere. The foundations of which have been traced back to the principles stipulated in the Constitution, and, notably, in Art. 2, which recognizes the fundamental rights of the person.

The regulatory balance has, however, been undermined by technological innovations, in that the need for the protection of personal data has encountered increasing difficulties, for the law, in responding effectively to the social inputs. Reasons for this include the continual new challenges arising from of the evolution of the communication media, as well as – with the advent of the Internet – the transnational dimension of this network. This has led to a complication of the legal response, which is anchored, essentially, at national level or at the very most, continental level (the reference made is naturally to the European Union).

[4] *Ivi*, pp. 120-121.

The clearest demonstration of the difficulties afflicting the law in this field lies in the fact that the rules adopted are becoming ever more analytical and ever more extensive. Examples include the articulation of the aforementioned Legislative Decree No. 196, 2003. Also, at European level, Directive No. 95/46/EC of the European Parliament and of the Council, 24 October 1995, on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Then, more recently, the EU Regulation No. 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons was approved with regard to the processing of personal data and on the free movement of such data[5]. This is not enough, however, to ensure that the law is really able to guide the actions of the individual. Not surprisingly, more and more legal provisions explicitly rely on codes of conduct and deontological rules, entrusted with the regulation for other sectors of great importance. This is to say that the Institutions are less and less able to *impose* rules and more and more often have to rely on the autonomy of private bodies, albeit "qualified private bodies" such as, for example, the Professional Associations.

From this perspective, the domain of social media is emblematic, from at least two points of view. Firstly, it is significant that in legislative texts, even the most recent ones, such as the EU Regulation, the protection of privacy on social media is not expressly and specifically regulated: therefore, for social media the legal regulation must either be obtained from general laws or from proceeding by analogy with other regulated areas. Secondly (and more importantly, for present purposes), in the context of social media, the established dynamic of referring to "qualified private bodies" cannot possibly be effective, other than in a very partial sense: the deontological rules and legal provisions may be applied to qualified private bodies (defined as persons performing special activities, such as service providers and persons who carry out professional activities on the social media). The area of privacy protection in relation to "common users" of social media, however, remains totally "uncovered" (as demonstrated by the household exemption, i.e. the non-applicability of the EU privacy legislation to persons who process personal data without commercial purposes and within a generally circumscribed group of individuals)[6].

[5] S. Gutwirth-R. Leenes-P. de Hert (eds.), *Data Protection on the Move, Current Developments in ICT and Privacy/Data Protection*, Springer Netherlands, Dordrecht 2016.

[6] P. Passaglia, Privacy *e nuove tecnologie, un rapporto difficile. Il caso emblematico dei* social media, *tra regole generali e ricerca di specificità*, in «Consulta *Online*», 3 (2016), pp. 338 ff.

This lack of legal and deontological cover in the case of the "common user" is particularly serious in light of the possibility (confirmed almost daily by the press) that those very users can be at the origin of major privacy violations. The critical issues that emerge can take one of two partially diverse forms, at least from a legal viewpoint, since there are two fundamental dimensions (A and B – below) upon which the action of the user can be differentiated.

First and foremost, because users of social media, upon entering the virtual community, automatically waive a share of their privacy (A). On this point, it is widely felt that there is a need to prevent individuals from giving rise to excessive waivers. However, a "protective" legislation would appear difficult to draw up, because the very fact of limiting the possibility of a self-regulation of the individual concerning his/her privacy runs the inevitable risk of being perceived as a limitation of the freedom of self-determination of the individual. Ultimately, therefore, as an attack on one of the cornerstones upon which the rule of law and the Liberal democracy rely. It follows that the law may intervene, generally if, and only if, there are good reasons to limit self-determination, particularly if other aspects come into play (for example, on the grounds of public safety). Even beforehand if the self-determination cannot be considered valid, as in the case of minors and persons who have been legally declared not competent.

The tension between privacy and social media, however, does not apply only in the perspective of self-regulation of the law: the "common users" of social media, although not subject to any legal and ethical constraints in terms of privacy, can actually cause serious damage to the privacy of others (B). In theory, the law could intervene in this type of conduct; however, a problem of effectiveness arises, since it is very difficult to "attack" social media behaviour effectively and without veering towards a politics of censorship[7].

Therefore, what emerges is that, with reference to the "common users" of social media, the law is sometimes (A) unable to intervene, whilst at other times (B) suffers from an incipient ineffectiveness. Such deficiencies cannot be remedied by deontological rules, for the simple fact that, since

[7] S. Di Guardo-P. Maggiolini-N. Patrignani (a cura di), *Etica e responsabilità sociale delle tecnologie dell'informazione. Etica e Internet*, 2, FrancoAngeli, Roma 2010, pp. 252-256. For a discussion on ethical and legal issues regarding *privacy* on the Internet, see also J. Berleur-P. Duquenoy-D. Whitehouse (eds.), *Ethics and the Governance of the Internet*, IFIP Press, Laxenburg 1999, pp. 38-53; J. Berleur, *Questioni etiche per la governance di internet*, in S. Di Guardo *et al.*, *op. cit.*, pp. 259-274.

we are dealing with "common users", those who do not engage in professional activities on the social media, these rules have no possibility of application. So ultimately, the task of behavioural guidance regarding social media can only be assigned to the field of ethics.

3. *Privacy and social media: ethics, the last fortress*

Any investigation, from an ethical view point, into the problems connected to privacy on social media, has to start with the process of the spectacularization, the "showcasing" of one's existence. This pervades contemporary society and is increasingly focused on a radical visibility[8]. This process has already been theorized by Debord (1967), who, prophetically, asserted that «reality emerges within the spectacle, and spectacle is real. This reciprocal alienation is the essence and support of the existing society»[9]. After all, «What appears is good; what is good appears. [...] the spectacle is *leading production* of present-day society»[10]. In this «Age of Access» we continue to use the same metaphor of the stage *à la* Debord, albeit an electronic stage in these modern times, upon which, Rifkin writes, we observe an alternation, in real time, of individual representations[11]. A stage which opens up to multiple personalities, «powerful metaphor for thinking about the self as a multiple, distributed system»[12].

[8] There have been some recent proposals which suggest tighter controls on published information, also through stricter access limitations. Cf. H.T. Tavani, *Philosophical theories of privacy: Implications for an adequate online privacy policy*, in «Metaphilosophy», 38 (2007), n. 1, pp. 1-22; Id., *Ethics and Technology: Controversies, Questions, and Strategies for Ethical Computing*, John Wiley & Son, New York 2011. Others, instead, propose the control of «contextual integrity», in relation to the distribution, appropriacy and pertinence of the information. H. Nissenbaum, *Privacy as Contextual Integrity*, in «Washington Law Review», 79 (2004), n. 1, pp. 119-158; Ead., *Privacy in Context: Technology, Policy, and the Integrity of Social Life*, Stanford University Press, Palo Alto (CA) 2009.

[9] G. Debord, *La société du spectacle*, Buchet-Chastel, Paris 1967, engl. transl. by K. Knabb, *The Society of the Spectacle*, Bureau of Public Secrets, Berkeley 2014, § 8.

[10] *Ivi*, §§12, 15.

[11] J. Rifkin, *The Age of Access: The New Culture of Hypercapitalism, where All of Life is a Paid-for Experience*, Jeremy P. Tarcher/Putnam, New York 2000, pp. 214-215.

[12] S. Turkle, *Life on the Screen: Identity in the Age of the Internet* (1995), Simon and Schuster, New York 2011, p. 14; E. Goffman, *The Presentation of Self in Everyday Life*, University of Edinburgh Social Sciences Research Centre, Edinburgh 1956. However, there are those, like Baudrillard, who sustain that we have already gone beyond the stage of the spectacle: since there is no distinction between public and private, individuals have become the recipients of a plurality of communication networks (J. Baudrillard, *L'autre par lui-même*, Paris, Galilée 1987, engl. transl. by B. Schütze, *Ecstasy of Communication*, Semiotext(e), New York 1988).

The spectacle thus appears to be the beginning and end of the communication on social media: the self is constructed and develops a sense both *in* and *through* relations with others, through accessing, or otherwise, the private universe of others and certain social media. This process of the spectacularization of the self has made the social media an emblematic place of what can only be described as «social showcasing»[13]. Putting oneself on display means also to expose one's private sphere, and as such, to risk having it turned into a commodity. Furthermore, this may result in the fuelling of dysfunctional behaviour, and not only on the part of corporations engaged in online marketing[14].

The amplified exposure of oneself has an immediate impact on the first type of relation outlined in the previous paragraph (A), that is, relative to the person entering his own data: it lowers the "warning threshold" of the individual, who is willing to publish his own data so that he may appear in the "showcase". In parallel, it is evident that there is much encouragement on the part of the *social platforms* towards users to share information, which is either personal, or relating to other people and entities[15]. These are voluntary, albeit imprudent, practices due partly to lack of information or misinformation[16].

Is it, therefore, always good or always bad to enter one's own personal data? Where should one draw the line? Or is everything reduced to a mere waiver of privacy protection?

First and foremost, it is necessary to get away from a vision based primarily on the aesthetics of the staged spectacle. We need to move, if anything, towards an ethical representation. This affirmation means that we need to assume that the surrendering of privacy, *hic et nunc*, could lead to

[13] V. Codeluppi, *Ipermondo*, Laterza, Roma-Bari 2012, pp. 84-97; Id., *La vetrinizzazione sociale. Il processo di spettacolarizzazione degli individui e della società*, Bollati Boringhieri, Torino 2007.

[14] N. Abercrombie-B. Longhurst, *Audiences: A Sociological Theory of Performance and Imagination*, Sage, London-Thousand Oaks 1998.

[15] Of interest, and confirmation of the ease with which individuals tend to publish even the most distinguishing of data, are the cases which are analyzed in C. Rizza *et al.*, *Interrogating Privacy in the Digital Society: Media Narratives after 2 Cases*, in «International Review of Information Ethics», 16 (2011), pp. 6-17; A. Acquisti-R. Gross, *Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook*, in P. Golle-G. Danezis (eds.), in «Proceedings of 6th Workshop on Privacy Enhancing Technologies», Robinson College, Cambridge 2006, pp. 36-58.

[16] S. Vallor, *Social Networking and Ethics*, in E.N. Zalta (eds.), «The Stanford Encyclopedia of Philosophy» (2016): https://plato.stanford.edu/archives/win2016/entries/ethics-social-networking/.

a permanent future loss of control regarding certain information (both due to, and thanks to, the continual availability of such information). This waiver may also have repercussions on other individuals, those who have not chosen to be visible in this virtual *agorà*, and certainly not to be subjected to the spectacle of their own lives. It is the case of parents who publish images of their children, who, as adults, may not feel represented by an identity built online without their consent or, indeed, disagree on principle with the spectacularization of their lives.

Finally, one must take into account that this waiver can be used by others to spread our data (voluntarily and/or involuntarily), perhaps even in a distorted way. Furthermore, via channels other than those originally chosen by us, without any prevision of the consequences that may result in our off-line day-to-day lives. Alternatively it could result in individuals entering information about others, thus creating multiple (and often false) identities.

Here we invite reflection on issues related to the second relational type mentioned above (B). Serious damage can be caused thereby, to the privacy of third parties. This relational type is more difficult to control, due also to its widespread use. Such a dimension, therefore, has inevitably to be restricted by ethical considerations.

Two key aspects emerge in particular (the first one fundamental to the existence of the second): the autonomy of technology and the mutability of the identity of an individual. Regarding the autonomy of technology, it should be noted that data change their ontological status, once inserted into the social media: what was once static information turns into autonomous agents (which is true for the internet in general). In some respects, this information can also turn into moral agents, since these can produce real consequences that could be qualified from an ethical point of view[17]. This data can develop in any direction and acquire a meaning that is different from the original one. In the wake of what has already been mentioned by Anders in relation to the artificial man-made devices produced during the Second World War, one must take into account that, also as aware and informed users of technology, individuals can still, in spite of themselves, become instruments of this same technology (even) against their will[18]. Similar fears, moreover, have been expressed by Jonas, according to whom one must take into account the responsibility on the

[17] L. Floridi, *op. cit.*, pp. 101 ff.

[18] G. Anders, *Die Antiquiertheit des Menschen, I: Über die Seele im Zeitalter der zweiten industriellen Revolution*, Verlag C.H. Beck, München 1980.

shoulders of today's individuals when working on technological developments for the future generations[19].

Furthermore, it is because of the autonomy of this technology which «creates itself»[20] that identity becomes mutable. This identity allows space for potential selves, which may even be turned into something different from the original self[21]. The most significant aspect from an ethical point of view, therefore, calls for quality, and in particular preciseness, of the entered data: if one publishes, deliberately or otherwise, false data regarding oneself, or if data is tampered with by another individual, a chain of misunderstanding and distorted information can ensue. This can potentially cause serious damage to third parties in their off-line day-to-day lives.

These changes can be implemented through subtle strategies. Counter-images of the self may be introduced, playing upon the ambiguity of certain data and the levity with which this data can sometimes be "shared" on the social platforms, which consequently receive and often redirect the data. Moreover, unlike in offline relations, the information or disinformation exchanged remains forever indelible in "cyber-memory". Furthermore, there may be an overlap between the various identities present on the social media. «Egocentric» communications about the self-become, involuntarily, «allocentric»[22]. Thus, from a self-presentation of the persona, a hetero-produced presentation can derive. Consider, for example, the tagging phenomenon, through which you can attach photos or texts to a person, without their prior consent (when, due to lack of experience of the system, the person has not asked for any notification and is therefore unaware). Even more subtly, a self-presentation can be used to form and convey an impression of a person, which is only slightly different from how they actually are offline. The author of the profile himself, or on the part of other "friends" may do this either. In the first instance, control is lost regarding exactly what is being disseminated. Furthermore, particularly in the latter case, (apparently) imperceptible changes are carried along, through a process which has a concrete impact, even offline, in terms of public access to our personal

[19] H. Jonas, *Das Prinzip Verantwortung. Versuch einer Ethik für die technologische Zivilisation*, Suhrkamp, Frankfurt am Main 1979, engl. transl. by H. Jonas and D. Herr, *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*, University of Chicago Press, Chicago 1984, pp. 92-93.

[20] A. Fabris, *Etica delle nuove tecnologie*, La Scuola, Brescia 2012, p. 55.

[21] K.J. Gergen, *The Saturated Self: Dilemmas of Identity in Contemporary Life*, Basic Books, New York 1991, p. 79.

[22] G. Riva, *I social network* (2010), Il Mulino, Bologna 2016, p. 27.

data[23]. The data recorded by the system are then used for advertising purposes (in the commercial sense) or news (information) targeted to a specific audience. These incentivate the individual to make purchases or to take interest in certain issues that perhaps would never have otherwise come to his attention. These aspects lead to continual comparisons with other individuals present on the social media. It creates a tendency to continually go one-step further in order to increase the number of relations (both strong but, in particular, weak). This happens in an undifferentiated context in which misunderstanding can be both frequent and dangerous, due partly to the large quantity of potentially publishable data.

Technological autonomy, therefore, raises yet another aspect, related to the handling of data which, if detrimental to an individual image, should never be used: one thing is to share certain information with "friends", quite another is to have it shared with "friends of friends", who in turn can forward the information to other "friends", and so on.

What, therefore, can be the motivation that draws us to behave in a way, which is detrimental both to our own privacy and to that of others (A and B)? In light of the above, the motivation behind certain spectacularization behaviour could lie in the human tendency of individuals (as pointed out by Riva) to desire an escape from anonymity, as well as a longing for personal recognition, combined with a necessity to meet those needs linked to relationships, self-esteem and self-actualization that Maslow places on the last steps of his famous pyramid[24]. However, these needs may become satisfied in a distorted way, or endanger both the subject himself as well as third parties.

Hence, everything can be linked to self-promotion deriving from a tendency towards narcissism, which can give rise to the (un)conscious transformation of one's image, into valuable goods. Creating intermediate spaces of "inter-reality", of «in-betweenness»[25], in which the boundary between public and private is increasingly less perceptible[26]. What is more, the desire to 'appear' seems sometimes to distract us from the ethical implications that can result from certain decisions.

[23] J. Palfrey-U. Gasser, *Born Digital. Understanding the First Generation of Digital Natives*, Basic Books, New York 2008, p. 42.

[24] A.H. Maslow, *A Theory of Human Motivation*, in «Psychological Review», 50 (1943), n. 4, pp. 370-396.

[25] L. Floridi, *op. cit.*, p. 25.

[26] J. Van Kokswijk, *Hum@n, Telecoms & Internet as Interface to Interreality*, The Netherlands: Bergboek, Hoogwoud 2003.

4. *The "last fortress" of ethics and the strengthening of the buttresses*

On the basis of this argument, there is the clear need for a rethinking both of how we act *on* social media as well as in our approach *to* social media.

So how can ethics guide us in the publication of our information, avoiding possible risks (A and B)? Upon which principles should we base our decisions in order not to harm our own privacy or that of third parties?

In reply to the first question, it should be taken into consideration that poor computer skills can result in a lack of control of the data entered. But even if digital skills were promoted, would it actually contain the problems that have emerged regarding privacy? Individuals should certainly know how social media function before using them, both from a technological and an operational point of view. Riva, for example, with reference to their use by minors, suggests the introduction of a license, just as for driving a car[27]. However, focusing principally on these skills is perhaps merely a shift back in the legal-deontological direction. Most probably, the social media, in view of their very ontological status, need to be conceived on an ethical basis, i.e. respecting the privacy of the individual, easy-to-use, transparent regarding the behavioral rules to be adhered to, where failure to comply could result in the degeneration and even the breakdown of a relationship.

Could then the answer be a responsible campaign to raise awareness regarding the appropriate use of social media? Certainly the one promoted by the Italian Data Protection Authority, based on a guide to social media (2009 and re-published in 2014), was aimed at promoting reflection on the meaning and the consequences of individual and collective action in the virtual *agorà* of the social media[28]. Particular attention is paid to the respect of the privacy of the individual. Nevertheless, the social media continue to be perceived and used as (pseudo) private spaces.

Neither of the afore-mentioned proposals, though indispensable in terms

[27] G. Riva, *op. cit.*, p. 167.

[28] Cf. http://194.242.234.211/documents/10160/10704/Opuscolo+Social+Network+pagina+singola.pdf: «with the objective of raising the awareness of users and providing them with food for thought as well as the tools for their own safeguard» (*ivi*, p. 3). This document is divided into a series of «warnings for internet surfers» (*ivi*, pp. 9 ff.) as well as questions to stimulate the self-responsabilization of the reader (*ivi*, p. 17), concluding with «10 tips on how not to get caught up in the trap» (*ivi*, pp. 23 ff.) and a glossary of the slang terms most commonly used on the web (*ivi*, p. 31). P. Galdieri, *Il trattamento illecito del dato nei* social network, in «Giurisprudenza di Merito», 44 (2012), n. 1, pp. 2697 ff.; P. Passaglia, *op. cit.*, pp. 345 ff.

of raising awareness, shields us completely from the risk of infringement of privacy, of losing control both of the published information and of the technological tool itself. It is not always clear at what point one should stop in order not to violate the other's space: without face-to-face interaction, the empathy and emotional openness that facilitate "good" communication are not developed. The «actualization» of a process, as Levy defines it, is intended as its occurrence and resolution in a space other than that of the network (in this case social)[29]. It is, in this sense, a litmus test of the performative scope of certain actions performed both *for* and *on* the social media. We run the risk, when not acting responsibly, of creating and encouraging a relational illiteracy. Considering the number of possible contacts, the qualitative aspect is neglected, since one is able to conceal emotional discomfort behind the construction of a certain virtual visibility. In online social relations, the signals transmitted by other channels are not present; significant and cognitive signals which are equally important in understanding the sense of the communicative exchange in all its complexity. Moreover, individuals seem less conscious of their online actions – almost as if they hadn't actually performed them – actions which offline they would never dream of carrying out.

So, which principles can guide our actions on social media? Without doubt we must return to a full restoration of the concept of responsibility, both in terms of what is done in relation to oneself and to others[30]. Furthermore, we must be answerable, above all, for the correctness and truth of the information conveyed, but also the authenticity of the exchange, which must be aimed at promoting understanding[31]. This requires our adaptation to a system, which, though at the outset showed only its positive aspects, has now also revealed its more negative side. It would appear the moment has arrived to attempt a re-semantization of the concept of publicity, which no longer means to make public to a select and limited group, rather to a potentially infinite public, and for a potentially infinite period of circulation. Equally, a re-semantization of the notion of privacy is required, to embrace a new meaning of the concept of private. Fundamental in this medial universe where the spaces appear indeterminate and ambiguous, due both to lack of knowledge, but also due to the ontological sta-

[29] P. Levy, *Qu'est-ce que le virtuel?*, Éditions La Découverte, Paris 1995, p. 15.

[30] A. Fabris, *Etica della comunicazione* (2006), Carocci, Roma 2014, pp. 47-51; M. Vergani, *Responsabilità. Rispondere di sé, rispondere all'altro*, Raffaello Cortina Editore, Milano 2015.

[31] D.M. Boyd-N.B. Ellison, *Social Network Sites: Definition, History, and Scholarship*, in «Journal of Computer-Mediated Communication», 13 (2007), n. 1, pp. 210-230.

tus of the social media. Speaking of privacy on social media could seem a contradiction in terms: private data is no longer something to be safeguarded *tout court*, rather something to be conveyed, albeit most certainly in a more aware and informed way. That which is considered fine to be made public, or otherwise, varies over time, in relation to society and to the individual himself[32]. So ethics must absolutely play a role in this (necessary) re-definition, especially if, particularly among young people,

> 'privacy' is not a singular variable. Different types of information are seen as more or less private; choosing what to conceal or reveal is an intense and ongoing process [...]. Rather than viewing a distinct division between 'private', young people view social contexts as multiple and overlapping. [...] Indeed, the very distinction between 'public' and 'private' is problematic for many young people, who tend to view privacy in more nuanced ways, conceptualizing Internet spaces as 'semi-public' or making distinctions between different groups of 'friends' [...]. In many studies of young people and privacy, 'privacy' is undefined or is taken to be an automatic good. However, disclosing information is not *necessarily* risky or problematic; it has many social benefits that typically go unmentioned[33].

The responsibilization (and awareness-raising) of individual users, which is currently the only real option on the part of Institutions and corporations, must be founded upon the new meaning that "make public" has taken on in the world of social media. Before the advent of the internet, to "make public" required mediation. Now anyone can transmit or transform information, highlighting certain aspects rather than others. Although nowadays, to "make public" on social media means to convey a radical transparency, at the same time, this transparency may be rendered opaque to the point of it taking on its own hue. This is what is happening, for example, even in the field of journalism with the "post-truths", about which so much has been written[34].

[32] Moreover, according to Acquisti, Brandimarte and Loewenstein, trasparency and control alone are not enough: «To be effective, privacy policy should protect real people – who are naïve, uncertain, and vulnerable – and should be sufficiently flexible to evolve with the emerging unpredictable complexities of the information age». Cf. A. Acquisti-L. Brandimarte-G. Loewenstein, *Privacy and Human Behavior in the Age of Information*, in «Science», 347 (2015), n. 6221, pp. 513-514.

[33] A.E. Marwick-D. Murgia-Diaz-J.G. Palfrey Jr., *Youth, Privacy and Reputation (Literature Review)*, Berkman Center Research Publication No. 2010-5, Harvard Public Law Working Paper No. 10-29, p.13, available at https://ssrn.com/abstract=1588163

[34] M. Del Vicario *et al.*, *The Spreading of Misinformation Online*, in «Proceedings of the National Academy of Sciences», 3 (2016), n. 113, pp. 554-559.

The manipulability of «artificial nature» can reveal extremely ambiguous aspects. «Artificial nature» is a term composed of two seemingly counterposed words. This, however, is not the case if we consider that the very aim of modern technology is to render artificial that which is natural[35]. Although it is a positive thing that social media should affirm areas for freedom of expression, such freedom must appeal to the co-responsibility of all the players involved. It is a responsibilization that must necessarily develop through (self-)limitation. However, intervention is required that could somehow limit our options, albeit responsibly, in order to develop a project of social participation, whilst maintaining a space of mutual respect[36]. A classic example of self-limitation is the need to protect the weakest members of society. This is particularly the case with minors, who run the risk of having their images circulated in a potentially uncontrolled fashion. These photos may even become the object of serious, often criminal, abuse. On this note, it is worth highlighting a recent initiative in Germany. A new Facebook page was opened up, dedicated entirely to the compromising photos of minors, which parents themselves had imprudently published on social media.

«Responsible freedom»[37], therefore, that takes account, not so much of the intentionality, as of the imputability of our choices and the consequences that may result, both inside and outside the social *agorà*. This could represent an opening towards the type of protection that, by the very nature of social media, the law is only able to offer up to a certain point, leaving the field of ethics with ample room for reflection. This would ensure that relations are established and maintained, which do not deviate into disinterest and indifference, but come back to the constitutive sense of the social media, a network for socializing, sharing, participation and connection in real-time[38]. These relations should be impressed upon all

[35] A. Fabris, *Etica delle nuove tecnologie*, cit., pp. 38-41.

[36] On this point, Vallor's perspective is of interest. She establishes an ethical behaviour on the *social networks* based on three "virtues", namely patience, honesty and empathy. S. Vallor, *Social Networking Technology and the Virtues*, in «Ethics and Information Technology», 12 (2010), n. 2, pp 157-170.

[37] A. Fabris, *Etica e internet*, in S. Di Guardo-P. Maggiolini-N. Patrignani (a cura di), *Etica e responsabilità sociale delle tecnologie dell'informazione. Etica e Internet*, 2, FrancoAngeli, Roma 2010, pp. 185-199, 196; V. Cesareo-I. Vaccarini, *La libertà responsabile. Soggettività e mutamento sociale*, Vita & Pensiero, Milano 2009.

[38] On the themes of indifference and virtual relations: A. Fabris, *Etica del virtuale*,Vita & Pensiero, Milano 2007, pp. 12 ff.; Id., *RelAzione. Una filosofia performativa*, Morcelliana, Brescia 2016, pp. 164 ff.

social media users, with no distinction, on an ethical level, between what, conversely, for the law is a very significant aspect: the renunciation of one's own privacy (through data input on the part of the individual himself) (A), and the utilization of data published by others (B). Thus privacy ought to remain, and as Floridi also asserts, albeit in a partial reattribution of the meaning, «should be considered a fundamental right and hence that, as for other fundamental rights, by default the presumption should always be in favour of informational privacy»[39].

Abstract

Nowadays social media play an increasingly important role in the relationship between ethics and the law. They have raised new issues regarding the concepts of both "publicity" (in the etymological sense of "making public"), and privacy. The limits of both the law and of deontology are becoming more and more evident, in this arena of the relations, which are established through the social media. This aspect implies the need for ethical reflection, focusing on the motivation that leads users to convey certain information – in primis *the desire for a spectacularization of one's life – as well as on the possible principles that may help guide informed choices. Among these would appear fundamental a reference to the concept of 'responsible freedom', and hence to the possible consequences which may arise as a result of certain choices, consequences both for oneself and other individuals, on social media as well as in our off-line day-to-day lives.*

Keywords: ethics; law; privacy; publicity; responsibility; social media; spectacularization.

Veronica Neri
Dipartimento di Civiltà e Forme del Sapere
Università di Pisa
veronica.neri@cfs.unipi.it

[39] L. Floridi, *An Interpretation of Informational Privacy and Its Moral Value*, in «Proceeding of CEPE 2005, 6th Computer Ethics: Philosophical Enquiries Conference, Ethics of New Information Technologies», The Netherlands: University of Twente, Enschede 2005.

Edizioni ETS
Piazza Carrara, 16-19, I-56126 Pisa
info@edizioniets.com - www.edizioniets.com
Finito di stampare nel mese di novembre 2017

I saggi dell'annata 2017 di "Teoria" sono stati valutati
in *double-blind peer review* da:

Andrea Lavazza
Centro Universitario Internazionale, Arezzo

Fabio Paglieri
ISTC-CNR Roma

Laura Palazzani
Università Lumsa

Giulia Piredda
Istituto Universitario di Studi Superiori di Pavia

Federico Gustavo Pizzetti
Università degli Studi di Milano

Maria Grazia Rossi
Università Cattolica del Sacro Cuore

Jennifer Ware
Wright State University Dayton

TEORIA

T

Rivista di filosofia
fondata da Vittorio Sainati

Ultimi fascicoli apparsi della Terza serie di «Teoria»:

XXXVII/2017/1 (Terza serie XII/1)
Linguaggio e verità / Language and Truth

XXXVI/2016/2 (Terza serie XI/2)
Etiche applicate / Applied Ethics

XXXVI/2016/1 (Terza serie XI/1)
New Perspectives on Dialogue / Nuove prospettive sul dialogo

XXXV/2015/2 (Terza serie X/2)
Relazione e intersoggettività: prospettive filosofiche
Relación e intersubjetividad: perspectivas filosóficas
Relation and Intersubjectivity: Philosophical Perspectives

XXXV/2015/1 (Terza serie X/1)
Soggettività e assoluto / Subjectivity and the absolute

XXXIV/2014/2 (Terza serie IX/2)
«Ripensare la 'natura' – Rethinking 'Nature'
2. Authors and Problems/Figure e problemi»

XXXIV/2014/1 (Terza serie IX/1)
«Ripensare la 'natura' – Rethinking 'Nature'
1. Questioni aperte/Burning Issues»

XXXIII/2013/2 (Terza serie VIII/2)
«Hope and the human condition – Speranza e condizione umana»

XXXIII/2013/1 (Terza serie VIII/1)
«Hegel. *Scienza della logica*»

XXXII/2012/2 (Terza serie VII/2)
«Spinoza nel XXI secolo»

XXXI/2012/1 (Terza serie VII/1)
«Conformity and Dissent - Conformità e dissenso»

XXXI/2011/2 (Terza serie VI/2)
«La formazione e la conoscenza ai tempi del web»

Questo fascicolo di «Teoria» si propone di prendere in esame le ricadute della scienza cognitiva dell'etica su una varietà di temi di metaetica, etica normativa, etica applicata e filosofia del diritto. Sono state dunque indagate criticamente le scoperte della neuroscienza cognitiva concernenti la responsabilità morale e legale.

The purpose of this issue of «Teoria» is to explore the relevance of the cognitive science of morality for a variety of topics in metaethics, normative ethics, applied ethics, and philosophy of law. In particular articles are concerned with how recent cognitive science findings affect our practices of attributing moral and legal responsibility.

Scritti di: Mario De Caro, Massimo Marraffa, Daniel C. Dennett, Felipe De Brigard, Lacey J. Davidson, Benedetta Giovanola, Rossella Guerini, Andrea Lavazza, Uwe Peters, Simone Pollo, Massimo Reichlin, Maria Grazia Rossi, Daniela Leone, Sarah Bigi, Elisabetta Sirgiovanni, Veronica Neri.

ISSN 1122-1259

€ 20,00

ISBN-13: 978-8846750716
9 788846 750716