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.