Philosophy and
Phenomenological Research

Philosophy and Phenomenological Research
doi: 10.1111/phpr.12537

Agency, Teleological Control and Robust Causation

MARIUS USHER
Tel-Aviv University

I propose a compatibilist theory of agency and responsibility, according to which an agent is responsible for an effect, if and only if, she is the earliest source of robust causation over it, via an action she carried out in the service of her long term interests. This theory deploys a notion of teleological control, which is a type of guidance-control of the agent over the effect and it involves action plans and means-end reasoning. The theory makes room for degrees of responsibility, and accounts for the distinction between compulsion and determination. The teleological control view is informed by neuroscience and cognitive theory, and while it is indifferent to the distinction between determinism and indeterminism, it contends that the property of natural laws relevant to agency is the presence of successive stages of attractor and bifurcation dynamics. While the former grounds robust causation over effects of actions, the latter limits the temporal range of robustness, allowing us to characterize responsibility in terms of the earliest sources of robust causation.

Compatibilists claim that agency and the type of responsibility involved in attributions of praise or of blame are compatible with determinism about the laws of nature – where determinism is, roughly, the thesis that the events of the past and the laws of nature permit only one physically possible future. A strong defense of this position against those who argued that determinism rules out the ability to do otherwise was mounted by Harry Frankfurt, who described scenarios that strongly suggest that the principle of alternative possibilities is not a necessary requirement for moral responsibility (Frankfurt 1969). In a Frankfurt type counter-factual intervener scenario, an agent is intuitively responsible even though she could not have done otherwise.[1]

A crucial task for compatibilism is to distinguish between events and actions for which an agent is morally responsible, and thus subject to praise or to blame, and those for which she is not. Following Frankfurt's argument, most compatibilists have developed source compatibilist theories (see Frankfurt 1971, Fischer and Ravizza 1998, Sartorio 2016a), in which the agent is the source from which her actions emanate and the responsibility for outcomes of actions is grounded on the actual-sequence of events leading to the action. Recently, however, a number of criticisms have pointed to difficulties

[1] For example, Black monitors Jones' decision and is ready to intervene to make her decide to A rather than B, if Jones shows any sign of deciding to B. In the actual case, Jones decides to A on her own (without any sign of B-ing) and Black does not intervene. As Black plays no causal role in Jones' action or decision, in the *actual-sequence* of events, the common intuition is that Jones is responsible for A-ing, although she could not have done or decided otherwise (Frankfurt 1969).

that beset prominent compatibilist attempts to ground responsibility on the actual sequence (Sartorio 2016b, section-3). First, source compatibilism has been challenged to account for the diminished responsibility (or lack of *ultimate* responsibility) of agents whose actions are subject to manipulation or to covert (non-constraining) control (Kane, 1996, ch. 5), and it has been argued that compatibilism cannot withstand the pressures from both remote (constitutive) and present luck (Levy 2009, 2012). Second, compatibilist theories need to answer the challenge from the teleological character of action, which was eloquently described by Scott Sehon: "Notions like action and goal-direction appear to have no role in purely physical descriptions of the world. Planets, rocks and elementary particles do not do things; if we are no different in principle than these things, then our status as agents who do things can be legitimately drawn into question." (Sehon, 1998: 197). Third, compatibilistic theories would gain further traction by accounting for the variation in the *degree* of responsibility – as reflected in attributions of praiseworthiness and blameworthiness (Nelkin 2017), which accompanies typical human judgments (Feltz 2013, Lagnado and Gerstenberg 2015, Lombrozo 2016, Murray and Lombrozo 2016, Sripada 2012).

Here, I present a novel approach for articulating the control condition for moral responsibility within a source compatibilistic approach.[2] I rely on the notion of robust causation, developed by James Woodward (2006) and also on a notion that I call *teleological-control*. I argue that this account addresses the manipulation and luck problems, by distinguishing between "normal" and manipulated (or indoctrinated) agents, in terms of sources of robust and teleological causation that cannot be traced back any further. A central aim of this account is to characterize the degree of responsibility that agents are often deemed to have for the effects of their actions, in terms of the robustness with which those effects are brought about. Finally, I maintain that while this account is indifferent to the metaphysical distinction between determinism and indeterminism, it does rely on a different distinction: *attractors* vs. *bifurcations* – two types of dynamical entities that are compatible with physical (and, in particular, deterministic) laws. I argue that this distinction grounds the notions of agency and responsibility and that it is well supported by experimental and computational research in psychology and behavioral neuroscience.

I begin with challenges posed by manipulation and luck for compatibilist theories of moral responsibility. Following (sections 2-3) I present my *teleological-control/robust-causation* account of the degree of responsibility an agent has for the outcomes of her actions. Then (section 4) I discuss how this account stands in relation to laws of nature, and I refine it to characterize responsible agents as the earliest sources of robust causation. Finally, I argue that it accounts for reduced responsibility of manipulated agents and to the problem of luck (section-5).

1. The problem of manipulation and the "luck pincer"

1.1 Manipulation scenarios, in which agents are influenced to adopt someone else's values (without coercion, but say, by positive reinforcement), have been argued to raise a difficult challenge for compatibilist theories. For example, Robert Kane (1996, ch. 5)

[2] While it is beyond of aims of this paper, it is possible that this analysis can also help to consolidate a compatibilistic account of Free-Will (see e.g., Deery and Nahmias, 2017).

argues that compatibilism has a difficulty explaining the difference between a person, who was determined to adopt (without coercion) the values of an indoctrinator, and the rest of us, whose values are determined by the laws of nature operating in a deterministic world. A vivid illustration of this difficulty was more recently presented by Derk Pereboom, who challenged compatibilists to distinguish, in terms of agency and responsibility, between four manipulation scenarios that vary in their degree, starting from a clear case of manipulation (case-1) and gradually progressing towards cases that appear to involve ordinary action (case-4). In each of the four cases, Professor Plum kills Ms. White for the sake of some personal advantage. I summarize these cases briefly below (with some adaptation; Pereboom 2001, 2005: 22-28).

Case-1: Professor Plum is manipulated by neuroscientists, who directly affect his mental states from birth via radio technology by making his reasoning rationally egoistic, and still sensitive to reasons. They do not implant *irresistible* desires and Plum remains receptive to relevant patterns of reasons, so that his reasoning should result in different choices in some situations, in which the egoistic reasons are otherwise. Let us assume that the controllers slightly, but consistently, increase the weight of the egoistic reasons in all his deliberations.

Case-2: Professor Plum is like an ordinary human being, except that he was created by neuroscientists who, although they cannot control him directly, have programmed him to weigh reasons for actions, so that he is often, but not exclusively, rationally egoistic.

Case-3: Professor Plum is an ordinary human being, except that he was determined by rigorous training practices of his home and community, so that he is often, but not exclusively, rationally egoistic.

Case-4: Professor Plum is an ordinary human being, subject to physical laws, which happen to be deterministic. However, while being raised under normal circumstances, he is often, but not exclusively, rationally egoistic.

In all of these cases, Professor Plum kills Ms. White as a result of his reasons-sensitive egoistic judgment. Pereboom's challenge to the compatibilist is to identify the border point on the slippery slope (from case 1-4), where the responsibility for the killing of Ms. White emerges, and to point out the crucial difference that makes the agent responsible on one side of the border and not responsible on the other (2005: 23).

Another recent manipulation challenge for the moral responsibility of agents in a deterministic world, is Alfred Mele's zygote argument, which moves the manipulator's actions before the birth of the agent. "Diana – a Goddess – creates a zygote Z in Mary. She combines Z's atoms as she does because she wants a certain event E to occur 30 years later. From her knowledge of the state of the universe just prior to her creating Z and the laws of nature of her deterministic universe, she deduces that a zygote with precisely Z's constitution located in Mary will develop into an ideally self-controlled agent (Ernie) who, in 30 years, will judge, on the basis of rational deliberation, that it is best to A and will A on the basis of that judgment, thereby bringing about E" (Mele, 2006; parenthesis added). Thus, a complete description of the state of the universe just after Diana creates Z – including Z's constitution together with a complete statement of the laws of nature – entails a true statement of everything Ernie will ever do, including an action (say stealing a wallet) that results in event E. The argument proceeds in three steps: (i) Ernie is not morally responsible for event E, because he was designed by Diana to just bring E about. (ii) there is no significant difference (with regards to moral responsibility) between the way Ernie's zygote comes to exist and the way any normal human zygote comes to exist

in a deterministic universe, (iii) in no possible deterministic world in which a human being develops from a normal human zygote is that human being morally responsible for anything he or she does. Note that this type of argument is also implicit in Pereboom's argument, with case-1 corresponding to premise (i) and the assertion of a lack of significant difference between case-1 and case-4 corresponding to premise (ii).

There are two strategies available for compatibilists to answer such manipulation arguments (McKenna, 2008; Kane, 1996). The first one is a *hard-line* response of denying premise (i) that the manipulated agent is not responsible for event-E. The second one, is a *soft-line* response of denying premise (ii) that there is no-difference between the manipulated agent and normal agent in a deterministic world. Recently, a soft-line answer to the manipulation argument was provided by Deery and Nahmias (2017). The account I propose shares much of the theoretical framework used by Deery and Nahmias. However, I will argue for a soft-line response to Pereboom's argument and for a hard-line response to the zygote argument.

1.2. The luck problem

A related challenge involves the impact of luck on moral responsibility. Two types of luck, present vs. remote, are important to distinguish in this context. Present luck is the luck that occurs immediately prior to an action and which is usually thought to pose a problem for libertarians. For example, in a "torn" decision, a person's will is strongly divided between two alternative actions (such as helping a person cross the street or hurrying up to an important meeting) and the decision may appear random (Balaguer 2014). Remote (or constitutive) luck, on the other hand, is involved in traits that agents are born with (Nagel 1979: 24-38) and, is thought to pose a problem for compatibilists. One compatibilist strategy for dealing with remote luck is to concede that agents are not responsible for the traits they are born with, but to insist that agents make themselves responsible for the traits they develop by making repeated decisions and taking responsibility for them (Mele, 2006). Neil Levy, however, has recently argued that the combination of remote and present luck raises a problem for compatibilist theories of responsibility that rely on the developing traits strategy (Levy 2009, 2015). Levy argues that the developing trait responsibility-taking strategy is subject to what he calls a 'luck-pincer': Either a decision that presents itself to someone is an easy one, because it is grounded in the traits the agent already has, but then its outcome is subject to constitutive luck, or it is a difficult decision, in which case, it is subject to present luck (involving factors, such as mood, subliminal stimuli, or probabilistic processes in the deliberation). Levy concludes that "between them, bypassing present luck and constitutive luck[3] play a decisive role in all our decisions" (Levy 2015: 641). Accordingly, all actions are subject to luck, which seems to preclude responsibility. Moreover, this problem is most severe for compatibilist theories that are history-sensitive (McKenna 2004) in order to account for cases of manipulation.[4]

---

[3] Levy's conceptualization of constitutive luck differs from the standard one, which is based on three components: chance, significance and lack of control (Rescher 1995, Latus 2003, Pritchard 2005). Levy (2012) replaces the luck component with a wider class that involves "traits that are exceptional across the grouping to which one belongs". I do not rely on this conceptualization here, as the luck-pincer argument stands for the standard definition too.

[4] History-sensitive compatibilism accounts for the lack of responsibility in manipulated agents as a result of their anomalous history.

To conclude, source compatibilism faces a challenge in dealing with cases of manipulation and luck. I proceed by discussing an important theory that offers answers to these challenges and that shares its basic motivation with the approach that I propose, although it differs in an important way. Then, I turn to my own way of spelling out the conditions for responsibility, which sets limits to the causal source of responsibility, providing a natural solution to the problems of luck and manipulation.

2. From guidance to teleological control

A central idea to many compatibilist theories is that responsibility requires control. John Martin Fischer has made an important distinction between two types of control: guidance-control vs. regulative-control. The former is available even in a deterministic world, while the latter – the ability to do otherwise – holding fixed the past up to the time of action and the laws of nature – requires access to alternative possibilities and is favored by Libertarians (Kane 1996).[5] Fischer, who endorses the conclusion from Frankfurt's counterfactual scenarios that responsibility depends on events taking place in the actual-sequence, has proposed that guidance-control should replace regulative control as a necessary and sufficient condition for moral responsibility. Fischer illustrates guidance-control with the case of a person driving a vehicle and taking a right turn at a crossroad intersection (Fig. 1, blue (right-arrow) trajectory). Even if the vehicle's steering-apparatus is subject to some blockage that prevents the wheels from turning left, in case the driver decides to turn right and executes this maneuver successfully, the agent is nonetheless responsible for that action, as far as she holds guidance-control over it (Fischer 1994:132-144). According to the theory developed by Fischer together with Mark Ravizza, guidance-control is to be understood in terms of two conditions: *reasons-responsivity* and *ownership* (Fischer and Ravizza 1998).

Reasons-responsivity (or reasons-sensitivity), which is more relevant to the present discussion, is the capacity of the mechanism that issued in the action, of changing course (Fig. 1, green, leftward trajectory) if *sufficient* reasons to do otherwise were to be provided. This was developed into the moderate reasons-responsivity theory, which requires that the mechanism that produces the action is 1) regularly receptive to reasons (regular-receptivity) and, 2) weakly reactive to reasons (weak-reactivity). The regular-receptivity corresponds to a capacity to recognize reasons in such a way as to give rise to an understandable pattern (from the standpoint of a third party who understands the agent's beliefs and values). The weak reasons-reactivity condition corresponds to there being at least one possible situation involving sufficient reasons (different from the actual case) favoring another action, in which case the mechanism would have reacted to produce that different action. The weak reasons-reactivity condition helps to account for the lack of responsibility in actions performed under addictions, phobias and inhibition-control deficits (e.g., frontal lobe patients), in which the mechanism could not produce a different action even if sufficient reasons were offered. The second component of guidance control theory – the mechanism-ownership – requires the mechanism issuing the action to be the agent's own (this is supposed to preclude cases of manipulations, like hypnosis or brainwashing; but see Stump 2002). Note that the reasons-responsivity theory is

[5] Some compatibilists do not require this type of freedom, but they require a conditional freedom (Moore 1912), a dispositional freedom (Fara 2008), or a Humean causation freedom (Beebee and Mele 2002, Berofsky 2012). There are objections to all these theories, which for lack of space I do not discuss here (but see Beebee 2013).

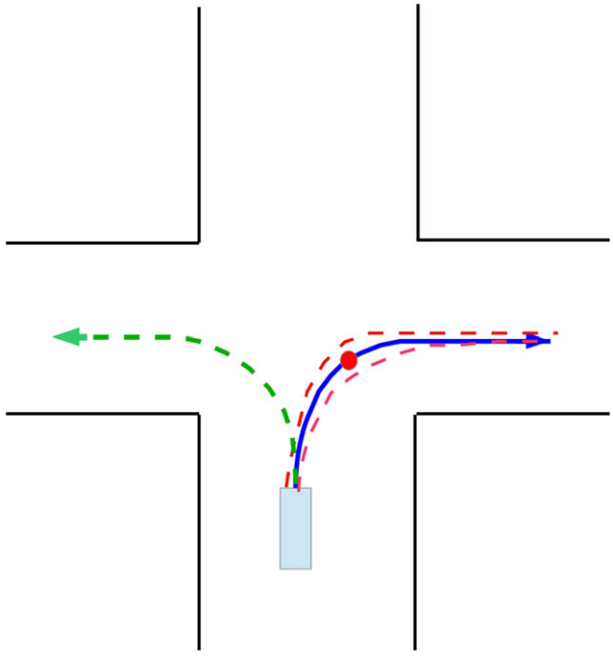

Figure 1. Illustration of guidance-control according to the reasons-responsivity theory (green) and according to teleological control (red). The blue curve is the actual trajectory (right-turn). The green and the red curves correspond to counterfactual trajectories within reasons-sensitivity and teleological control, respectively. The red circle (in the middle of the blue trajectory) is an obstacle that needs to be avoided via a local (minor) perturbation (red trajectories). Note that this can be done without having the vehicle turn the wheels leftward. The counterfactual trajectory in the reasons-responsivity theory (green curve), requires a global (major) change: 'doing otherwise'.

mechanism- rather than *agent*-based, as it defines the condition for moral responsibility in terms of the mechanism that issued the action: in a Frankfurt scenario the agent could not do otherwise (due to the counterfactual intervener) even if sufficient reasons were available, but the mechanism could, as the intervener is not part of this mechanism.

Here I propose a different analysis of the capacity of the driver who takes a right turn in the example above, which does not require the ability to do otherwise, given reasons to do otherwise. Instead of weakening the *could do otherwise* requirement, I propose to jettison it altogether. There are a number of motivations for this move. First, as recently discussed by Carolina Sartorio (2016b), the Fischer and Ravizza mechanism-based account of reasons sensitivity appears to be in conflict with the lesson from Frankfurt's scenarios that responsibility should be grounded on events in the actual sequence and, as discussed by Michael McKenna (2001, 2013), it depends on an unspecified theory of mechanism individuation. Both McKenna (2013) and Sartorio (2014) have made suggestions of how to modify the reasons-responsivity theory from a mechanism- to an agent-based one. However, there are independent reasons to doubt that the weak-reactivity condition is either sufficient or necessary for moral responsibility. On the one hand, Alfred Mele argued for a lack of sufficiency, illustrating this with the case of a person suffering from severe agoraphobia, a condition associated with the irresistible desire to stay at home (Mele 2000, 2006). The agoraphobe may nevertheless leave home under extreme

conditions, say, a raging fire, and thus satisfies the Fischer-Ravizza condition (of weak reason-reactivity) for being morally responsible for not attending his daughter wedding in the absence of fire, even though he is apparently not so. On the other hand, Pereboom has argued that weak-reason reactivity is not necessary for moral responsibility (Pereboom 2006). He asks us to consider an over-committed 'Kierkegaardian' agent whose ultimate commitment to some moral principle, say, telling the truth, would make her act against what she might perceive as sufficient reasons to do otherwise. Finally, note that one of the initial attractions of the weak-reactivity condition was to secure the lack of responsibility in cases of addiction. However, Harry Frankfurt and more recently Chandra Sripada have argued that whether an addict is responsible for A-ing (A being addiction related) is not a matter of whether she can resist A-ing, but rather whether she is a willing or an unwilling addict (Sripada 2017).

While some of the ingredients of the reasons-responsivity theory are important for moral responsibility (I accept that regular reasons-receptivity and inhibitory control – a mechanism related but not identical with weak reasons-reactivity – are requirements for an intact capacity of practical reasoning, which is necessary for moral responsibility; see next section), there are good reasons to explore an alternative way to characterize the guided control that agents have over their action in the actual sequence without invoking what Pereboom called an 'ersatz' principle of alternative possibilities (Pereboom 2006: 201). Rather than the ability to change course, if sufficient reasons apply, I see guidance-control as the ability to maintain course and achieve a goal, in the face of a variety of possible impediments, such as unexpected obstacles in the environment or the actions of other agents. This points to an important property of actions, their *teleological* character (see Wilson 1989, and Sehon 1997). Interestingly, Frankfurt has pointed out this important feature of intentional action in a paper that has been somewhat neglected in the responsibility debate (Frankfurt 1978). Frankfurt starts by arguing for a deficiency in the standard causal account of actions, according to which actions and mere movements can be differentiated "by nothing that exists or is going on at the same time as those events occur, but by something quite extrinsic to them – a difference at an earlier time among another set of events entirely" (1978: 157) – their causal antecedents (mental states, such as beliefs and desires). He then shows that this leads to difficulties in dealing with cases of deviant causation.[6] Frankfurt's proposal is that the problem of understanding intentional actions arises from the attempt to locate the distinctive features of actions in their antecedents, rather than in the process by which the action unfolds, and which needs to be *teleologically* guided by the agent: "When we act, our movements are purposive. This is merely another way of saying that their course is guided" (p. 159). Furthermore, he characterizes guidance as a "causal mechanism, whose readiness to bring about compensatory adjustments tends to ensure that the behavior is accomplished." (1978: 160; see also Nagel 1977).

In the following, I will refer to this capacity as *teleological-control* to distinguish it from guidance-control, which is today mostly associated with the reasons-responsivity theory.[7] In the driver example, teleological-control involves the driver's ability to

[6] Frankfurt deploys the classical deviant causation case of man who intends to spill his glass of water in order to signal to his confederates to begin a robbery, but this thought makes him anxious to the degree that his hand trembles spilling the glass (Davidson 1973). While the spilling is caused by the mental state of the agent, it is not teleologically guided.

[7] In previous work, I labeled this property of intentional action *teleological-guided control* and I argued that it is central to attributions of agency and responsibility (Usher 2006). This label, however, is too similar to Fischer's and may obscure important differences between our views.

intervene when needed, to keep the vehicle on track, so as to achieve the same goal under potential perturbations, such as an unexpected obstacle on the road that would divert the vehicle off its course. This is illustrated in Fig. 1 (red-dashed, right-turning, trajectories). Note that avoiding such obstacles requires an ability to be sensitive to occurring reasons (such as the red-obstacle in Fig. 1), in order to make corrections to the planed trajectory. However, this involves local changes (without alteration of the start/end points) by "turning right" at different points, rather than global changes to the actual trajectory. Indeed, it is this ability to avoid obstacles (to a certain degree) that a driving instructor is interested in when assessing a driver's competence. To state this in possible-world terms, an agent has teleological-control over an effect of her action, only if she would bring about the effect, not only in the actual world but also in a certain set of similar worlds. While this still involves a modal property that departs from the actual sequence, I will argue (in section-4) that it is grounded in facts (attractors) that are part of the actual sequence (Sartorio 2016). Before doing this, I now turn to show how this framework enables us to quantify the degree of responsibly that an agent has for the effects of her action by deploying a recent theoretical development within the theory of causation, the notion of robust causation, developed by James Woodward (2006).

3. Teleological Control and Robust Causation

I apply robust causation and teleological control in order to explicate degrees of moral responsibility (as reflected in judgments of blame or praise), which are reduced under manipulation. However, as we will see, robust causation and teleological control are not sufficient for moral responsibility. They provide, however, a measure of the causal-responsibility that an agent has for the outcomes of her actions, which is among the most important ones (say, from a legal perspective). Moreover, as Deery and Nahmias (2017) recently advocated the issue of causation was, unfortunately, neglected in the free-will/responsibility debate, resulting in an unprincipled notion of causal-efficacy and causal sourcehood. I start with a characterization of causal responsibility before describing additional conditions for moral responsibility that involve what can be broadly characterized as a "Compatibilist-friendly Agential Structure" (CAS; McKenna 2008).

The starting point is an *intervenionist* account of causation, which characterizes causation on the basis of interventions on *control-variables* and the observation of their effects (Campbell 2010, Pearl 2000, Woodward 2003).[8] The main thrust of Woodward's robust (or insensitive) causation theory (Woodward 2006) is that not all causal chains of events are equal with regards to causation. While some are *robust, i*.e., *insensitive* to changes in the background circumstances, others are extremely sensitive. Background circumstances are of the essence here because all causal chains between events that do not correspond to the complete state of the world are subject to background circumstances. Pressing on the gun's trigger would cause a bullet to be shot, so long as rain did not wet and corrode the mechanism – a background circumstance; shooting a person in the leg would cause death, as long as there was no emergency help available – another background circumstance. As Woodward illustrates, while the causal chain of shooting a person in the heart and her death is relatively robust, the causal chain of shooting her in the leg and her

[8] See Campbell (2010) for a discussion of mental causation, where mental states are the most relevant control variables for an explanation of human behavior.

death is sensitive (thus non-robust), as it depends on a more specific (singular) set of background circumstances. Woodward shows that this distinction also accounts for differences in the perceived strength of causation between effects of direct actions and effects of omissions, or of other cases of sensitive non-robust causation, such as double prevention. We can now put forward the following proposition to characterize the role of robust-causation in determining the causal degree of responsibility an agent has for outcomes of her action:

> ***(RC1)*** The *degree* of responsibility an agent has over an outcome of her action is determined by the *robustness* of the causal chain from the mental states (intention and beliefs) involved in the production of the action to the outcome.

Two important features of (RC1) need to be explained. First, I limit the causal-chain to begin from the formation of intentions and states of belief, since they are instrumental to the execution of action plans; going further back to desires and values would contradict the intuition that we appear to have full responsibility for actions that we make after 'torn' decisions, in which opposed desires motivate us in different directions (Balaguer 2014). In a "torn" decision, a person's will is strongly divided between two alternative actions. In such cases, the causal-link between the divided desires before an intention for action was formed (as a result of a decision), and the consequences of the action are not robust (and thus is highly sensitive to background conditions). Since I believe that agents responsibility is not reduced under torn-decisions, I suggest that the robustness in RC1 corresponds to the causal chain from intention to outcomes (see also Bratman 1981, for an argument that intentions have a special reason-giving status for carrying out actions).

Second, we can quantify robustness in terms of the fraction of possible worlds within a local neighborhood of the actual world (corresponding to variations in background circumstances with the action kept fixed), in which the effect occurs; alternatively one might measure (or at least rank), distances to the nearest similar world in which the effect does not occur; the larger this distance the more robust the causal contribution of the action to the effect.[9] As I argue below, this causal responsibility measure depends on the teleological control that the agent deploys in the action planning and execution. Although Woodward does not discuss teleological goal-directed action, the examples discussed here (and in the next section) illustrate the fact that for events that are the result of an agent's action, teleological control is the most important source of robust causation.

Consider a typical case of an action that involves moral responsibility: a case of murder. I claim that it is one of the features of a murder that the murderer would also cause the event in a set of similar worlds (say, by making adjustments to her plan, so as to bring about the intended goal in the face of some variability in the movements of the victim). One may object to this, by considering a 'shy' murderer, who gives up her action at the slightest deviation from the scenario she planned when she perceives any danger. But I argue that some

[9] Pritchard (2005) discusses the role of "veritical" epistemic luck in the formation of justified true beliefs that are not knowledge. The robust (sensitive) causation of action poses a parallel luck problem. Just as in the epistemic case, luck plays too much a role in determining true-belief, so in sensitive causation of action it plays too much of a role in determining the effect of the action. In both cases, this can be cashed out by requiring that the consequence (true belief or action-result) would occur not only in the actual world but in a neighborhood of similar ones.

robustness needs to remain in place for an action to qualify as a murder. For example, the murderer must intentionally kill her victim, and this should imply that if the victim moved from point A to point B, the murderer would point her gun accordingly. Indeed, intentional actions are teleological: they usually involve a plan, which is supposed to achieve its goal, at least, under some variations in background circumstances. This can be supported by appealing to practices of ascribing intentions: if we knew that an agent pulls the trigger on her victim in a singular situation and not in any other situation (not even one with an irrelevant minor difference, like a change of the victim's shirt color), we would be justified in doubting that the agent had 'killing the victim' as her intention.

A crucial consequence of relying on teleological-control to deliver robust causal chains is the ability to account for the difference in responsibility in the following three cases: (1) A shoots B intentionally in the heart, (2) A intends to kill B but rather than shooting Y directly, she lets him escape through a dangerous jungle (in which there is a 10% death rate), (3) — the multiple shot murder: A shoots B multiple times until she is satisfied that B is dead. While in (1) the shooting (and the intention to shoot) *robustly* causes B's death under a variety of circumstances (such as those involving B's position), this is not the case in (2), where death is highly sensitive to background circumstances (animals in the jungle and resources of B) that dilute the responsibility of A, and thus B's death appears subject to luck. While in case (2) there appears to be a reduced degree of responsibility, in case (3) there appears to be an enhanced one. Here the nearest possible-world in which death does not occur is even more remote (death occurs in a larger neighborhood of nearby possible worlds). Note, however, that the degree of responsibility is not changed if A shot his victim after a torn decision, despite the fact that there are nearby similar worlds in which A decided to abort. As mentioned above this invariance is accounted for by the requirement that robustness involves a causal chain that begins with the formation of an intention. In the next section I further discuss the rationale for this with respect to a tracing condition.

Robustness, via the deployment of teleological control, is also at play in other intuitive cases where there is a contrast of high vs. low degree of responsibility. A soccer player, scoring a goal would get more credit if she controlled the ball (say, by dribbling and then shooting into the goal) than if the ball deflected from her back; in the latter case there is little robustness, as most changes to the path of the ball or the locations of the opponent players would change the outcome. Even ballistic actions can have robustness when planning and skill are deployed. Compare a 'Beckham-style' free kick in soccer that bends around the defensive wall with one deflected off the wall, both resulting in a goal. Unlike the latter, Beckham's curling kick is robust with respect to changes in the positions of the opponent players (within some limits).

I have argued so far that the condition of teleological control, as measured by the robust causation between an agent's intentional plan that triggers an action and an outcome of that action, determines the causal degree of responsibility that an agent has for that outcome. However, while teleological-control and robust causation are factors that determine the degree of responsibility (that is relevant for judgment of praise or blame), they are not sufficient for *moral* responsibility. A paranoid schizophrenic who plans and commits a murder in order to "save the world" after having heard voices instructing her to do so is not morally responsible, even if she deploys robust causation and teleological control.

Additional conditions are thus needed for moral responsibility. Several options exist, which have been widely discussed in the literature.[10] Morally responsible agents need to have an intact capacity for practical reasoning. This involves sensitivity in recognizing reasons and foreseeing consequences of one's actions, as well as a capacity to inhibit prepotent or automatic responses that conflict with the practical reasoning. Such an inhibition mechanism is lacking in frontal lobe patients, who sometimes carry out actions that conflict with the outcome of their practical reasoning. The inhibition mechanism, however, does not require that a Kierkegardian agent ever aborts an action that was the result of practical reasoning. Moreover, for responsibility, the action should be such that it advances the agent's cares (a set of motivational states lying at the foundation of a hierarchy of motives; Sripada 2015) or long-standing policies (Bratman 1987). It is beyond the scope of this paper to provide a full specification of the practical reasoning capacities, which I refer to as "Compatibilist-friendly Agential Structure" (CAS) (McKenna 2008). The CAS requirement rules out the schizophrenic who is impaired in her ability to make adequate perceptual and cognitive inferences about the world (Hug, Garety and Hemsley 1988), as well as frontal lobe patients who are unable to inhibit sexual gestures, which the patients understand to be inappropriate (Duffy and Campbell 1994). Finally, CAS rules out responsibility of people who suffer from addictions or from phobias (which they do not endorse), since their actions do not promote the agents' cares or long term-policies. While CAS and robust-causation jointly determine the moral responsibility that agents have over the outcome of their actions, I suggest that the degree of responsibility (relevant for attributions of praise and of blame) is determined by the robust-causation, with the CAS forming an enabling condition.[11]

There are two important prerequisites for teleological control of the type I have discussed. One is a planning capacity, which includes means-ends and error correction strategies; the second is a decision-making capacity, which includes a capacity to represent the environment and to select sub-goals that promotes the agent's long term policies. In the next section, I turn to characterize these capacities and discuss how they fit with certain properties of laws of nature. As I will show, this analysis sets temporal limits on robust causation and on teleological control, which are central for offering a solution to the problems of luck and manipulation.

4. Grounding teleological control: attractors, bifurcations and responsibility tracing

The implication of the metaphysics of natural laws (deterministic or indeterministic) for the existence of agency and responsibility (as needed for attributions of praise and blame)

[10] There are two major contending approaches to moral responsibility. The more prominent one considers control as the most important condition (Fischer and Ravizza 1998). Another approach that originates with Frankfurt (1971) focuses on the requirement that the action should express the agent's character or deep-self (Sripada 2015; see also Wolf, 1993, for a combined approach). It is beyond the scope of this paper to arbitrate between these approaches. Moreover, I believe that elements from both are required.

[11] I believe that this is enough to account for most cases of moral responsibility, and is consistent with previous literature. However, future work may explore an extension in which the CAS conditions contribute to the degree of responsibility. One such idea was suggested by Coates and Svenson (2013), based on the reasons-responsivity theory. Accordingly, responsibility for an action is reduced the further we have to move in the possible world space, so that an alternative action takes place. This accounts for the intuition that a depressed person who fails to perform an action due to her depression has a reduced responsibility. However, it also counterintuitively, predicts reduced responsibility for actions that are the outcome of robust 'Kierkegardian' convictions, compared to torn-decisions (in which the distance to alternative-action worlds is the smallest).

is a topic of intense controversy (Kane 1996, Roskies 2006). The irrelevance of determinism to moral responsibility was clearly articulated by Fischer, who stated that "it is implausible that our status as agents, who do things and are responsible for their effects, should hinge on subtle ruminations of theoretical physicists" or that "our status as genuine agents should depend on whether natural laws have associated with them (say) probabilities of .99 or 1.0″ (2012: 118).

I maintain that what is required for robust causal chains and teleological-control is not determinism or indeterminism, but rather successive stages of *attractor* and *bifurcation* dynamics (see e.g., Complexity Labs 2016, Kelso 1997). Attractors and bifurcations are dynamic entities, which are evident in natural systems, involving the property of *convergence* or *divergence,* respectively. I will illustrate these dynamic properties below, but before doing so, I want to clarify the motivation for appealing to this notion in the context of agency and responsibility. In particular, I want to clarify the reason why attractor and bifurcation dynamics are more relevant to responsibility than determinism.

I propose that robust causation requires teleological control, which in turn requires a property that is stronger than determination – a *reliable or counterfactual type of* determination. (Note that I am not saying that such a reliable-determination has to be universal across space and time, but only that it temporally holds in the agent-action system during action planning and execution; in section 4.3, I discuss *bifurcations* which sets temporal limits to determination). To illustrate the difference between these two types of determinations consider an example from William James, in which he contrasts the behavior of an intentional agent (Romeo) with that of a non-intentional object (iron filings):

> "Romeo wants Juliet as the filings want the magnet; and if no obstacles intervene he moves towards her by as straight a line as they. But Romeo and Juliet, if a wall be built between them, do not remain idiotically pressing their faces against its opposite sides like the magnet and the filings [when a card is placed between them]. Romeo soon finds a circuitous way, by scaling the wall or otherwise, of touching Juliet's lips directly. With the filings the path is fixed; whether it reaches the end depends on accidents. With the lover it is the end which is fixed, the path may be modified indefinitely (James, 1890, p. 20).

What distinguishes the intentional agent (Romeo) from the non-agent (fillings) is a capacity to deploy a counterfactual determination characterized by *equifinality*; Lombrozo 2010). The type of dynamic system that possesses counterfactual determination is an attractor. Attractors appear in non-linear dynamical systems and are characterized by a set of possible trajectories, each corresponding to the future state development of the system, starting from a different initial state (in the modal space of possible worlds). As illustrated in Figure 2, attractors correspond to cases in which the system's possible trajectories (i.e., their future states) *converge* towards a final state – the attractor's center – which exhibits robustness by absorbing perturbations. As the arrows show, the system starting from various initial states converges towards the attractor center.

To illustrate the difference between mere determination and counterfactual determination, I show in Figure 3, two cases in which the actual chain of events connect A to B. While on the left panel – which illustrates a case of *multifinality* – small changes to A result in changes in B, this is not the case for the attractor case (right panel) – which illustrates a case of *equifinality*, as the attractor absorbs perturbations to the actual state.

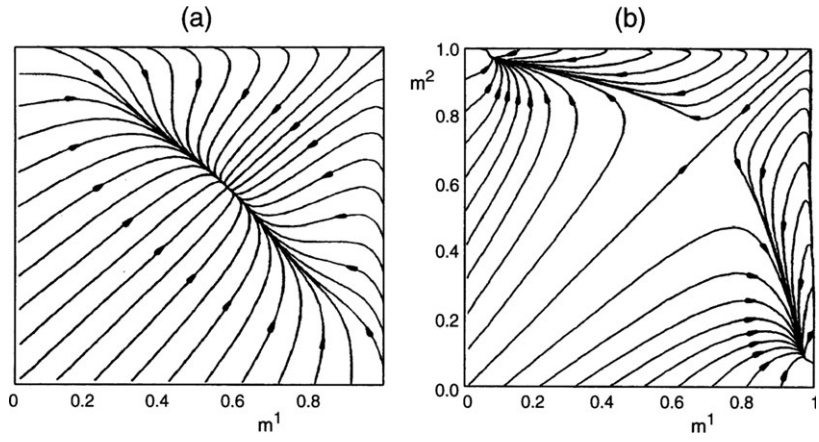

Figure 2. Two examples of attractor dynamics in a neural-systems (Horn, Usher 1990). The 2D (m1/m2) space corresponds to the activations of two competing neural assemblies. The arrows show how the state of the system (activation of two neural assemblies) changes in time. a) single-attractor (the state converges to the attractor center for all starting-points), b) two-attractor system with two possible convergence states depending on the starting point of the system. The border between the two attractor basins is a bifurcation – starting points near the bifurcation show divergence in their future trajectories

I contend that the presence of an attractor is a grounding fact for the actual-sequence trajectories that follow from it, as the attractor distinguishes between sequences that occur by chance (A to B in Fig. 3 left) and those that do not (A to B in Fig. 3 right) and that in order to account for the basic distinction between agents (who do things) and objects (which merely move), the world has to contain attractor dynamics.

Attractors emerge in dynamical systems as soon as the dynamical process includes a certain type of feedback loops. Such feedback loops can appear in non-biological systems, as a result of a non-linear coupling of dynamic variables, but they mostly emerge in biological organisms – such as animals – endowed with a perceptual apparatus and an action production system. Unlike plants, animals generate feedback loops via their cognitive system, by relying on means-ends and error-correction strategies to achieve goals (see Newell and Simon 1972). Below I briefly describe two types of attractors that play a role in agency and robust, teleological control, before I consider their implication to a theory of causal responsibility.

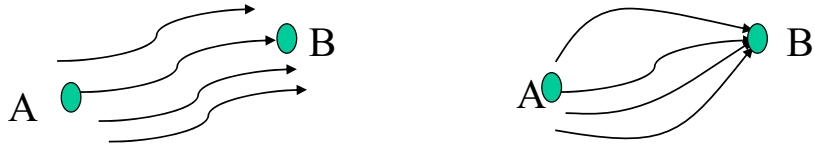

Figure 3. Two cases of deterministic dynamics, which account for the same A → B causal chain. Left: perturbations in the starting point of A (corresponding to variations in background circumstances) result in perturbations in B; Right: same perturbations are absorbed by the attractor (centered on the goal-B), implementing teleological behavior.

4.1. Intention states and self-policies (or cares)

An important type of attractor, relevant to teleological-control, is based on a basic feedback loops at the cognitive (neuronal) level, and is a precursor to having action plans. I follow here Michael Bratman (1987) and Richard Holton (1999), who have discussed in detail the role of *stable* intentions in allowing the deployment of a planning agency. According to standard cognitive science theory, intentions are the end-states of decision processes and the start of actions or new decisions, and once formed they are relatively immune from re-consideration and thus from revision. Since attractors are resistant to perturbations, they are the stability bearers of our mental life, endowing intentions with the necessary stability to enable agents who possess them, to control and coordinate future actions. As reasoning and deliberation take time and effort, it is rational not to revise intentions back and forth. As Holton phrased it, "once we have decided in which restaurant to eat, it is a good idea to let the matter rest, without endlessly discussing pros and cons; and this is true, even though it occasionally means that we shall go to a restaurant that was not the best choice." (Holton 1999: 248). Not doing so has a high price on mental effort and makes interpersonal communication a nightmare.[12] Research in psychology and behavioral neuroscience backs up these theoretical analyses, showing that we are more likely to be persuaded by evidence in the early stages of a decision than toward its end, when a stable decision-state (and intention) is formed, as indicated by commitment biases (Bronfman et al. 2015, Nickerson 1998). Nevertheless, too much stability corresponds to unwise stubbornness and achieving optimal action control performance involves a compromise between flexibility and stability.[13]

In addition to implementing intention states, attractors also characterize stable self-governing policies or cares (e.g., being committed to liberal humanism, or trying to be less impatient with others), which control the weights given to various reasons for specific actions and crystallize pressures from various elements of one's psychic stew into a more decisive attitude (Bratman 2000; see Roskies 2016 for discussion of brain mechanisms of self-government). Furthermore, it might be plausibly argued that the robustness property of attractors grounds the stability of intentional self-governing policies for human agents, enabling us to take an *intentional stance* towards bearers of attractor states and predict their behavior to a remarkable degree (Dennett 1981). To borrow an example from Dan Dennett, we can predict with remarkable accuracy the behavior of person from a phone conversation with his wife telling him "Oh, hello dear. You're coming home early? Within an hour? And bringing your boss to dinner? Pick up a bottle of wine on the way home, and drive carefully" (1981, p. 69). Moreover, unlike a Laplacean superscientist-prediction, intentional stance predictions are quite robust with respect to variations in various details (traffic, etc). I maintain that this robustness is the result of attractor dynamics that mediate intentional plans (see further below).

4.2. From attractors to robust (teleological) causation

While attractor systems are prevalent in neural systems (Fig. 2), they also occur in non-biological systems, which are subject to certain type of non-linear coupling of the dynamical

[12] According to Holton, the excessive revision of intentions is a characteristic of the weakness of will and of capriciousness.

[13] See Holton (1999) for an account of the conditions determining the need to revise intentions.

variables that characterize the system. For example, a tornado may be seen as a (quasi) stable attractor that generates a number of robust effects on its environment (as long as it lasts). In order to obtain teleological control of the environment, however, the attractor has to involve not only a process within an organism (say, an intention-state that is realized in a person's brain), but rather a specific coupling of the organism with the environment. For example, a stable intention-state, cannot bring about (on its own) to desired states of the world. When coupled, however, with the environment, as part of a sensory-motor action generation (closed-loop) system, intention states do achieve robust teleological control of the relevant part of their environment (Kelso 1997). The key property of this coupling is the ability to detect the state of the environment and to perform actions that carry out the necessary compensatory corrections needed to bring the organism closer to her goal. To illustrate this, I show in Figure 4 that such a closed loop system – in this case characterizing a prey-predator interaction – has attractor properties. To do so I present a computer simulation of the prey-predator pursuit dynamics. In each simulation, the prey (blue) starts at coordinate (500,1000) and the predator (red) at coordinates (0,0). With each time-step, the prey moves at constant velocity (direction changes at random times), while the predator moves with a higher velocity

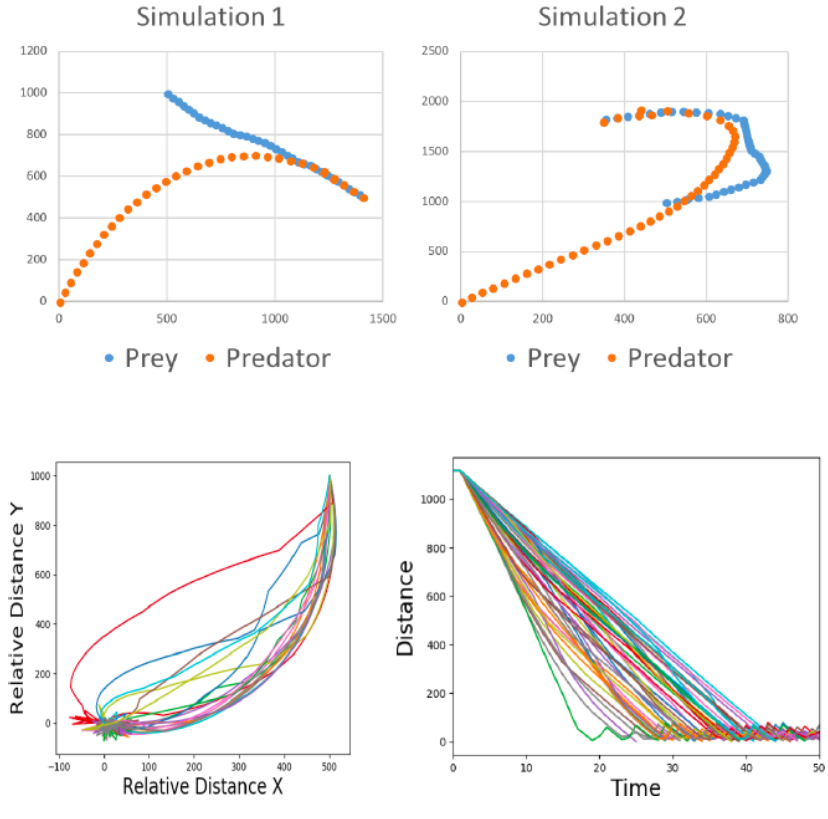

Figure 4. A dynamical simulation for a simplified prey-predator system. Two trajectories are shown in the upper panels (prey-blue, predator-red). The bottom left panel shows the relative distance (as a vector with x-y coordinates) for 20 such trajectories, all starting in the upper right corner and showing clear convergence to zero-distance. The bottom-right panel shows that in all these trajectories the distance between the prey and the predator converges to zero as the attractor is reached.

towards the location of the prey at that moment (sensory-motor loop). In both simulations (Fig. 4 upper panels) we see that, because of the closed (sensory-motor) loop, the predator (red) converge on the prey (blue). Moreover, this takes place despite the prey taking different trajectories. In the two lower panels I show twenty such simulations, for the separation vector (bottom-left panel) and for the Euclidian distance (bottom-right panel) between the predator and its prey, which converge to zero, as the attractor is approached.

This predator-prey illustration is somewhat limited. The predator will abort as soon as its sensory-motor loop with the prey is disrupted, say, by an occluder. In human agency, however, one can deploy more abstract (and thus more powerful), but functionally equivalent, *means-ends* reasoning processes, as part of action plans, which execute the corrections that the agent needs in order to reach a target, subject to variability in background circumstances (Newel and Simon 1972). Means-ends reasoning is an efficient strategy of humans problem-solving, which also deploys a closed perception-action loop in the *problem-space*. To do so, the agent needs to possess a goal, a perception of her present state and to compute "operations" reducing the "distance" between the two. In the example above, while a predator aborts its pursue of the prey when the latter is occluded by an obstacle, a more powerful agent, who possesses abstract knowledge about the victim's goals, may rely on this to infer the prey's likely move and plan how to pursue it.

To conclude, I propose that any system displaying genuine teleological-control via means-end reasoning and action plans must be governed by attractor dynamics over the time interval that corresponds to the execution of the action plan, and that this stability is what grounds the ability of agents to robustly affect their environment. Unlike determinism, which, as Fischer correctly points out, appears unrelated to our agency capacity, attractors and bifurcation are a prerequisite for any type of agency, such as the forming of intentions and their execution.

4.3. From attractors and bifurcations to responsibility tracing

By contrast to attractor dynamics, our mental life is also subject to bifurcation processes ranging from simple decisions (which apartment to buy) to complex and much slower personality forming processes (e.g., becoming devoted to human rights and liberal-humanism or to extreme nationalism). Bifurcations correspond to dynamic situations in which the trajectories are highly *divergent* and sensitive to the smallest perturbation. The presence of successive stages of convergence and divergent dynamics (attractors and bifurcations, respectively) is, I suggest, the key property of the world, consistent with the laws of nature, which enables us to see the world as containing agents and outcomes, with agents conceived as sources of responsibility (or robust causes) for the outcomes. Based on the assertion of a succession of attractors and bifurcations in the natural law, we can add an additional responsibility condition, *RC2*. I start with a simplified binary condition, before I consider the more refined graded version (*RC3*).

> ***RC2***: an agent has responsibility for an action when the action can be teleologically/robustly traced to certain mental states of the agent, but no further. At some stage, one reaches a bifurcation: no antecedent before this bifurcation can *robustly* explain the action.

Accordingly, an agent is responsible for the effects of her action when she is the *earliest source of robust causation* leading to the effect.[14] As targets of responsibility often involve a hierarchy of agents, this gives a novel interpretation to Robert Kane's metaphor of ultimate-responsibility, in terms of President Harry S. Truman's famous phrase, "The buck stops here," illustrating the principle of no further outsourcing of responsibility.

The presence of convergence/divergence dynamics, which sets limits on the chains of robust-causation, is what fixes responsibility, by contrast with either determinism or indeterminism. According to mere determinism, the responsibility for any event is traced back (if at all) to the Big-Bang. In a world that exhibits successive attractors and bifurcations, on the other hand, responsibility is traced back to the earliest source of robust control. This distinction has important consequences, which allows us to address many problems. I will argue that this allows us to account for the reduced responsibility in cases of manipulation or indoctrination. By combining the conditions *RC1, RC2* and the condition of promoting long term aims, we can formulate *RC3*.

> ***RC3***: when an agent exhibits teleological-control in bringing about an event that promotes her long-term goals (or cares), while her goals are subject to teleological-control by another agent (for example, an agent that is manipulated), the degree of responsibility she has for the effects of her action depends on the difference between the amount of teleological control that she deployed and the one that was deployed on her by the manipulator.

I will next show that this condition allows us to address a number of manipulation scenario objections that have been raised against compatibilism.

5. Addressing the problem of manipulation and the luck-pincer

The robust causation theory allows us to meet the challenges posed by manipulation scenarios as well as by Levy's luck pincer, because the theory foregrounds a central difference between mere determination by physical laws and control by other agents (manipulations or indoctrination), in terms of the robustness of the causal chain.

5.1. Manipulation arguments

Mere causal determinism involves sensitive (non-robust) causation that depends on variations in background circumstances (Fig. 3, left). Agent-control, on the other hand, requires robust causation that is stable under background variations (Fig. 3, right). By tracing responsibility to the earliest source of robust-causation, the robust causation theory provides a soft-line response that denies the lack of a difference between the normal agent (Pereboom case-4), and agents that are subject to manipulation (Pereboom case 1-3), accounting for the intuition of reduced responsibility in cases of manipulation or indoctrination.

In particular, when an agent is subject to teleological guidance control that is externally applied, like Professor Plum, in Pereboom's case 1 the agent is controlled and thus

[14] In an article that was developed independently of the present one, Deery and Nahmias (2017), use a similar condition. They rely on Woodward's principle of causal invariance, to determine causal sourcehood. Accordingly, X=x is the causal source of Y=y, iff X bears the strongest causal invariance relation to Y among all the prior causal variables; where causal invariance is determined by the relation being stable across a wide range of changes in background circumstances. The theory I propose emphasizes the role of attractors in the realization of teleological control and may differ from that of Deery and Nahmias in its answer to the zygote argument (see section-5).

some of the responsibility for her action resides with the earlier source of teleological control. Note that in case-1, the manipulator robustly controls the values and actions of the manipulated agents – she can make them develop such values not only in the actual world but in similar ones, by applying a variety of interventions. Using the teleological-control account, we can explain Pereboom's intuition that there is no clear border that delimits responsibility from non-responsibility cases on the apparently slippery slope described, yet contend that the responsibility of the Plum character in the various scenarios is a matter of degree. By contrast to case-1, in case-4, the Professor's character is the earliest source of robust causation to which the effects of his actions can be traced.[15] In case-3 (training/indoctrination), the indoctrinator has partial control, whose strength depends on the effectiveness of the training/indoctrination. In case-2, a partial control over the Professor is also achieved, if the "programming" blocks him from taking actions that are not "rationally egoistic".[16] This is consistent with the common intuition of reduced responsibility for agents that have been subject to effective indoctrination, which is supported by studies in experimental philosophy (Feltz 2013, Murray and Lombrozo 2016, Sripada 2012). Pereboom is thus right about the slope, but the robust causation account shows why it is not slippery.

Thus, according to the teleological-control/robust causation theory, normal and indoctrinated agents differ in a major respect. While both have their traits determined, in the case of the normal agent (Professor Plum in case-4), the determination involves sensitive causation that depends on myriad factors (genes, society, etc); hence the agent remains (via her mental states) the earliest source of robust causation for the actions she carries out, whereas for the indoctrinated agent the robust causal chain extends to the indoctrinator. While one may contend that educational systems also induce a conformity of values, one may note that this is an issue of degree, and that a good education encourages the ability to entertain multiple viewpoints, and to reason on the basis of evidence. Indeed, an important signature of good education, as distinguished from indoctrination, is the presence of diversity in the values and ways of thinking that it engenders, which reflects sensitive (non-robust) causation. This stands in contrast to indoctrination, whose signature is a uniformity of values, reflecting robust causation, stemming from the indoctrinator.

The account given here for responsibility under manipulation contrasts with that of the Fischer and Ravizza reasons-responsivity theory. While the teleological/robust-causation account predicts a variable degree of responsibility that decreases from the normal (case-4) to the "abnormal" (case-1), according to the reasons-responsivity theory, Professor Plum is responsible in all the cases. As Fischer explains, "Plum has taken responsibility for the manipulation mechanism; after all this is the mechanism on which he always acts, and when an individual develops into a morally responsible agent, he takes responsibility

---

15 Note that receiving advice from a friend to kill Ms. White does not reduce Plum's responsibility in case-4, because such advice does not exert a robust impact on his intention (unless the advice is part of a systematic indoctrination, which moves us into case-3). Thus the earliest source of robust causation remains his intention (or decision) to kill Ms. White.

16 As argued by Berofsky (2005) and Mele (2005), it is not obvious that programming does not interfere with CAS capacities, such as the ability to critically evaluate. Here I follow McKenna (2008; p. 151), who suggested that Plum may have developed his egoistic values based on years of study of Hobbes ethics. Still, if the programming affected this process, this means that it provides a systematic bias towards such values (say, by suppressing alternative ethical outlooks). In such a case, Plum is subject to teleological control by the programmers (his values are robustly determined by them).

for his actual-sequence mechanism, even if he does not know their details. Furthermore, the desires on which Plum acts are not irresistible," (Fischer 2004: 156). I maintain that the graded responsibility picture, which the robust causation theory delivers, matches better with typical human judgments, as indicated by recent experimental studies (Feltz 2013, Murray and Lombrozo 2016, Sripada, 2012).

While this analysis agrees with Deery and Nahmias (2017) approach to Pereboom's manipulation case – we both offer a soft-line response – I propose that for the zygote argument it is best to take hard-line strategy (denying premise-i, that Ernie has no responsibility over event E). By contrast, Deery and Nahmias deny premise (ii) of the argument, that there is no difference (with regards to responsibility) between a normal person and one (such as Ernie) whose zygote was designed by Diana (the Goddess) to bring about event E (e.g., steal a wallet), given her perfect knowledge of the state of the universe and of the laws. To support this, they argue that the causal invariance – a measure of stability to changes in background circumstances – between Diana and E is stronger and thus trumps the causal relation between Ernie's decision to steal the wallet and his action. The latter is motivated as follows: "The reason that Diana's decision (DD) bears the strongest causal invariance relation to s*teal* is that Diana can (we are assuming) ensure that Manny steals as she intends across the widest possible range of changes to C" (where Manny is the Deery and Nahmias equivalent of Mele's Ernie, and C corresponds to background circumstances). Can we assume, however, that Diana can ensure the theft under changes in background circumstances? Remember, that the way in which Mele describes Diana creating the zygote that developed into the person that 30 years later will steal the wallet, was based on Diana's perfect knowledge of the state of the world at that time and the laws of nature (which are assumed to be deterministic). Obviously, if we are to make any intervention on background circumstances (say bringing about rain on the day of the theft), there is nothing that Diana could do to "ensure" that Manny steals the wallet. This is because her prediction is conditioned (by assumption) on the actual state of the world at the time she made it and therefore she has no way to extend to counterfactual interventions.[17] To state this differently, by assumption, Diana's action causes event E only in the actual world (and not in any other similar worlds). I contend that, in contrast to Diana, *online* manipulators who are ready to intervene when needed (case-1 in Pereboom) and who hold teleological control over the actions of their subjects, do undermine their responsibility. Unlike Diana, they teleologically control the agent to act as they wish by counteracting any changes or interventions in a way that mere physical prediction cannot achieve.[18] Moreover, if Diana was to rely on programming (rather than on physical prediction) this may also allow a soft-line response of reduced responsibility of the manipulated agent, as the programing can systematically bias the formation of values (leading to robust outcomes – i.e., which take place not only in the actual world).

5.2. The luck challenge

Consider now Levy's luck pincer. This is a problem for compatibilism only if one accepts that constitutive luck threatens moral responsibility. A good way to illustrate this threat is by using a scenario deployed by Saul Smilansky, in which a thief who

[17] I leave to the readers to decide if the notion of a more powerful Goddess who ensures events in counterfactual worlds is logically coherent.

[18] It is possible that Deery and Nahmias would, like us, deny premise (i) and not (ii), if Diana's powers are interpreted as I suggested.

admits of having committed a theft out of laziness to work for a living and indifference to other people's suffering, argues he should not be held morally responsible because it was not his choice to be born lazy or indifferent to others' suffering (Smilansky 2000). I believe that this threat can be resisted. I follow Hurley (1993) and Rescher (1995) who argued that the threat presupposes a characterless agent who is randomly allocated with traits, while it is more plausible that the configuration of traits, or at least the most central of them, is what defines the agent's identity.[19] The robust causation theory can build on this to provide a way out of the luck-pincer. As it provides a principled distinction between teleological control and mere determination (Fig. 3), it rules out responsibility in cases of manipulation by external teleological-control and allows traits not subject to such control (i.e. those associated with "constitutive luck") to screen out sources of present luck. Accordingly, while agents do not satisfy the highest possible responsibility criterion of making themselves into what they are out of nothing (Strawson 1994),[20] they nevertheless satisfy a more modest criterion of being the earliest source of robust causation for the consequences that follow from their actions.

The teleological/robust causation theory also explains the intuition that while drastic life changes that appear to involve luck (Arpaly 2003, Nagel 1979) do not appear to reduce the agents' responsibility (McKenna, 2008, p. 156), changes that are the result of manipulation do (Sripada, 2012). In the latter case (but not the former), the responsibility can be traced back to events that precede the agent's actions or decisions.

6. Conclusions

I presented an account of agency and responsibility that is agnostic on the issue of whether physical laws are deterministic or indeterministic. The central part of the theory is the distinction between teleological control (grounded by attractors) and mere (sensitive) causation, with the former, but not the latter, mediated by robust causal chains. This results in an agent-based account that is grounded in the actual-sequence and traces responsibility for an event to the earliest source of robust causation. The account explains why the moral responsibility for outcomes appears to come in degrees and provides a source compatibilist account of the problems of luck and manipulation.[21,22]

---

[19] A different way to put this is that one is not 'lucky' to not be born with different traits, because the chance condition does not apply: the agent does not exist in the similar worlds, in which those traits are changed. Levy (2012, 2016) has recently proposed a different conception of constitutive luck, which for lack of space, I cannot discuss here.

[20] Strawson calls this "responsibility of the *heaven-and-hell* variety".

[21] See http://people.socsci.tau.ac.il/mu/usherlab/files/2018/07/Robust_Ommisions.pdf for an initial account on how this theory can be extended to account for responsibility for effects of omissions.

[22] Special thanks are due to Nick Zangwill for countless discussions and critical suggestions on various versions of the manuscript. I also want to thank Zohar Bronfman, John Martin Fischer, Michael Herrmann, David Lagnado, Neil Levy, Yair Levy, Alfred Mele, Adina Roskies, Nicholas Shea, Chandra Sripada, and David Wideker for very helpful comments on previous versions of the manuscript and Aaron Kravitz for assistance with the computer simulation and English editing. I acknowledge funding from the Binational (Israel-USA) Science Foundation (2014612).

References

Arpaly, Nomy. (2003). Unprincipled Virtue: an Inquiry into Moral Agency. New-York: OUP.

Balaguer, Mark. (2014) *Free Will*. The MIT Press.

Beebee, Helen. (2013) *Free Will*. Palgrave Macmillan.

—— and Mele, Alfred. (2002) 'Humean Compatibilism'. *Mind*, 111, 201–223.

Berofsky, Bernard. (2006) 'Global Control and Freedom'. *Philosophical Studies*: 131, 419–445.

——. (2012). *Nature's challenge to Free Will*. Oxford, OUP.

Bratman, Michael, E. (1987) *Intentions, Plans and Practical Reason*. CSLI Publications.

—— (2000). Reflection, planning and temporally extended agency. *The Philosophical Review*, 109: 35–61.

Bronfman Zohar Brezis Noam, Tsetsos Konstantinos, Donner Tobias, and Usher Marius. (2015) 'Decisions reduce sensitivity to subsequent information'. *Proceedings. of the Royal Society London, B*, 282.

Campbell, John. (2010) Control variables and mental causation. *Proceedings of the Aristotelian Society*, 110: 15–30.

Coates, Justin and Svenson, Philip. (2013) 'Reasons-responsiveness and degrees of responsibility'. *Philosophical studies*, 165: 629–645.

Davidson, Donald. (1980) *Essays on Actions and Events*, Oxford: OUP.

Deery, Oisín and Nahmias, Eddy. (2017). Defeating Manipulation Arguments: Interventionist causation and compatibilist sourcehood. *Philosophical Studies*, 174(5): 1255–1276.

Duffy, J. D. and Campbell, J. J. (1994). The regional prefrontal syndromes: a theoretical and clinical overview. *Journal Neuropsychiatry Clinical Neuroscience.*, 6(4): 379–387.

Dennett, Daniel. (1981) *True Believers: The intentional strategy and why it works*. In A. F. Heath (ed.), *Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford*. Clarendon Press. pp. 150–167.

Fara, Michael. (2008) 'Masked abilities and compatibilism'. *Mind*, 117: 843–65.

Feltz, Adam. (2013) 'Pereboom and premises: asking the right question in the experimental philosophy of free-will'. *Consciousness and Cognition*, 22: 53–63.

Fischer, John Martin (1994). *The Metaphysics of Free Will*. Malden MA: Blackwell.

—— (2004). 'Responsibility and Manipulation'. *Journal of Ethics*, 8: 145–177.

—— (2012). 'Semicompatibilism and its rivals'. *Journal of Ethics*, 16: 117–143.

——, and Ravizza Mark. (1998). *Responsibility and Control: a Theory of Moral Responsibility Cambridge*, UK: CUP.

Frankfurt, Harry. (1969) 'Alternative Possibilities and Moral Responsibility'. *Journal of Philosophy*, 66: 829–839.

——. (1971) 'Freedom of Will and the Concept of a Person'. *Journal of Philosophy*, 68: 5–20.

——. (1978) The problem of action. *American Philosophical Quarterly*, 15: 157–162.

Horn, David, & Usher, Marius (1990). 'Excitatory–inhibitory networks with dynamical thresholds'. *International Journal of Neural Systems*, 1, 249–257.

Holton, Richard. (1999) 'Intention and the weakness of will'. *Journal of Philosophy*, 99: 241–262.

Hurley, Susan (1993). 'Justice without constitutive luck'. In A. P. Griffith (Ed.), *Ethics*. Cambridge, CUP,179–212.

Huq, S. F., Garety, P. A., and Hemsley, D. R. (1988) 'Probabilistic judgments in deluded and non-deluded subjects'. *Quarterly Journal of Experimental Psychology A*, 40: 801–12.

Kane, Robert. (1998). *The significance of Free Will*. Oxford, OUP.

Kelso, J. S. (1997). *Dynamic Patterns: The Self-Organization of Brain and Behavior*. MIT press.

Lagnado, David and Gerstenberg, Tobby. (2015) 'A difference-making framework for intuitive judgments of responsibility'. *Oxford Studies in Agency and Responsibility*, D. Shoemaker(Ed.).

Latus, Andrew. (2003) 'Constitutive luck'. *Metaphilosophy* 34: 460–475.

Levy, Neil. (2012) *Hard Luck. How luck undermines Free Will & Moral Responsibility*. Oxford, OUP.

——. (2009) 'Luck and history sensitive compatibilism'. *Philos. Q.* 59(235), 237–251.

——. (2015). Luck and Manipulation Cases: A Response to Professor Haji. *Dialogue*, 54 (4): 633–646.

Lombrozo, Tania (2010). Causal explanation pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, 61: 303–332.

McKenna, Michael. (2001) 'Review of "Responsibility and control: a theory of Moral Responsibility by John Martin Fischer and Mark Ravizza"', *Journal of Philosophy*, 98: 93–100.

——. (2004) 'Responsibility and globally manipulated agents'. *Philosophical Topics*. 32: 169–92.

——. (2008) 'A hard-line reply to Pereboom's four-case manipulation argument'. *Philosophy and Phenomenological Research*, 77(1): 142–159.

——. (2013) 'Reasons-responsiveness, Agents and Mechanisms'. In *Agency and Responsibility*, Vol. 1, D. Shoemaker, Ed., Oxford Univ. Press, pp. 151–183.

Mele, Alfred. (2000) 'Reactive Attitudes, Reactivity, and Omissions'. *Philosophy and Phenomenological Research*, 61: 447–452.

——. (2005). 'A critique of Pereboom's four-case argument for Incompatibilsm'. *Analysis*, 65: 75–81.

——. (2006) 'Fischer and Ravizza on Moral Responsibility'. *Journal of Ethics*, 10: 283–294.

——. (2006) *Free Will and Luck*. Oxford, OUP.

Complexity Labs. (2016) "Bifurcations & Attractors," http://complexitylabs.io/bifurcations-attractors/.

Moore, G. E. (1912) *Ethics*. Ed. W. H. Shaw. Oxford, OUP, 2005.

Moore, Michael. (2010) *Causation and responsibility: an essay in law morals and metaphysics*. OUP.

Murray, Dylan and Lombrozo, Tania. (2016) 'Effects of manipulation on attribution of causation, free-will and moral responsibility'. *Cognitive Science*: 1–35.

Nagel, Thomas. (1979) 'Moral luck'. In *Mortal Questions*. New York: Cambridge Univ. Press, 24–38.

Nagel, Ernest. (1977) 'Goal-directed Processes Biology', *The Journal of Philosophy*, 74, pp. 271.

Nelkin, Kay Dana. (2017) 'Difficulty and degrees of moral praiseworthiness and blameworthiness', *Nous*, 50: 356–378.

Nickerson, Raymond. (1998) 'Confirmation bias: A ubiquitous phenomenon in many guises'. *Review of General Psychology*, 2(2): 175–220.

Newell, A. and Simon H. A. (1972) ''*Human Problem-Solving*''. Englewood Cliffs, NJ: Prentice Hall.

Pereboom, Derk. (2001) *Living without Free Will*. Cambridge CUP.

——. (2005) 'Defending Hard Incompatibilism'. *Midwest Studies*, 29, 228–47.

——. (2006) 'Reasons-responsiveness, alternative possibilities, and manipulation arguments against compatibilism: reflection on John Martin Fischer's My Way'. *Philosophy Books*, 47, 198–212.

Pritchard, Duncan. (2005) *Epistemic Luck*. Oxford, Claredon Press.

Rescher, Nicholas. (1995). *Luck: The brilliant randomness of everyday life*. Farrar, Straus & Giroux.

Roskies, Adina. (2006) 'Neuroscientific challenges to free will and responsibility'. *Trends in Cognitive Sciences*, 10:419–23.

——. (2016) 'Decision-Making and Self-Governing Systems'. *Neuroethics*.

Sartorio, Carolina. (2014) 'Sensitivity to Reasons and Actual Sequences'. In *Oxford Studies in Agency and Responsibility* (Shoemaker, ed., 2015)

——. (2016a). '*Causation and Free Will*'. Oxford: OUP.

——. (2016b) 'A partial defense of the Actual Sequence Model of Freedom'. *Journal of Ethics*, 20: 107–120.

Sehon, Scott. (1997) 'Deviant Causal Chains and the Irreducibility of Teleological Explanations'. *Pacific Philosophy Quarterly*, xxviii (1997): 195–213.

Smilanky, Saul. (2000) *Free-Will and Illusion*. Oxford, OUP.

Sripada, Chandra. (2012) 'What makes a manipulated agent unfree'. *Philosophy and Phenomenological Research*, 85: 563–593.

——. (2015) 'Self-expression: A deep self theory of moral responsibility'. *Philosophical Studies*: 1–30.

——. (2017) 'Frankfurt's Unwilling and Willing Addicts'. *Mind*, 126: 781–815.

Strawson, Galen. (1994) 'The Impossibility of Moral Responsibility'. *Philosophical Studies*, 75: 5–24.

Stump, Elenore. (2002) 'Control and causal determinism', in S. Buss and L. Overton Eds., *Contours of Agency*, Cambridge, MIT.

Usher, Marius (2006). 'Control, choice and the convergence / divergence dynamics: a compatibilistic probabilistic theory of free will'. *Journal of Philosophy*, 103, 188–213.

Wilson, George. (1989) *The Intentionality of Human Action*, Stanford: University Press.

Woodward, James. (2006) 'Sensitive and Insensitive Causation'. *The Philosophical Review*, 115: 1–50.

Wolf, Susan. (1993) *Freedom within reason*. New York, NY, OUP.