# CONTROL, CHOICE AND THE CONVERGENCE/DIVERGENCE DYNAMICS: A COMPATIBILISTIC PROBABILISTIC THEORY OF FREE WILL*

A theory of free will needs to account for the type of control that human agents exercise in free actions. Understanding this type of control is important for central issues in moral philosophy, such as those concerning responsibility and desert; for one can hardly be thought to be morally responsible for actions one does not control.[1] Yet a proper understanding of control is missing in both compatibilist and libertarian theories. Is the existence of alternative possibilities for actions, necessary or sufficient for this type of control? If it is neither, would there be any way to explain in what way control over actions is consistent with causal determination but is diminished in situations that involve *covert nonconstraining* (CNC) compulsion, such as Skinnerian indoctrination or value engineering? As observed by Robert Kane (*op. cit.*), the clashing intuitions of some of the parties in this debate have brought the field to a somewhat deadlocked position.

Here I argue that the control needed for exercising the type of free will required for responsibility and autonomy does not depend on whether determinism or indeterminism is true. Instead, I will try to show that the critical element for the presence of responsibility and autonomy is a more complex dynamical property involving a succession of convergence and divergence dynamics. The convergence dynamics (attractors) will be shown to enable teleological (goal directed) behavior, while the divergent dynamics involve bifurcations in

* Special thanks to Nick Zangwill for comments on successive drafts of the paper and countless discussions and suggestions that helped me develop and clarify these ideas. I also want to thank Ariel Kernberg, Eytan Ruppin, Orly Shenkar, and Dan Tidhar for very stimulating discussions and comments, and Alfred Mele and Saul Smilansky for a critical reading of the manuscript. A simplified version of some of these ideas in the context of neurocomputational models was presented in a technical report (Marius Usher and Eytan Ruppin, "Free Will in Light of Chaotic Systems," TR 197/91 (1991), Institute of Computer Science, Tel-Aviv University, Israel).

[1] For good discussions, see Randolph Clarke, "Incompatibilist (Nondeterministic) Theories of Free Will" (The Stanford Encyclopedia of Philosophy, 2005); Robert Kane, *The Significance of Free Will* (New York: Oxford, 1996); John Martin Fischer, *The Metaphysics of Free Will: An Essay on Control* (Cambridge: Blackwell, 1994); Alfred Mele, *Autonomous Agents: From Self-control to Autonomy* (New York: Oxford, 1995); Saul Smilansky, *Free Will and Illusion* (New York: Oxford, 2000).

0022-362X/06/0000/000–000 

choice and in character formation. This opens the door to a compatibilist theory that does not clash with powerful intuitions of open-ended choice in deliberation.[2] In my approach I use a probabilistic framework for two reasons. First, *physical* laws may be indeterministic, in which case the causal relation between the brain states involved in choice and the ensuing actions is probabilistic. Second, even if the *physical* laws are deterministic, it does not follow that mental states are sufficient for uniquely determining decisions and actions, due to the way that the mental states are realized in physical states and to the probabilistic nature of the environmental causes. This will involve appealing to the notion of *constitutive luck*,[3] which is central to this account.

I start by examining some aspects of free will that an agent needs to possess in order to be responsible for her actions, focusing on the question of whether she needs to be able to act otherwise, and on some restrictions posed by psychological overdetermination (impatient readers who are familiar with these issues can jump to the next section). I then deploy a theory based on a mechanism of *guidance control* that enables teleological behavior, and which assumes that action selection and control involve a succession of bifurcation/attractor states. I argue that the theory solves a number of puzzles about autonomy and CNC-compulsion.

## I. THE PRINCIPLE OF ALTERNATIVE POSSIBILITIES

A central issue in the current debate between compatibilists and libertarians is whether there must be *alternative possibilities* for actions that we are morally responsible for. Let us call the statement that alternative possibilities are necessary for responsibility the *principle of alternative possibility* (PAP). A key issue in the debate on PAP is Harry Frankfurt's counterfactual scenario, where, although there is no alternative possibility to an action, once the action is performed we have a compelling intuition that the agent is responsible.[4] The scenario involves a counterfactual intervener, who overrides the agent's libertarian choice (say, by directly activating her brain choice cen-

[2] For eloquent illustrations of the difficulty to give up on these intuitions, see Thomas Nagel, *The View From Nowhere* (New York: Oxford, 1986); John Searle, "Free Will as a Problem in Neurobiology" (talk given at Royal Institute of Philosophy in February 2001, based on "Consciousness, Free Action and the Brain," *Journal of Consciousness Studies,* x (year?): ???–??).

[3] Nicholas Rescher, *Luck: The Brilliant Randomness of Everyday Life* (Pittsburgh: University Press, 1995).

[4] Harry Frankfurt, "Alternate Possibilities and Moral Responsibility," this JOURNAL, LXVI, 23 (December 1969): 829–39.

ters), only if she showed a *sign* of choosing *B* but not, in the actual case, when she chooses *A*. In the actual case, as no external intervention took place and the agent carried her choice and action on her own, we feel that the agent is responsible despite the fact that the alternative action (or even choice) was blocked by the counterfactual intervener. This argument is open to the objection that despite the intervener's effort, the agent still maintains some *flicker of freedom*; she may still *try* to choose otherwise and much debate has focused on the question of whether such a flicker of freedom is robust enough to ground responsibility.[5] For example, if the sign (or its absence) is deterministically related to the available options, *A* or *B* (as is needed if the contravener is to prevent the alternative action on the basis of the sign), one may argue that the presence (or absence) of the sign *is*, in effect, the choice, and thus the counterfactual intervention comes too late; the agent *does* have a robust alternative after all. More recently, a more refined counterfactual intervention scenario has been presented by Derk Pereboom.[6]

Although some libertarians insist on the flicker of freedom objection, I believe that robustness considerations in such scenarios give us a good reason to prefer an account of responsibility that does not rely on PAP, unless an independent argument in favor of PAP convinces us otherwise. One of the best articulated arguments in support of PAP is the *no-matter-what* principle advanced by Peter van Inwagen.[7] According to it, "if it is a fact that *p*, an agent is morally responsible

[5] For detailed discussions, see Fischer, *The Metaphysics of Free Will*, and his "Recent Work on Moral Responsibility," *Ethics*, CX (October 1999): 93–139; Kane; Eleanor Stump, "Libertarian Freedom and the Principle of Alternative Possibilities," in Daniel Howard-Snyder and Jeff Jordan eds., *Faith, Freedom and Rationality* (Totowa, NJ: Rowman and Littlefield, 1996), pp. ???–??; S??. Goetz, "Stumping for Widerker," *Faith and Philosophy*, XVI (1999): 83–89; Mele, "Soft Libertarianism and Frankfurt-style Scenarios," *Philosophical Topics*, XXIV, 2 (1996): 123–41; Mele and David Robb, "Rescuing Frankfurt-style Cases," *Philosophical Review*, CVII (1998): 97–112; David Widerker, "Libertarianism and Frankfurt's Attack on the Principle of Alternative Possibilities," *Philosophical Review*, CIV (1995): 247–61.

[6] Pereboom, "Alternative Possibilities and Causal Histories," *Philosophical Perspectives: Action and Freedom*, XIV (2000): 119–37. In the scenario he imagines, the sign is a moral thought in favor of *B*, which occurs indeterministically and is *necessary* for the choice of *B*, but does not deterministically cause it; the agent may still indeterministically choose either *A* or *B*. If the agent chooses *A* (on her own), because the moral reason did not occur and thus no intervention took place, the agent is responsible, and the alternative possibility—in which the moral thought occurred—does not seem robust enough to ground the agent's responsibility. This is because, unlike in the original scenario, the sign does not settle the choice: despite the occurrence of the moral thought the agent may still decide against it.

[7] Van Inwagen, "Fischer on Moral Responsibility," *The Philosophical Quarterly*, XLVII, 188 (July 1997): 373–81. A similar principle, the *W*-principle (for "what-else-could-I-

for the fact that *p* only if that agent was once able to act in such a way that it would not have been the case that *p*" (*ibid.*, p. 376). To support the *no-matter-what* principle, van Inwagen presents a number of variants of Frankfurt scenarios, but without counterfactuals, such as:

> ($S_1$) "I am supposed to take the serum upriver to the plague-driven village. But I get drunk and miss the boat. Taking the boat is the only possible way to get to the village. Soon after the boat leaves the dock, it strikes a rock and sinks. Hundreds of villagers who would have been saved by the serum die" (*ibid.*, p. 378).

Van Inwagen argues that the agent cannot be held responsible for the death of the villagers (though he acted *irresponsibly* and is guilty of dereliction of duty) for the obvious reason that he could not have saved them anyhow. Furthermore, the same conclusion should apply, according to van Inwagen, to the counterfactual Frankfurt scenarios. To evaluate the *no-matter-what* principle, consider a situation involving causal overdetermination. The following scenario is modified (to include overdetermination) from van Inwagen.

> ($S_2$) Gunnar and Rifler desire Ridley's death and plan, independently, to murder him. It so happens that, without knowing of each other, both of them shoot Ridley at exactly the same moment. The bullet shot from Gunnar's gun and the bullet shot from Rifler's rifle hit Ridley in his chest and kill him. A pathological investigation shows that Ridley was still alive when both of the two bullets hit him (it took him few seconds to die) and that each of the bullets would undoubtedly and independently have caused his death. Are either Gunnar or Rifler "morally responsible for the fact that Ridley died and his children are now orphans" (*ibid.*, p. 379)?

The interesting feature of this scenario, is that by appealing to the no-matter-what principle, both Gunnar and Rifler can claim *not* to be responsible for Ridley's death. Whatever Gunnar had chosen, Ridley would still have died as the result of the other bullet. But this clearly leads to a highly counterintuitive conclusion that no one is respon-

---

do"), was advocated by Widerker, "Frankfurt's Attack on the Principle of Alternative Possibilities: A Further Look," *Philosophical Perspectives: Action and Freedom*, XIV (2000): 181–201. I agree with Fischer's criticism that in the absence of independent evidence, this principle begs the question against the compatibilists, who insist that the relevant issue is what "I ought to do" and not whether "I can do it" (Fischer, "Recent Work on Moral Responsibility"). It may seem that the type of argument one can put in support of the *W*-principle is the one articulated by van Inwagen and therefore the conclusions of the analysis below will apply too.

sible for Ridley's death. This indicates that there are situations (such as causal overdetermination)[8] where the association between responsibility and the no-matter-what principle does not hold and thus that the intuition behind this principle can be resisted.[9]

The rejection of PAP is also consistent with the views of some libertarians who concede that, while responsibility can be upheld in Frankfurt scenarios in the absence of robust alternative possibilities, the lack of determinism does rule out responsibility because it does not allow the agent to be the true *source* of the action. As formulated by Michael McKenna,[10] "Source incompatibilists hold that determinism does rule out free will. But it does so, not because it rules out alternative possibilities, but instead, because, if true, the sources of an agent's actions do not originate in the agent but are traceable to factors outside her" (*ibid.*, p. 201). A related position has been taken by Kane, who concedes that some responsibility-bearing actions are determined by antecedent mental states (or willings), but insists that some indeterminism in the causal chain that precedes an action is necessary for the agent to possess *ultimate responsibility* (UR) for that action. This is because he fears that otherwise, the determinism would shift the source of the explanation of the action beyond the agent: "if these willings were in turn caused by something else, so that the

[8] In cases of overdetermination, one may allow "sum events" (the two shootings) as causes in order to justify the intuition that each of the agents has partial responsibility for the death of Ridley. This, however, does not detract from the point that the no-matter-what principle, as applied to each agent alone, is invalid.

[9] The same logic can be used to develop a Frankfurt-type scenario that eliminates the need for a sign predictive of the choice. Consider for example, the following variation on a scenario proposed by Mele and Robb. The agent is faced with a choice between two options, *A* and *B*. Unknown to the agent, the intervener initiates in the agent's brain a deterministic process, *D*, parallel and independent from the indeterministic process, *L* (for Libertarian), assumed to produce a free-choice in the libertarian sense. The *D*-process hits the *A* choice-node (the intervener wants the agent to choose *A*) and the *L*-process hits the *A* or the *B* node (indeterministically) at precisely the same moment, *t*. Either of them (in isolation) is sufficient to cause a decision; however, *D* is stronger, so that in case of divergence its effect prevails. In the case that both processes hit the *A* unit, we seem to be in the situation where the agent chooses *A* and is responsible for it, despite the fact that she could not have *chosen* otherwise (as the *D* process would have overwritten the *L* process). One objection to this could be based on the no-matter-what principle. The agent could claim to be absolved of responsibility since, even if his *L*-process chose otherwise, the same action would have been performed. This case, however, is rather similar to scenario ($S_2$) above, where two agents shoot (and kill) a person independently. It thus seems that the agent at least shares responsibility for her action despite the fact that she could not have chosen otherwise.

[10] McKenna, "Robustness, Control and the Demand for Morally Significant Alternatives: Frankfurt Examples with Oodles and Oodles of Alternatives," in McKenna and Widerker, eds., *Moral Responsibility and Alternative Possibilities: Essays on the Importance of Alternative Possibilities* (Burlington, VT: Ashgate, 2002), pp. ???–??.

explanatory claims could be traced back further to heredity or environment, to God, or fate, then the ultimacy would not lie with the agents but with something else" (*op. cit.*, p. 4).

In order to further support the necessity of indeterminism for UR, Kane gives his *manipulation* argument. According to Kane, we think that an agent, whose actions are the outcome of CNC control by another agent, lacks autonomy and control and UR over her actions, even when those are caused by her own (induced) motivations. Kane uses the illustration of B.F. Skinner's Walden Two, a Utopian community where peoples' values are engineered by behavioral scientists without inducing any sense of coercion. According to Kane, the Walden Two members lack autonomy and UR over their actions, which issue from motivations induced by the indoctrinator. Yet, as Kane argues, it is not clear in what way determination by physical law differs from CNC control, since in both cases the agent's actions appear to be determined (thus lacking alternative possibilities) and beyond her control. According to Kane, only an indeterministic process taking place during character formation, from which the action issues, can salvage UR (*op. cit.*, p. 4).

There are two problems with these arguments. First, the concept of UR has been argued to be logically incoherent and, second, it is not clear that indeterminism can help, as it runs up against the luck counterargument.[11] The former point has been compellingly articulated by Galen Strawson,[12] who argued that an agent acts the way she acts because of the way she *is*, and thus to have UR, the agent needs to be responsible for the way she is. This leads to infinite regression or to a controversial notion of self-causation. Although indeterminism in the choice mechanism may allow the agent to act in a different way (and perhaps form a different character) from the way she actually does, this solution is vulnerable to the luck objection. As argued by Mele, "If there is nothing about the agent's powers, capacities, states of mind, moral character, and the like that explains the

[11] Good presentations of this argument appear in: Clarke, "Free Choice, Effort and Wanting More," *Philosophical Explorations,* II (1999): 20–41; Mele, "Review of *The Significance of Free Will* by Robert Kane," this JOURNAL, XCV, 11 (November 1998): 581–84, and his "Ultimate Responsibility and Dumb Luck," *Social Philosophy and Policy,* XVI (1999): 274–93; van Inwagen, "Free Will Remains a Mystery," *Philosophical Perspectives: Action and Freedom,* XIV (2000): 1–20; but for a counterargument, see Mark Balaguer, "A Coherent, Naturalistic, and Plausible Formulation of Libertarian Free Will," *Nous,* XXXVIII, 3 (2004): 379–406.

[12] Strawson, "The Unhelpfulness of Indeterminism," *Philosophy and Phenomenological Research,* LX, 1 (January 2000): 149–55.

difference in outcome, then this difference is just a matter luck."[13] Notice, however, that the argument from luck, can be deployed even against compatibilistic theories, since inborn biological factors or environmental influences during infancy (that influence the forming character) are also beyond the agent's power.[14]

I believe that a strong conception of UR may, indeed, be incoherent. This issue is controversial,[15] and I will not insist on it here. I will argue instead for a weaker version of ultimate responsibility (UR*), which does not involve self-causation. Rather, UR* requires that the action is one that the agent produces by exerting a certain type of control (to be elaborated below) and which, in addition, is not uniquely determined by anything[16] besides or antecedent to the agent. This prevents Kane's concern about the shift of responsibility beyond the agent, and I will argue that it is not blocked by determinism. Central to the theory I deploy is the acceptance of *constitutive luck*, which should help us resist the idea that heredity or influences of early infancy undermine responsibility. As eloquently discussed by Nicholas Rescher, "identity must precede luck" (*op. cit.* p. 157); we are not born as "bare particulars" that are later allocated with properties. Rather, the properties we are born with are part of our identity and thus are the basis for attributions of responsibility and desert. In section IV, I will argue that this concept of constitutive luck applies also to indeterministic events that affect moral choices of the type considered by libertarians.

Turning now to the manipulation argument, there are two strategies available to compatibilists. The first one, *hard compatibilism*, is to deny the claim that CNC manipulation necessarily eliminates responsibility.[17] As argued by D???? Blumenfeld (*ibid.*), the most persuasive manipulation scenarios are the ones where the agent's personality is suspended by the controller, who temporarily imposes his motivational structure on the agent. Hard compatibilists admit that, in such situations, when the agent recovers (her previous per-

[13] Mele, "Review of *The Significance of Free Will* by Robert Kane," p. 583.

[14] See Smilansky.

[15] See Balaguer for arguments that indeterminism in torn decisions does not undermine control and authorship of actions.

[16] Excluding "the state of the whole universe," which is a noninformative *source of explanation.*

[17] See, for example, Frankfurt, *The Importance of What We Care About* (New York: Cambridge, 1988); D???. Blumenfeld, "Freedom and Mind Control," *American Philosophical Quarterly,* XXV, 3 (July 1988): 215–28; Tomis Kapitan, "Autonomy and Manipulated Freedom," *Philosophical Perspectives: Action and Freedom,* XIV (2000): 1–103 (pages right?).

sonality), she is not to blame for actions she did during the interval she was under the control of the manipulation. They claim, however, that a manipulated agent whose value system is transformed forever, and who (due to the manipulation) fully identifies with the evil acts she does as a result of her new (and now persistent) motivational states, is responsible (and can be blamed) for her acts. The second strategy, *soft compatibilism*, concedes that responsibility is undermined even in manipulation scenarios of the latter form.[18] I believe that which of these two strategies we endorse hinges to a large degree on whether we think that the manipulated agent and the pre-manipulated one share the same personal identity.[19] Other cases of manipulation, such as the indoctrination of young children, where a pre-manipulated agent did not exist, trigger more ambiguous intuitions.[20]

In light of these conflicting intuitions, I take the manipulation argument for PAP to be indecisive. Nevertheless, I believe that this argument does pose a serious challenge to compatibilism. Even if responsibility is not undermined by manipulation, as hard compatibilists hold, another feature of a free agent, her autonomy, does seem to be undermined in CNC manipulation cases. Agents strongly resent to having their autonomy violated,[21] and this is an important property of free agency that needs to be accounted for. The challenge is to account for the difference between CNC control and causal determination of action by physical law, with regard to the agent's autonomy. This requires an account that does not involve indeter-

[18] See Mele, *Autonomous Agents*; Ishiyaque Haji, *Deontic Morality of Control* (New York: Cambridge, 2000). Mele provides a number of colorful illustrations of CNC manipulation. One such example involves a young philosopher, Beth, whose values and pro-attitudes are modified (at the request of her dean) by "new-wave" engineers to match those of Ann (a more industrious philosopher in the same department). According to Mele, after the modification, although Beth and Ann became psychological twins, Ann is autonomous (and responsible for her actions) while Beth is not.

[19] Both Mele and Haji, for example, hold that personal identity is maintained under manipulation.

[20] Consider, for example, Haji's scenario, of a monk raised at the Franciscan order, and who fully endorses (and is happy with) the value system he was indoctrinated with. I believe that whether the monk is responsible (or not) for good actions that follow from his engineered values is debatable.

[21] Consider the following case of manipulation, discussed by Patricia Greenspan ("The Problem with Manipulation," *American Philosophical Quarterly*, XL (2003): 155–64). In this scenario, the students in a psychology class have conditioned their behaviorist professor to move out from his preferred corner of the room by manifesting various signs of comprehension and of attentiveness. In this case, although the professor makes a free and uncoerced choice, his decision is manipulated to fit someone else's end and thus his autonomy is violated.

minism and can nevertheless distinguish between determination and compulsion.[22]

To conclude, thus far: in the absence of a decisive positive argument favoring PAP,[23] I believe that a theory of free will and responsibility should not assume that, for any action the agent is responsible for, there has to be an alternative possibility that is accessible from the state of the world that preceded the action. (Note, that rejecting PAP does not rule out indeterminism playing a role in our being able to justify the commonsense intuition that we are *initiators* of actions, as suggested by some *soft libertarian* theories.[24])

The theory I propose here is therefore not constrained by the need to respect PAP and is thus compatibilist about determinism and moral responsibility. Nevertheless it does not presume the truth of determinism. Although it may be impossible to establish metaphysical claims, solely from scientific evidence and theory, I believe that quantum mechanics makes it not unlikely that the laws of physics are indeterministic. While it is generally accepted that compatibilism does not depend on the truth of determinism (and that it only insists that determinism does not undermine free will and responsibility), the task of showing that it is robust enough to account for control and responsibility in a world where the relation between the mental states of agents and their action is probabilistic, has not been much addressed. The aim of this paper is to do so, while at the same time providing a compatibilist distinction between cases of physical determination and compulsion.

## II. DOXASTIC FREEDOM AND PSYCHOLOGICAL OVERDETERMINATION

A compatibilistic theory that does not respect PAP needs, nevertheless, to respect some weaker "freedom" principles. Frankfurt's origi-

[22] Obviously such an account cannot rely on the distinction between determination by another agent versus determination without such an agent; for example, in *Autonomous Agents*, Mele, who offers a compatibilist account of autonomy, agrees that if an agent's values are modified as a result of natural forces (say due to strange electromagnetic fields during to a trip to the Bermuda triangle) the agent is no more autonomous (and deserving of blame/praise for actions performed under those modified values) than an agent whose values were explicitly set up by evil value engineers.

[23] Note also that contra the libertarian intuitions for PAP there are equally compelling intuitions in support of the idea that we would be responsible for actions even in a deterministic world. In his "Recent Work on Moral Responsibility," Fischer has argued that if we were to find out, from a panel of physicists, that the laws of nature are ultimately deterministic, we would not feel that all basis of responsibility attribution and moral justice has evaporated.

[24] See Mele, "Soft Libertarianism and Frankfurt-style Scenarios"; Clarke, "Modest Libertarianism," *Philosophical Perspectives: Action and Freedom*, XIV (2000): 21–45.

nal suggestion was that moral responsibility does not require alternative possibilities, but that it requires, nevertheless, that the agent has not performed her action *only* because she could not do otherwise. I rely here on a particular interpretation of this idea presented by Robert Cummins,[25] according to which if an agent believes that a desire is irresistible and this belief is part of the reason why she performed the action, she is not responsible for so acting. But, if the fact that the agent *could not do otherwise* did not play any role, as a psychological variable, in the deliberation, the agent is responsible for her action. Accordingly, responsibility is compatible with physical determinism, since an agent "may not know whether determinism is true, and even if she knows, she need not know which acts are determined, and thus the lack of an open alternative plays no role in her deliberation" (*ibid.*, pp. 412–13).

In line with this analysis, I assume that an agent is responsible for an action only if knowledge or belief in lack of alternative possibilities is not a part of the deliberation that leads to the action. For example, we tend to absolve of responsibility a guard who does not oppose an armed robbery because he incorrectly believed (due to good reasons) that his pistol was empty of ammunition. Although, in reality the guard had an alternative possible action, he believed he did not have one, and this was the decisive factor in his not opposing the robbery.[26]

This doxastic condition (the agent believing an alternative option to exist), is too weak for responsibility, as it is satisfied in cases of psychological overdetermination, such as addiction and phobias.[27] In order to rule out responsibility in such situations, within a compatibi-

[25] Cummins, "Could Have Done Otherwise," *The Personalist,* LX, 4 (October 1979): 411–14.

[26] The importance of epistemological cognitive requirements of responsibility has also been emphasized by Carl Ginet ("The Epistemic Requirements for Moral Responsibility," *Philosophical Perspectives: Action and Freedom*, XIV (2000): 267–77), and by Kapitan ("Doxastic Freedom: A Compatibilistic Alternative," *American Philosophical Quarterly,* XXVI, 1 (January 1989): 31–42). Kapitan formally developed a principle first suggested by Daniel Dennett (*Elbow Room* (New York: Oxford, 1984)), according to which the relevant sense of *can* for responsibility is relative to the knowledge of the agent; "something is epistemically possible for Jones if it is consistent with everything Jones already knows," epistemic possibility supplying the "useful notion of can" (p. 184). Kapitan furthermore formulated two conditions that are necessary for a free action (in the doxastic sense): *efficiency* and *contingency*. The former involves the agent's presumption that he would *A*, were he to choose to *A* (and the converse). The latter involves the agent's presumption that the choice is as yet contingent (where the contingency is evaluated relative to the set of cognitive states held by the agent during the choice).

[27] Psychological overdetermination is related to actions performed under "irresistible" desires; see Mele, *Springs of Action* (New York: Oxford, 1992), chapter 5, for further discussion.

list framework, John Fischer[28] has introduced the requirement of reason sensitivity. Accordingly, while responsibility does not require the ability to do otherwise (in an identical situation), it does require sensitivity to the input of reasons: the agent should be able to do otherwise, if some (additional) "sufficient" reasons are provided.[29]

To conclude, I will assume that a theory of free will and responsibility needs to allow agents two types of freedom: doxastic ability to do otherwise, and reason sensitivity. In addition, since an agent can only be responsible for *intentional* actions that she controls, the theory needs to account for intentional control. The next section presents the theory, focusing on the key aspect of teleological control and its relation with determinism, indeterminism, and stable behavioral patterns that underlie the predictability of events.

### III. TELEOLOGICAL GUIDANCE CONTROL AND PREDICTABILITY

Intentional explanation of actions typically assume that actions are made for reasons and that having such reasons (more precisely, a combination of desires and relevant beliefs) is a causal factor in the generation of actions.[30] For example, we may explain the intentional action of an agent going into the kitchen, as being *caused* by her desire for coffee and her belief that she can find it in the kitchen. Yet, it seems that there is an important aspect of intentional explanation of actions that is not automatically captured in a crude causal account, and this is their teleological or goal-oriented character.[31] Thus we perceive the action of going into kitchen as directed towards a goal (for example, "the agent went into the kitchen *in order to* get coffee"). Here I contend that any intentional action is teleological (and goal oriented), even if its goal does not extend beyond the

[28] Fischer, *The Metaphysics of Free Will*.

[29] See Fischer and Mark Ravizza, *Responsibility and Control* (New York: Cambridge, 1998) for a more refined version of this account, which states that "an agent is morally responsible for an action insofar as it issues from his own, moderately reason-responsive mechanism" (p. 86), and A.C. MacIntyre, "Determinism," *Mind*, vol? (1957): 28–41, for an early similar proposal. Note also that even in typical cases of phobia, one may be responsive to some extreme/exceptional reasons: throwing flames into an agoraphobic's house, could persuade the latter to finally leave the place (Mele, "Soft Libertarianism and Frankfurt-style Scenarios"). The input sensitivity thus needs to take place not only under such "exceptional" situations.

[30] Donald Davidson, "Actions, Reasons, and Causes," this JOURNAL, LX, 23 (November 21, 1963): 685–700.

[31] See, for example, George M. Wilson, *The Intentionality of Human Action* (Stanford: University Press, 1989); S??. Sehon, "Teleology and the Nature of Mental States," *American Philosophical Quarterly,* XXXI, 1 (January 1994): 63–72, and Sehon, "Deviant Causal Chains and the Irreducibility of Teleological Explanations," *Pacific Philosophical Quarterly,* LXXVIII (1997): 195–213.

action itself (say "raising one's arm," or "driving a car on a specified trajectory") and that a necessary condition for such actions is the deployment of a type of control, which I label *teleological guidance-control* (TGC) and which I argue to be necessary for responsibility.[32] I start with an informal and intuitive grasp of what TGC is, using an example from Tomis Kapitan.[33] Having done so, I move to present a substantive account of the physical features of this type of control and its relation with causal laws and their properties, such as *predictability*, *determinism*, and *indeterminism*.

Consider the following scenario presented by Kapitan.

> ($S_3$) "…the pilot and the co-pilots of a passenger plane suddenly die, en route, due to a poison ingested before takeoff. The head steward, apprised of the dreadful situation, is faced with the task of guiding the plane to a safe landing. He knows nothing about flying the plane, but were he to press certain buttons and levers, and manipulate the steering mechanism in certain way—actions he is able to perform—the plane would land safely on the designated runway. As it is, he fiddles madly with the controls and manages to do something that results in the plane landing, though, unfortunately, not safely. All aboard perished, except for the steward himself, who survived with minor injuries. Should he be blamed for not bringing it about that the plane landed safely? Was he responsible for bringing it about that it landed in a way that all the passengers were killed?" (*ibid.*, p. 423).

In his analysis of this situation, Kapitan locates an important condition for an action to be under agent control. The way in which the action is issued needs to be *reliably* determined (or ensured) by what the agent has done. While the pilot has *guidance* control over the landing of the plane (she was able to ensure its landing), the steward did not (even if by luck he stumbled over the correct sequence of movements that landed the plane safely). Notice here, the key requirement of *reliable* determination. Plain determination is not enough. If the world is deterministic, both the steward and the pilot landings are determined (by the laws of nature and the state of the world). But how do we distinguish reliable from plain determination?

One way to cast this issue would be in terms of *predictability* (I will address the underlying metaphysics in a moment). We could say that for an event (or state of affairs) to be *reliably* determined by the state of a system at a given moment, its happening needs to be predictable

[32] The term "guidance control" is also used by Fischer, but in a different sense (see the section on "reason sensitivity."

[33] Kapitan, "Modal Principles in the Metaphysics of Free Will," in J. Tomberlin, ed., *Philosophical Perspectives: Metaphysics*, x (1996): 419–45.

(in some sense) from the state of the system at that moment. Such an idea was used by Daniel Dennett[34] in his *intentional stance* theory, in which we attribute intentional states to systems (or organisms) whose behavior shows a reliably predictable pattern (which is stronger than what can be predicted from Newtonian equations and is not derivable a la Laplace). Animals, plants, and even lightning bolts and thermostats, are relevant examples. An important feature of the kind of predictability in question is its being robust given variations (as illustrated in the pilot scenario). Characterizing control and responsibility in terms of this kind of predictability, however, has the disadvantage that it seems to make control and responsibility relative to one's scientific skills. In response to this worry, Dennett argued that intentional predictability is grounded on the fact that there are objective patterns that "impose themselves, not quite inexorably, but with great vigor, absorbing physical perturbation and variation" (*ibid.*, p. 27). What the nature of these objective patterns is, however, was not explained.

A main aim of this paper is to account for the objective structure that mediates the patterns enabling domains with predictability in the word of physical objects and organisms. The upshot, I argue, is that this requires a *teleological* system with goal-directed behavior, which will reach the same end state in face of perturbations. As many critics have noticed, however, "there is a prima facie tension between the common sense account of ourselves as agents and the scientific view of human beings as physical objects. 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 put into question."[35]

I endorse a causal theory according to which teleological behavioral patterns arise within a causal dynamical system. Such a theory, therefore, needs to account for the apparent tension between causal and teleological behavior, which is probably due to the fact that there appears to be no proper place for teleology in the physical laws of nature. A central principle of both deterministic (Newtonian mechanics and electromagnetism) and indeterministic (quantum mechanics) physical theories is that the change in the state of a system (and thus its future) depends only on its present state. Both theories have also fully reversible dynamics; no place is thus left for "ten-

[34] Dennett, *The Intentional Stance* (Cambridge: MIT, 1987).

[35] Sehon, "Deviant Causal Chains and the Irreducibility of Teleological Explanations," p. 197; see also Nagel, pp. 113–19, for a similar discussion.

dencies" to reach an end state, as was the case in Aristotelian physics. It is perhaps for these reasons, that teleological systems are most often characterized in relation to a "design" or a function.[36] Accordingly, an artifact is teleological because of its doing what it was designed to do; we and other biological organisms are teleological because of our doing what the Darwinian evolution "designed" us to do. While I agree that evolution helps to explain how teleological systems (of biological kind) emerged, I believe that their proper characterization should be independent of the evolutionary origin.[37] Here I propose a different account of teleological systems, possessing guidance control (TGC) and consistent with causal laws.[38]

Here is the main idea. Teleological systems require a property that is stronger than determinism and which involves a *counterfactual* type of determination, called an *attractor*. TGC systems (or agents) generate stable behavioral patterns, in which events (or states of the world) are being determined by a goal state of the system (agent), in a set of possible worlds similar to (and including) the actual one; the fact that the event is determined to take place in a set of counterfactual situations (similar to the actual one) reflects the requirement of being *reliably* determined. Critically, neither determinism nor indeterminism is sufficient or necessary for TGC, which is typically manifested at a macro-level of description, involving collective states. I explain these ideas below, starting with the conditions that enable the appearance of teleological goal-directed systems.

Consider, first, the mixing of a liquid between two connected containers, originally filled with hot and cold liquid, respectively. Under the second law of thermodynamics, an irreversible process of flow takes place between the containers, equalizing the temperatures. Although the behavior of individual molecules is subject to reversible causal laws (and in practice, contra Laplace, unpredictable), the temperature of the liquid in the two containers, a collective (or

[36] Dennett, *The Intentional Stance*; Ruth Millikan, "On Swampkinds," *Mind and Language,* XI (1996): 103–17.

[37] The over-reliance of theories of content on evolutionary teleology leads to some counterintuitive consequences. For example, swampmen (creatures that are accidentally created out of molecules in a swamp, with body and brain structures identical to ours, but without our evolutionary theory) have no beliefs and desires (see Millikan). Liberating teleology from its link to evolutionary design opens the way for swampmen liberation (L??. Anthony, "Equal Rights for Swamp-persons," *Mind and Language,* XI (1996): 70–75; Usher, "Comment on Ryder's SINBAD Neurosemantics: Is Teleofunction Isomorphism the Way to Understand Representations?" *Mind and Language,* XIX (2004): 241–48).

[38] See also Mele, "Goal Directed Action: Teleological Explanation, Causal Theories, and Deviance," *Philosophical Perspectives: Action and Freedom*, XIV (2000): 279–300.

macro) state variable, shows a teleological pattern: it *tends* toward a state of equilibrium. The teleology of this equilibrium state is due to its finality or *convergence* type of behavior; the macrostate *converges* toward the equilibrium, since it ends there even if perturbations are applied trying to change its course. (Other examples of collective macrostates displaying teleological behavior are the vortex of a whirlpool that sinks whatever small objects get close to it, or a lightning bolt that is attracted to the best object that conducts electricity in the ground. In both cases, although the trajectory of individual molecules is subject to causal physical laws (but is in practice unpredictable due to chaotic dynamics), the collective state that corresponds to the vortex (its center and angular momentum) or to the lightning (its center and momentum), displays a stable and predictable behavior. An end-state with this property is called an *attractor*. In an attractor dynamics, the "space of possibilities" shrinks as trajectories converge towards the attractor (see Figure 1 in the following section), leading to a decrease in *entropy*.[39] Crucially, the property of convergent or attractor dynamics is orthogonal to the determinism/indeterminism distinction—one can have determinism with or without convergence, and one can have indeterminism with or without convergence.[40] The role of attractor states in guidance control of actions is illustrated in the following example.

Consider a simple goal-directed system: a teleguided rocket. The mechanism that enables the rocket to display teleological behavior by "pursuing" the target (seen here as its goal) is based on a feedback loop and an error correction mechanism. This mechanism stabilizes the rocket's trajectory towards the target despite external perturba-

[39] The entropy (or the Shannon information) reflects the degree of uncertainty of the system (or the number of possible microscopic configurations, consistent with the macroscopic description (for example, temperature). Under the second law of thermodynamics, the entropy increases. This is the case for the mixing of liquids, but not in the other two examples (see next footnote).

[40] The deterministic Newtonian/Hamiltonian dynamics satisfy the Liouville theorem, according to which the volume of an area in the "phase space" (the space of all possibilities of the variables that define the system) remains constant as the trajectories evolve, and a similar property exists in quantum mechanics. In an attractor dynamics, the volume in the configuration space shrinks as the trajectories converge, producing a decrease in entropy. In either, a Newtonian or a quantum world, it is possible, however, that a subsystem (corresponding to a collective variable, for example, the tornado vortex) behaves according to attractor dynamics provided that the rest of the system, which involves other variables, compensates for it (by increasing entropy). Collective-state variables that control actions in the brain correspond to population codes or cell assemblies, whose behavior is characterized by attractor dynamics (see Figure 1). Other variables that are irrelevant to the controlling behavior (for example, heat released by the body) involve a divergence dynamics compensating for the decrease in entropy.

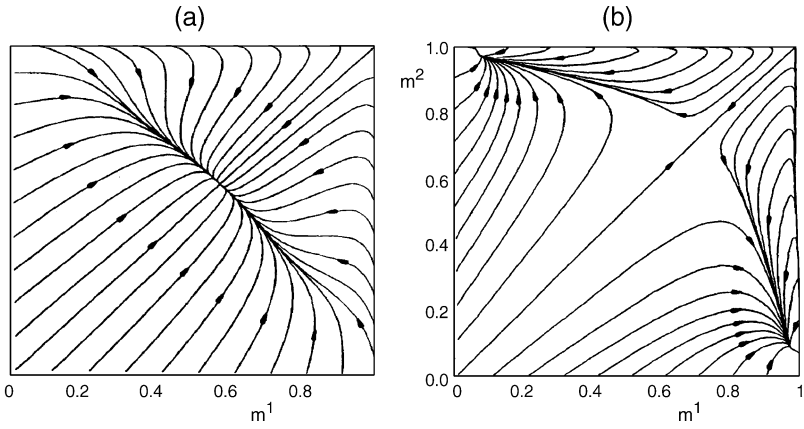


Figure 1: Attractor dynamics in a deterministic nonlinear neural system. The dynamics are indicated by the arrow flow, showing convergence to attractor points. (a) The single attractor corresponds to a scenario where all the trajectories converge at the same end point, corresponding to a case of overdetermination. If the various starting points correspond to informational-relevant inputs (reasons or data in favor of one or the other of the options) for a specific decision, the single attractor implies a total lack of input sensitivity; one cannot not do otherwise, even with different input relevant to the situation. (b) A multiple-attractor scenario, where although the trajectory is bound to end up in one of the two attractors, there are bifurcation points (input-relevant situations) where a different attractor is reached. Each attractor has a *basin of attraction*—a nonzero volume of states that map into that attractor. This ensures that input sensitivity is not too weak, holding not only for the most exceptional situations (see footnote 33??).[41]

tions. If a gust of wind makes the rocket deviate from its planned trajectory, the correction mechanism will bring it closer to it. This confers a teleological behavior on the rocket: it has a type of purpose or finality. When the rocket is launched repeatedly at the same target, the actual trajectory and the forces applied may vary, but the convergence upon the target prevails. *Stability under perturbations* is a property that distinguishes the rocket from nonteleological type of projectiles, which may happen to hit upon a target but do not *converge* on it. As with the water temperature, the rocket mechanism is characterized by an attractor. This means that it reaches its target not

[41] Reproduced from D. Horn and Usher, "Dynamics of Excitatory-inhibitory Networks," International Journal of Neural Systems, 1 (1990): 249–57.

only in the world corresponding to the actual situation, but also in similar possible worlds (say, when its trajectory is subject to wind perturbations or when the target attempts to escape). This makes such control mechanisms highly effective in achieving their goal, and although their action is purely causal, they generate the appearance of a teleological, purpose-oriented process.

The need for a correction-stabilization mechanism, in the action control of organisms, is obvious in an indeterministic world. This need exists, however, even if determinism holds. Since agents cannot have full knowledge of all the environmental inputs, a certain degree of uncertainty and randomness needs to be accepted as an inevitable feature of the environment.[42] Therefore, the problem that an agent faces is not how to get from $A$ at $t_1$ to $B$ at $t_2$ given the precise state of the world at $t_1$, but rather how to get from $A$ at $t_1$ to $B$ at $t_2$ given the (limited) information possessed and in spite of a variety of unexpected interventions or perturbations during the interval. This requires a property that is stronger than determinism and involves a counterfactual type of determination. I propose that this property is necessary for the intentional control (of the TGC type) that is used in the situation described by Kapitan. It is in virtue of being able to guide the plane to landing in face of a variety of counterfactual situations (wind changes, other approaching planes, and so on.) that the pilot (but not the steward, even if by luck he managed a safe landing) is in control (has TGC) of the plane.[43]

One aspect of TGC is crucial. TGC is not only a property of a situation (in terms of an attractor). When TGC exists one can also locate a *center* that generates it and which is responsible for the entropy reduction. This center is the collective state that governs the attractor (for example, the error-correction mechanism for the rocket, or the vortex center and its rotation for the whirlpool). For an agent to be in control of her action, it is critical that the TGC is produced by an intentional state (for example, an intention) of the agent. If, on the other hand, the TGC is *externally* produced, then the agent is not controlling her action but, rather, she is being controlled. Being externally produced means that the attractor variable is not an

[42] See Rescher.

[43] Control can be detected by a decrease in the entropy (or increase in Shannon information) of the system. The control mechanism of the rocket ensures that a large set of possible configurations at $t_1$ is mapped into a small set of configurations (or a single state) at $t_2$. Such entropy reducing (or information generation) mechanisms are therefore quite peculiar and provide a sensitive method to characterize the notion of teleological control and to distinguish it from causal determination.

item within the mental life of the agent. The obvious case is when the controlling variable is physically external to the agent (say, as when a person is caught in a whirlpool and cannot escape it). In other cases, the controlling variable may be physically internal to the agent, but external to her mental states (for example, when the agent's emotional state is uniquely determined by an ingested substance or when a brain tumor makes the agent behave inconsistently with her own beliefs and values[44]). This will be discussed in section v in relation to manipulation scenarios.

An important question is: If TGC needs to be internally produced, why is internal production not alone sufficient? Suppose that an agent intentionally raises her arm. In what sense, does one need counterfactual control (in order to ensure that she reaches the same end state under a variety of conditions) when we are dealing with a token case? Moreover, if in other circumstances, the agent fails to raise her arm, does that mean that she did not act freely or intentionally in the present case?[45] This question parallels the question of whether causal determination is sufficient to account for intentional actions, or whether a stronger teleological requirement is needed.[46]

Consider the following two situations.

($S_4$) *Intentional action of a highly constrained agent.* The agent is highly constrained in her action abilities: all she can do is press (or not press) a button, which will trigger a mechanism shooting a person at a prespecified time (the timing is programmed in advance and the agent does not control it). The agent presses the button and kills the person.

I contend that, for the shooting action to be intentional it must be the case that, if the agent was informed (and she trusted the information) that the triggering mechanism had been reversed and

[44] One may wonder if environmental factors, such as warm weather or dopamine-rich diets may involve a type of external environmental control (I thank the editors for raising this issue). I believe that it is important here to distinguish between those factors that have a *biasing* effect and those which uniquely determine their effect. For example, if warm weather makes people less enterprising or energetic, but it allows for variation in the energy profile among individuals, then it is only one of many factors, which together, contribute to the agent's character (weather has a biasing effect). If, one the other hand, the factor has a very specific effect and it eliminates individual differences (everyone who ingested the substance develops the same character or set of values), then it can be seen as a form of external control by the environment. In practice, the likelihood of this occurring (except in situations, such as the ingestion of a substance triggering uncontrollable aggression, which totally wipe out the agent's control due to lack of reason sensitivity) is vanishingly small without a teleological mechanism.

[45] I wish to thank the editors for raising this question.

[46] Sehon, "Deviant Causal Chains and the Irreducibility of Teleological Explanations."

that pressing the button will now block the bullet that would have been otherwise shot (at the prespecified time), then she would have refrained from pushing the button. Accepting this, however, indicates that even under the most limited conditions (with a repertoire of only two ways to achieve an action), how the agent would behave in a perturbed condition has a bearing on whether her present action is intentional (note that the action is the *shooting* and not the *pressing the button*). To qualify as intentional, the action needs to have been generated out in a teleological way, so that it can be achieved under some perturbed situations.[47]

($S_5$) *Deviant causal chain.* The agent intends to knock over her glass of water in order to distract her debate opponent. However, her intention upsets her and makes her feel so nervous that her hand shakes uncontrollably, striking the glass and knocking it to the floor.[48]

In this case, although the action of dropping the glass is caused by the intention to drop it, no counterfactual reliability exists, and therefore it is not an intentional action. It is precisely because nervousness diminishes TGC, that it cannot *reliably guide* the same action to take place in perturbed situations.[49] I next argue that the scheme for control and responsibility that I presented is robust enough to function even in an indeterministic world and that a succession of attractor and bifurcation stages is essential for responsibility.

## IV. BIFURCATIONS, CHOICE, REASON SENSITIVITY, AND NEURAL DYNAMICS

The property of attractor dynamics does not apply constantly over time. (If it did, people would be much more predictable, and boring, than they are!) Rather, one typically finds a process that involves successive stages of attraction (where perturbations converge) and bifurcations (where they diverge). While the attraction stages correspond to goal states (intentions for action or stable character traits) that reliably predict (and guide) behavior, at bifurcation points the agent's behavior is less predictable and shows a higher sensitivity to the input of reasons (in Fischer's sense, that new evidence would modify the choice). Such bifurcations can appear in two important

[47] An intentional action involves some knowledge condition of the goal; note that a mechanism of error correction (such as that of the guided rocket) needs to possess a *representation* of the target and a procedure of action toward it.

[48] This is adapted from Mele, *Springs of Action*; Sehon, "Deviant Causal Chains and the Irreducibility of Teleological Explanations."

[49] See further discussion, in Sehon, "Teleology and the Nature of Mental States" and "Deviant Causal Chains and the Irreducibility of Teleological Explanations."

ways in the genesis of free actions. First, they take place when the agent is faced with a choice between conflicting courses of actions, each of which is supported by some reasons the agent endorses. Second, bifurcation processes take place in the process of acquiring moral values (this is a slower process, related to character formation, where conflicting sources of information are interpreted).

One system with both attractor and bifurcation dynamics is a multi-attractor system. Figure 1 illustrates such a system, using computational studies of neural cell assemblies. The requirement of multiple attractor dynamics is important for ruling out responsibility in cases of psychological overdetermination and is consistent with Fischer's proposal that responsibility requires sensitivity to the input of reasons. (Notice, that I do not require that every decision starts from a bifurcation point, but only that an alternative attractor exists.) Research in neural networks has shown that such multi-attractor dynamics can also be *intermittent*, with time intervals of stability (convergence towards an attractor) followed by intervals of instability and bifurcations, where the convergence dynamics is switched into one of divergence, before a new attractor state is reached.[50] Moreover, neural systems can form new attractor states as a result of self-organization.[51]

It is important to note that such bifurcations, at the level of mental (psychological) states, are possible even if physical determinism holds. Consider the situation where at time $t_1$ the mental macrostate of the agent is close to the boundary between the two choice attractors in Figure 1b. Even if *physical* determinism is true (all tokens of the same type of brain state, $B_1$, at $t_1$, evolve into tokens of another type of brain state, $B_2$, at $t_2$) psychological determinism does not need to hold (different tokens of the same type of psychological macrostate, $M_1$, at $t_1$ may evolve into tokens of multiple types of mental states at $t_2$–$M_2$, $M_3$, and so on). This is due to the fact that there are alternative brain microstates at $t_1$ (for example, $p_1$, $p_2$) that are equally sufficient for the mental macrostate $M_1$ (at $t_1$) but, at bifurcations, the dynamics are sensitive to the differences between the microstate tokens. For example, these microstates can deterministically evolve from $t_1$ to $t_2$ as: $p_1 \rightarrow q_1$, $p_2 \rightarrow q_2$ with $q_1$ and $q_2$ sufficient for different psychological macrostates ($M_2$ and $M_3$, respectively).[52]

[50] O. Hendin, D. Horn, and Usher, "Chaotic Behavior in an Excitatory-Inhibitory Network," *International Journal of Neural Systems*, I, 4 (1991): 327–35.

[51] See for example, W. Dong and J.J. Hopfield, "Dynamic Properties of Neural Networks with Adapting Synapses," *Network: Computation in Neural Systems,* III (1992): 267–83.

[52] See also Nick Zangwill, "Daydreams and Anarchy: A Defense of Anomalous Mental Causation," *Philosophy and Phenomenological Research* (in press).

Additional reasons for bifurcation processes in choice or character formation are indeterminism and the presence of "noisy" input or epistemic uncertainty, such as a tune on the radio bringing to mind a memory that biases the agent one way or the other. A simple way to take such factors into account (standardly used in neurocomputational models of choice) is by adding a source of noise on top of the dynamics of the choice macrostate (for example, quantum indeterminism may lead to minute stochastic fluctuations in synaptic transmission events).

Does such open-endedness during deliberation undermine control and responsibility? The first thing to notice is that due to the attractor dynamics that are robust to perturbations, the effect of indeterminism (as well as that of epistemic noise) shows only at attractor boundaries. Therefore the objection that indeterminism diminishes control by making the agent choose or act against her "better judgment"[53] does not apply (in the absence of attractor dynamics, the noise may modify the "better judgment"). At bifurcations, however, the agent's will is divided. If the agent has competing reasons for two courses of action, no matter which of these reasons she decides to act on, the agent has authorship over the action.[54]

In agreement with Kane and Mark Balaguer (*op. cit.*), I believe that indeterminism does not undermine the fact that, when an agent performs an action and this action is probabilistically caused by her intentional states (for reasons consistent with the agent's character, motivation, and so on), the agent bears responsibility for her action. Assume for example that given an agent's character and current situation, the probability of the agent *A*-ing is 0.1 and the probability of her *B*-ing is 0.9. Consider one of the rare cases where the agent *A*-ed. I maintain that as long as there is any probability (larger than zero) for performing an action as a result of a mental state (and for reasons related to it), when the agent performs the action she is fully responsible for doing so. The agent's responsibility for the action is grounded in the fact that she has mental states that are directed towards that action and make it more likely.

Nevertheless, as Kane admits, although probabilistic causation explains the action as a causal outcome of the agent's mental states, it cannot explain why the agent chose *A rather than B*. Since such a *contrastive* explanation is unavailable, the fact that the agent chose *A rather* than *B* remains random. But as I argued above, this does not

[53] Mele, *Autonomous Agents*, p. 203.
[54] See Balaguer for further aspects of this argument.

undermine the responsibility for *A*-ing, if *A*-ing is consistent with the agent's character and motives. If, however, we inquire about the agent's UR—that is her responsibility for having the character and the motives that she has—I agree with the compatibilist critics that the lack of a contrastive explanation is likely to become problematic: at birth there is little mental repertoire to ground responsibility for indeterministic decisions, and if the developing character depends on them (as Kane requires), the way this character turns out may well be considered to be a matter of luck.[55] As we saw earlier, however, the problem of luck is not exclusive to indeterministic theories; deterministic theories are faced with the same problem, when challenged about hereditary and biological factors that affect an agent's character. The solution I prescribe (and which diverges from Kane's) is to allow that there is *constitutive luck* in both cases. Accordingly, an agent's character is an evolving system, which is affected by (and absorbs) a variety of "random" factors, indeterministic events being only one possible subclass.[56]

To summarize: indeterminism is not necessary for responsibility and control. However, it also does not undermine them, if it takes place at bifurcations. I contend that some type of TGC mechanism is necessary for any intentional responsibility-bearing action. Such mechanisms correspond to attractors in dynamical systems with variable intervals of persistence (from minutes to years)[57] and they yield to bifurcations, which lead to novel attractors. The responsibility for an action can then be traced back to the relevant attractor that has TGC over it. For example, we may trace back the responsibility for a

[55] Bernard Berofsky, "Ultimate Responsibility in a Deterministic World," *Philosophy and Phenomenological Research,* LX, 1 (2000): 135–40.

[56] The theory I propose here is neutral on the question of whether indeterminism, although not needed for responsibility, has a value in allowing agents to "make contributions to the world, that they are not determined to make" (Mele, "Soft Libertarianism and Frankfurt-style Scenarios," p. 135). Whether such indeterministic contributions (or initiations) have a value depends, however, on what we say about the following. Suppose that after a hard open-ended bout of deliberation, we discover that one of the following scenarios is true: (i) the choice was ultimately (in)determined by an irreducible probabilistic factor in the brain choice mechanism (a la Balaguer); (ii) the choice was ultimately determined by a minute external influence (say, a tune accidentally heard on the radio that brought to mind a thought that made you prefer one of the options). Do we really feel that we made a difference to the world in (i) but not in (ii)? I tend towards the view that the degree of authorship and initiation in these two situations is the same, so indeterminism does not matter: it is merely one type of constitutive luck.

[57] Transient attractors correspond not only to mental processes such as intentions for actions or stable character traits, but even to meteorological phenomena such as tornadoes, or social processes such as Communism and Nazism.

building's breakdown to the tornado that destroyed it, or to the person who demolished it; crucially we stop here, as we are not able to track the responsibility beyond bifurcation points. Consider a world where no such convergent/divergent dynamics take place. If that world was deterministic (and lacking attractor/bifurcation dynamics), then one could trace the responsibility for an effect all the way to the state of the world at the Big Bang; no other state of the world is "special" in its role at bringing that effect about. By contrast, a world with successions of bifurcations and attractors has special centers of responsibility. In such a world, even if the transition from a bifurcation to an attractor is luck dependent or random, the attractor becomes the ultimate feature that explains why its associated TGC patterns take place, in a way that nothing antecedent to it can do (that is, the attractor has UR*). Even if their origin is grounded in constitutive luck, transient attractors are entities of which it is correct to use the slogan of Harry Truman, "The buck stops here." In the following section I discuss situations where an attractor state was itself TGC-induced by another attractor state; for example, the tornado may have been engineered, or an agent's character created by indoctrination.

## V. AUTONOMY AND MANIPULATION

The challenge facing compatibilism is to account for the distinction between determination and compulsion. In the previous section, I used the concept of TGC to account for the distinction between actions that are under agent control and actions that are *merely* determined by physical laws. Actions that are under agent's control are the outcome of TGC that originates *within* the choice and action production mechanism of the agent. There are situations, however, when the source of the TGC is *external* to the agent. Consider, for example, a manipulated agent. The manipulator may use a variety of techniques (setting up the situation, conditioning, and so on) to make the agent produce a desired action. In a recent article, Patricia Greenspan (*op. cit.*) discusses a variety of manipulation scenarios that violate the agent's autonomy, with the conclusion that one of their central features is that the manipulee is used as a means to the manipulator's end. Note that this involves teleological and counterfactual determination: the manipulator will ensure the agent will produce the action in a variety of nearby possible worlds. This does involve TGC, whose source is *external* to the manipulated agent, as its origin is not part of the agent's mental states. In this situation, the agent is not in control, but rather she is being controlled by the source of TGC. Being under external control that leads to an action is being compelled to do it.

Notice, that being subject to compulsion does not require a cogent agent to do the manipulation, but rather takes place whenever possibilities are reduced (entropy reduction) by an attractor type dynamics whose source is external to the agent. Clearly, other cogent agents are the most likely sources that may exert such entropy-reducing control. However, a noncogent attractor system, such as a tornado vortex or a lighting bolt, could achieve the same.[58]

The central challenge is to distinguish between the "benign" type of influence that occurs in any aspect of normal life (for example, children's education) and the pernicious type (manipulations) that we consider to be autonomy violating and (arguably) responsibility reducing.[59] Compare normal education with the type of manipulation involved in an effective indoctrination (or value engineering) that uses a combination of methods based on positive reinforcement (as described Skinner's Walden Two), manipulation of informational input, and consensus group dynamics.[60] An agent who was subject to such indoctrination can still exert TGC to produce actions that satisfy her goals and motivations, and to this extent she is responsible for these actions; had the agent (or corresponding TGC state) been disabled, the action would not have occurred. In my view, we can agree with hard compatibilists that indoctrinated agents are responsible for their actions. However, I think we can also agree with soft compatibilists and libertarians on two important aspects: First, indoctrinated agents are not autonomous, and second, although they are responsible for their actions, they are not the *last source* of responsibility: this is the indoctrinator who shares the responsibility for his agent's actions.

To further examine the difference between the autonomy of the "normal" and of the indoctrinated agent, consider two agents who exert TGC at present, but who differ profoundly in the way their motivational and belief states were acquired. One acquired her motivational states under normal education, while the other acquired her motivational states under indoctrination. The value systems and

[58] Although highly unlikely, a nonTGC event such as the electrical noise in a plane navigation system could neutralize the pilot maneuvers and bring the plane to a safe landing at a different destination than intended. Similarly, an agent may have a uniquely determined change of character (but which maintains her ability for "normal" action control, as does the Beth-character in Mele's autonomy violating scenario, see footnote 22?) due to an ingested substance. The likelihood of such autonomy undermining non-TGC events, however, is extremely small.

[59] Kane, *The Metaphysics of Free Will*.

[60] See examples in Greenspan.

motivational states of both agents depend on some external factors. For both, the combination of values that characterizes their mental makeup is determined by the environment. A first difference, one may discern, involves the "complexity" of the *determining function*. For the "normal" agent the determining function involves a myriad of environmental and biological factors including innate (genetic) ones, while, for the indoctrinated agent, the determining function can be specified by a few factors corresponding to the intentions of the indoctrinator at a previous time.[61] Should this matter? Or should one rather say that, after all, determination is determination?

Note, however, that there is a more important difference between the two agents, which shows up when we examine counterfactuals. Unlike the normal agent, the mental configuration of the indoctrinated agent is *counterfactually determined* by the indoctrinator. By contrast, the mental configuration that the normal agent ends up with, would diverge significantly given minute variations in environmental input or in biological makeup. The character formation of this agent is the result of a (luck dependent but *constitutive*) self-organizing process, which results into a stable character. Unlike with the indoctrinated agent, the normal agent's character is not *reliably determined* by something besides or antecedent to the agent: she has UR*.

The soft compatibilist can then propose that autonomous agents are those whose TGC mechanisms were not subject to a *fully* determining TGC, whereas manipulated agents' were.[62] Thus, although both agents are responsible for their actions, only the normal agent is autonomous; the indoctrinated agent shares responsibility with the indoctrinator, who is the TGC source of her motivations.

In practice, an indicator of the presence of indoctrination (and loss of autonomy) can be obtained by examining the degree of *diversity* of value configurations of agents raised in a particular environment, and the predictability of their configurations (on the basis of the education program). It is in the nature of a good education program

[61] The determining function maps the state of the world at a previous time, $t_1$, to the mental state of the agent. With the indoctrinated agent, a time $t_1$ (when the indoctrinator has formed his plans) exists, where the determining function requires a small number of variables (the intentions of the indoctrinator) and other physical variables are screened out, while with the normal agent, the determining function at $t_1$ is sensitive to variations in a much larger set of physical variables.

[62] Note that a biasing process, such as warm weather, does not make the biased character insensitive to other factors and thus is not enough to eliminate variation and undermine autonomy. Note also that this is a simplification, since indoctrination (as well as autonomy) is not an all or nothing affair, but rather is a matter of degree.

that it allows (or even encourages) diversity and that despite the "best" efforts of the educator, the emerging character of a developing agent is bound to surprise us. On the other hand, highly uniform societies (where differences between agents are minimized) are a strong indicator of the presence of indoctrination and of the loss of autonomy.[63]

## VI. CONCLUSION

The account of free will and responsibility proposed here shares features with a number of compatibilist and libertarian theories. In particular, it makes use of doxastic freedom and of the reason-sensitivity principles for responsibility. A part of this account, however, was to develop a theory of teleological control, based on attractor/bifurcation dynamics consistent with causal mechanisms, which can function within an environment characterized by uncertainty. Within this theory, intentional control is characterized by a counterfactual determining causal process (teleological guided control, TGC) whose center is a psychological state. Bifurcation processes, where input sensitivity is maximal, taking place in action selection or in character formation, do not undermine responsibility, if one allows for constitutive luck. Crucially, the theory can distinguish between causal determination, which does not undermine autonomy and CNC manipulations, which do. Finally, the theory does not contradict the powerful intuition that processes of deliberation are open ended.

MARIUS USHER

University of London, Birkbeck

[63] A manipulated society, such as that described in Aldous Huxley's *Brave New World* can also "set up" diversity by giving each agent a different indoctrination program. Thus one needs to know more about how diversity was produced in order to conclude that no manipulation was at work.