Convergence/divergence dynamics grounds action control, agent-autonomy and responsibility[1]

Marius Usher & David Lagnado
Department of Psychology Department of Psychology
Tel-Aviv University, Israel UCL, London
marius@post.tau.ac.il d.lagnado@ucl.ac.uk

Abstract

Theories that attempt to explain the nature of agent control and responsibility are often framed in relation to the ontological character of natural law: *determinism* vs. *indeterminism*. This distinction, however, has not proved fruitful, since agent control, autonomy and responsibility are threatened both by determinism (is determination by laws of nature different from non-coercive determination by a manipulator agent?) and by indeterminism (the *problem of luck*). Here I will argue that a much more relevant distinction to control/responsibility and agent-hood is obtained from two different types of properties of natural law: convergent vs. divergent dynamics (attractors vs. bifurcations).

Introduction

There are two serious challenges to the development of a coherent account of action and moral responsibility, which address a set of very powerful and compelling psychological intuitions. The first, is the basic intuition that we make, at least sometimes, free actions for which we are morally responsible. This intuition raises a problem because it seems to be inconsistent both with the assumption of determinism and of the indeterminacy of physical laws. According to a standard argument, physical (or biological) determinism entails that all actions are determined by a long sequence of events that were in place before the agent was born, and over which she has no control. Furthermore, it is difficult to account for the difference between 'free' actions that are the result determination by laws of nature and 'non-free' actions that are the result of a non-coercive manipulation or indoctrination[2]. Indeterminism, on the other hand, presents us with an equally challenging obstacle: the *problem of luck*[3]. The alternative strategy, of denying the existence of free-will and moral responsibility is not only counter-intuitive but also inconsistent with the nature of rationality in practical and theoretical reasoning[4].

I argue here that the problem of free -will and of moral responsibility results from a second, more basic problem, that of distinguishing between actions and mere body movements, between agents and objects. To quote Scott Sehon (1997): "*there is a prima facie tension between the common-sense account of ourselves as agents and the view of ourselves as physical objects. Notions like action/goal-directed behaviour appear to have no role in physical description of the world. Planets, rocks, elementary particles do not **do** things; if we are not different than these things, then our status as agents who **do** things can be put into question*" (Sehon, 1997)[5].

I will argue that the key to providing a coherent account to both of these intuitions – that of free-will and that of agent-hood – is to focus on a different aspect of physical law, convergent vs divergent

1 We wish to thank Nick Zangwill for very helpful discussions and criticism of these ideas.
2 Kane R. (1996). The significance of Free-Will.
3 Ibid; Mele A. (2006). Free-will and Luck.
4 Searle J. (2001). Free-will as a problem in neurobiology. Philsophy, 76, 491-513.
5 *Pacific Phil. Quart.*, LXXVIII, 195-213.

dynamics, and use this to develop a theory of action via a process of robust-causation or *guidance-control*. I will start with explaining guidance control, before showing how one can use this notion to ground moral responsibility. A key aspect of this approach is to insist that a condition for moral responsibility is that the agent has robust causation over the action, which implies causation not only in the actual world, but also in a set of similar worlds. Second, I will examine the problem of luck and suggest a solution that embraces luck as a constitutive part of the agent. Finally, I will show how this scheme solves a number of challenging puzzles about freedom and responsibility.

Guidance-control and robust causation

The standard account of action, conceives of actions as being caused by adequate mental states (beliefs, desires, etc) of the agent. As first discussed by Harry Franfurt[6], this implies that actions and mere movements can only be distinguished by events that took place *before* they came into play (i.e., by their causes), rather than in themselves, for example, in the way in which they are proceeding at present. This leads to difficulties in dealing with cases of deviant causation, such as the case of a 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 and so his glass spills. Frankfurt's proposal is that the problem arises from the attempt to locate the distinctive feature of actions in their antecedents, rather than in the process by which the action unfolds, and which needs to be actively and teleologically guided by the agent.

This idea was further developed in my recent work[7], in which I proposed that teleological-guided control (TGC) is a necessary condition for moral responsibility and that it can help to account to the distinction between determination and manipulation. This means that an agent can only be morally responsible for an action on which she had guided control, in the sense that she brought about that action via a robust-causal path[8]. Bringing about an action in a robust way, means being ready to intervene with appropriate adjustments, to correct potential deviations between the actual process and the intended one. Accordingly, robust causation or guidance-control is being exercised by agents who wish to bring about an event, in the face of potential environmental perturbations. An important feature of this approach is to require that to be responsible for an action (say, a murder), the agent must have caused the event not only in the actual world, but in a set of similar worlds (say, by making adjustments to her plan, to bring about the intended goal, in face of some degree of variability in the movements of the victim).

Interestingly, the relation between responsibility attributions and multiple world contingencies, has also been proposed within the context of probing the degree of responsibility that agents have in cases with over-determination. This framework is based on a computational theory of causality and responsibility[9], which equates the degree of responsibility an agent has for an effect that was the result of over-determination by the action of a number of agents, via the minimal number of changes, *N*, required in order to make the effect counterfactually dependent on the agent (i.e., to make the agent pivotal). Recent experimental research indicates that people's attributions of moral responsibility in such situations are indeed sensitive to cross-world contingencies[10]. Here I suggest that the dependency of responsibility for an effect on the agent extends to situations without over-determination by multiple agents. In fact, actions performed by agents never have exclusivity on the determination of effects, as they always conspire with myriads of environmental influences. It is thus proposed that to qualify as

6 Frankfurt H (1978). The problem of action. *Am. Phil. Quarterly, 15*, 157-162

7 Usher M (2006). Control, choice and the convergence-divergence dynamics. J. of Phil Vol. CIII, (4), 188-214.

8 See also Kapitan T. (1996). Modal Principles in the Metaphysics of Free Will. *Phil. Perspectives*: Metaphysics, X: 419-45.

9 Pearl, 2000; Halpern & Pearl, 2005; Woodward, 2004; Chockler & Halpern, 2005; Gerstenberg & Lagnado, 2009)

10 Zultan, Gerstenberg & Lagnado (2011)

responsible for an effect, the agent must reach some threshold of robustness over the effect, which can be quantified via the magnitude of the environmental perturbations that can be tolerated without jeopardizing the effect. Another possibility is that rather than being a threshold function of robustness, the responsibility is a graded function of it. To illustrate this, a football player (soccer in the US), has guidance control over the ball, if s/he can control it under a variety of perturbations in the game. When scoring a goal, she gets more credit if she did control the ball (say, dribbling and then shooting into the goal) than if the ball deflected from her back. In the latter case, there is no robustness whatsoever, as any change of the path of the incoming ball, would reflect differently, nullifying the outcome. A similar example is the attribution of a higher degree of responsibility for a premeditated murder than for a murder out of the "heat of the moment".

From the perspective of ontology, it is important to consider what is the aspect of physical reality that grounds teleological guidance control. I propose that the ingredient of the physical reality that enables the robust causation needed for teleological guidance control, and which allows us to use an *intentional stance* in predicting the behavior of rational agents[11] is *attractor* dynamics. Attractors and bifurcations are dynamic properties, which appear in complex non-linear systems. The former correspond to situations in which trajectories *converge* towards a final state -- the attractor – which is robust to, and absorbs, perturbations. It is suggested that to account for the basic distinction between agents and objects (actions and events), the physical world has to produce, at some level of description, attractor dynamics. Such dynamics emerge as soon as the physical process includes error correction loops. For example, a guided rocket (Fig. 1, right panel), as opposed to a bullet (left panel).

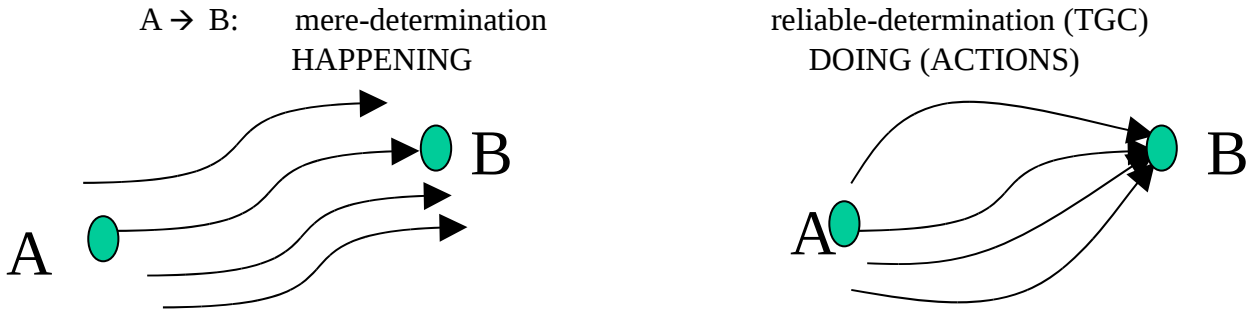

Fig. 1. Attractor dynamics can ground intentional teleological behavior, such as that performed by agents and distinguish them from particles and planets. This is because with attractors (right panel, but not with other deterministic systems, left panel), future states are determined (and can be predicted) in a reliable way, i.e., under similar but counterfactual conditions.

Decisions, indeterminism and the luck problem

While agents have guidance control over actions via intentions and goal plans, these intentions often emerge as a result of complex deliberations that are shaped by conflicting reasons and motivations. One may, for example, agonize about whether to order chocolate cake or fruit salad for desert, involving a variety of heath and emotional considerations. Even if the action of ordering fruit-salad is under the control of the agent, once her decision was taken and her intention formed, the free-will problem can be framed in relation to the control that the agent had over the decision itself. Unlike the intention-action system, which is characterized by stabilizing error-correction loops, decision processes are characterized by *divergent dynamics* that are extremely sensitive to perturbations in informational input[12]. Recent research in the neural substrate of decision making has highlighted the role of attractors and bifurcation in the decision mechanism[13]. In our current research we are testing the hypothesis that

11 Dennett D. The intentional stance.

12 Kiani R, Hanks TD & Shadlen M (2008). J. of Neurosci, 3017-29; Tsestos, Gao, McClelland & Usher (in prep.).

13 E.g., Wang, X. J. (2008). Decision making in recurrent neuronal circuits. *Neuron*, *60*(2), 215–234.

the sensitivity of the decision process to information changes dramatically after a preliminary decision[14]. To illustrate this, consider a property buyer who after spending many months and effort viewing and deciding between few properties, makes his decision and conveys it to the agent, and now in the time interval before the exchange of contract is signed, she tends to ignore information about the property that could be relevant to her. If this introspective observation is confirmed, this will establish the presence of two qualitative different stages during the production of free action. The first one, the process of decision, is highly sensitive to informational input (divergent) and ends when some decision criterion is reached, triggering a second stage of intentional planning that is of a convergent (stabilizing) nature.

Given the input sensitivity of the decision process, one may argue that the outcome of decisions is often subject to luck, and thus while agents may exercise guidance control to realize their intended actions, they have little control over what they decide to do, undermining thus their moral responsibility. Consider, for example, a person who is torn in the dilemma of helping a person in need on the street or ignoring him in order to arrive on time at a meeting. Obviously, if the decision process is input sensitive, the outcome may appear to be a matter of luck. It is important to note, however, that while the luck problem is more often discussed as a threat to indeterministic theories of free-will (Kane, Mele), it is as acute for deterministic theories. Even in a deterministic world, the outcome of decisions may change with the most minute environmental variations (an advertisement heard on the radio or a discussion we may have had in our childhood, etc). Furthermore, much of our values and beliefs are subject to factors, such as genetics or education, over which the agent has no control whatsoever[15].

It is important to distinguish between two kinds of control during decisions. When an agent deliberates between two conflicting actions, A and B, both of which are supported by reasons and motivations, and finally decided to A, the agent can be said to possess control over doing A. This is because, while the decision process is probabilistic, it is still caused by mental states that justify and motivate the decision; without the agent who possessed that mental states the action would not have taken place. If, however, the decision process is probabilistic – if repeated multiple times it will end sometimes with A and other times with B, or there are worlds similar to the actual ones where the agent As and others where she Bs – the agent can be said to possess no control over *doing A rather than B.* This is termed the *cross-world luck* problem[16].

I believe that the cross-world luck problem is unavoidable and that it threatens almost equally indeterministic and deterministic free-will theories. To illustrate the latter, consider the thief, who in court, while admitting that he stole out of selfishness and a reluctance to work for a living, argues that it was not his wish to be born selfish[17]. Obviously, very few people find this argument compelling. One way to resist it[18] (Usher, 2006; Mele, 2006; Rescher) is to embrace luck as a constitutive factor in the life of agents. The idea is that our identity is not independent of our properties and thus we cannot be lucky or unlucky to find ourselves with this or that mental property, but rather the mental properties are constitutive of identity. Having been born unselfish would make the thief a quite different person.

The same idea can be used to rescue free-will and responsibility against the cross-world luck problem.

---

14 Usher M, Tsetsos K & Lagnado D. (in preparation).

15 Greenspan P. (2003). The problem with manipulation, *Am. Phil. Quart., XI,* 155-64.

16 Mele A. (2006).

17 Smilanky S. (2000). Free-Will and Illusion.

18 Usher (2006); Mele (2006); Rescher N. (1995). Luck: the brilliant randomness og everyday life.

The actions of agents are partly subject to constitutive luck, such as those determined by genes and educations, and also by neuronal-luck that may cause the agent to choose *A rather than B* in a specific situation. The key idea is that agents are temporally extended entities (Bratman), which via their decisions, absorb values via the process of identification with their choice, which become a constitutive part of their personality (see also Mele, 2006).

The presence of successive stages of convergence and divergent dynamic is, I suggest, the critical property of physical reality that enables us to divide the world into agents and effects, with agents conceived as sources of responsibility (or robust causes) for events. The idea is that an agent has responsibility for an action, when the action can be teleologically tracked back to the agent, but no further. This scheme extends to non-biological agents. For example, the responsibility for a house destruction may be traced back to the tornado that destroyed it, or to the terrorist who has planned its destruction. In both cases, one reaches a bifurcation (the tornado formation that cannot be further backtracked): no antecedent before this bifurcation can *robustly* explain the action, justifying Trueman's famous phrase: “The buck stops here”. This picture differs from the situation that would unfold in a deterministic world in the absence of convergence/divergence dynamic stages, where responsibility tracks back (if at all) to the Big-Bang.

Manipulation scenarios

Manipulation scenarios are among the most challenging arguments that have been raised against the compatibilism of free-will and determinism. I will briefly review two such arguments, before showing how the convergence/divergence theory can resolve them. The first argument, formulated by Robert Kane[19], asks us to imagine a community (say, Skinner's Walden-II), in which people are raised under a non-coercive indoctrination program, based on positive reinforcement, that makes (without failure) the people adopt the value system of the indoctrinator. Although when grown up these agents may face decision dilemmas and make "free" actions, there is a sense in which their decisions and actions are not autonomous and the ultimate responsibility for their actions resides with the indoctrinator. Kane then calls the compatibilist to explain in what way we, who are governed by deterministic laws of nature, are different from the agents in the Walden-two community.

The second anti-compatibilistic argument was formulated by Alfred Mele[20]. He asks us to imagine that Diana (a Goddess), plants a zygote embryo into a woman, Mary, which is designed such that, given Diana’s complete knowledge of the laws of nature and the state of the deterministic world, the embryo will develop into a person, Ernie, who in thirty years from now will do E. Mele suggests that the compatibilists will have difficulties to answer the powerful intuition that, like in the Walden-II example, Ernie is only a marionette, and the ultimate responsibility for his E-ing and even, perhaps, for all the actions he will ever perform, rests with Diana.

The divergence/convergence framework solves these problems, first by making a clear distinction between determination and manipulation. While the former only requires that an event-E is determined by the laws of nature and the past state of the present-world, manipulation requires robust causation, or guidance control over the event. In particular, when an agent is subject to teleological guidance control that is *externally* generated, like the members of the Walden-II community, she is *controlled*, and then the ultimate responsibility for their action resides with the earlier source of TGC, the manipulator himself. The important distinction is that the indoctrinator possesses robustness over the values and actions of the manipulated community members; he can make these members develop such values not only in the actual world, but in similar ones, by being able to adjust to a variety of environmental

19 Significance of Free Will.

20 Free-Will and Luck.

changes. Consider now, the situation of Ernie and Diana, and let us focus on the following issue. Does Diana have robust causation and TGC over the event-E, and over all the other actions that Ernie will ever produce? If the answer in no, as should be the case, if she just considered the presentworld when creating the zygote that will develop to E, then the anti-compatibilistic intuition can be resisted[21], because without a prior robust causation of the mental states of Ernie, the responsibility for his actions cannot be backtracked further. There is one sense, however, in which we could think of Diana as holding stronger powers over Ernie's future. This is if she, seeing a number of possible future worlds, has created the zygote, so as to produce-E, in each of them (we can for now leave it open whether this is metaphysically possible). Assuming it is, she is in the situation of the Walden-II indoctrinator, of having robust-causation over the event-E, and the responsibility for its production can be further backtracked to her. Interestingly, however, she is unlikely to have robust causation over any other event in Ernie's life, since even a Goddess cannot robustly cause (i.e., in multiple worlds) more than a single event; those events could appear via different and intersecting trajectories, in their respective worlds. If this were Ernie's predicament, he would be in a situation that was nicely illustrated in Kurt Vonnegut's Sirens of Titan.[22]

Discussion

The scheme suggested here for free-action and moral responsibility is agnostic about the issue of whether physical law is subject to determinism or indeterminism[23]. Instead it posits two key elements of the physical world that allow us to view it in terms of agents who hold responsibility for certain events. The first element is the existence of attractor dynamics that ground robust causation and responsibility for some events. The second ingredient is the existence of successive stages of attractor and bifurcation dynamics that create boundaries on how far one can backtrack responsibility for events. Such attractors and bifurcations can map, not only onto the specific decisions/intentions that a person carries out, but also onto more stable personality/ideological traits (e.g., humanism/racism) and bifurcations in the process in which such personality traits are formed. Using this framework, we suggest that the responsibility for an action can be tracked back to the earlier attractor that was instrumental in the action-TGC. Furthermore, external TGC on an agent is what characterizes situations that correspond to lack of control (for the agent on which the control is applied) and reduced responsibility, such as manipulation or indoctrination.

This account allows one to distinguish between cases of indoctrination and of normal determination (such as education), where an agent's motivational states are formed. For the normal agent (subject to a myriad of forces and influences within the society) her values are a result of a self-organizing process, which cannot be tracked further back (no previous robust causation). Thus, although the personality of a normal agent depends in a complex way on the society in which she evolved, there is no previous *robust* determination possible, and thus she is *autonomous*. With the indoctrinated agent, on the other hand, a previous robust determination exists that points to the indoctrinator. Thus, while the normal agent's constitution can be said to be subject to *constitutive luck* (she would be different if one of the myriad of factors changed), no such luck existed for the indoctrinated one. Indeed, an important signature of good education vs. indoctrination is the presence of diversity vs. uniformity in the values and thought styles that they engender.

---

21 See also Fischer JM (2011). the Zygote argument revisited. *Analysis Advance access, 1-6*.

22 In the Sirens of Titan, the human civilization is guided in its development by a powerful spaceship of a robot civilization that being shipwrecked on Saturn's moon Titan, has calculated it will be faster to guide human development on earth and make it develop science and manufacture a needed exchange part for the spaceship, than to ship this from its home civilization: http://en.wikipedia.org/wiki/The_Sirens_of_Titan

23 See also Mele (2006).