Stochastic libertarianism:

How to maintain integrity in action without determinism

Thomas Müller* and Hans Briegel†

Preprint version, 12 January 2014‡

Contents

Abstract

Theories of free agency based on indeterminism—that is, libertarian theories—are often accused of undermining an agent's integrity: If an action is due to indeterministic happenings, how can it be called the agent's action to begin with? Isn't a deterministic connection between an agent's circumstances and her action needed to maintain her integrity?

*Fachbereich Philosophie, Universität Konstanz, Fach 17, 78457 Konstanz, Germany; email: `Thomas.Mueller@uni-konstanz.de`.

†Institut für Theoretische Physik, Universität Innsbruck, Technikerstraße 25, A-6020 Innsbruck, Austria, and Institut für Quantenoptik und Quanteninformation der Österreichischen Akademie der Wissenschaften, Innsbruck, Austria; email: `hans.briegel@uibk.ac.at`.

‡This paper is fully collaborative; authors are listed antialphabetically.

We claim that a meaningful notion of agency does not need determinism. In this paper we introduce stochastic libertarianism, a novel theory of free agency under indeterminism. Based on a physically motivated, stochastic model of the temporal evolution of a deliberation process, stochastic libertarianism views indeterminism as a core resource for meaningful agency rather than as a threat. We counter the supposed threat by explicitly discussing Van Inwagen's replay argument, exposing a flaw in the argument that is due to insufficient attention to temporal details. Our approach can also explain how a stochastically libertarian agent developing over time can exhibit highly realiable behavior. We claim, therefore, that integrity in action does not need determinism.

1 Introduction

Allow us to introduce Bob. He's a normal guy, he likes ice cream like most of us. It's a sunny day, and Bob passes an ice-cream parlor with a couple of friends. His favorite flavors are vanilla and chocolate; he likes them both the same, and usually doesn't know which one to choose. Today he ends up taking vanilla, but he might as well have taken chocolate.

One of Bob's friends, Alice, has been secretly keeping track of Bob's ice cream choices and wonders just on which grounds he makes them, as she cannot see any pattern whatsoever in them. She has heard about the free will debate and all that and wonders whether her secret record may in fact prove that Bob really doesn't have any control over what he is doing. And since Bob does not seem to be aware of this situation, Alice even wonders whether he can really be called the author of his ice cream choices. At some point she reveals her thoughts to Bob who, both surprised and amused by her worries, freely admits that he doesn't really know himself why he chooses vanilla over chocolate at one instance and the other way round at another. But he insists and assures her that it is in fact him who makes the decisions. He happily agrees that one could consider his choices purely random. In fact, sometimes he even tosses a coin in front of his friends. But he feels that this does not affect his integrity as an agent. After all, who cares? He enjoys what he gets in either case, so why should he rack his brains over such an irrelevant decision?

We would like to defend Bob. We think that his way of dealing with the situation, under circumstances as described, indeed make sense and does not compromise his integrity in any way. When the outcome of a decision does not matter, randomness is a simple and useful tie-breaking resource. This is so regardless of whether the randomness can be traced back to some microscopic happening in Bob's brain, or whether it is enforced by tossing a coin (or sending a photon through a beam splitter).

There are cases in which some outcome has to be reached, no matter which one, just in order to avoid deadlock—think of Buridan's ass. We take it that most philosophers would agree that some mechanism for tie-breaking is needed in such cases. And if the outcome doesn't matter, the way in which the outcome is reached doesn't matter either. If randomness is easily available, it makes sense to employ it to break the tie. But does it make sense to employ randomness in situations in which the outcome *does* matter? Assume that Bob suspects that, but isn't sure whether, he doesn't tolerate vanilla ice cream too well. In such a situation, it matters what he does, but there could still be reasons for both options—he might want to make sure once and for all, and therefore expose himself to vanilla to test his reaction, or he might want to stay on the safe side, and therefore go for chocolate. Can randomness still be a legitimate resource in such a situation, or in situations in which even more is at stake?

We take it that faced with this question, most philosophers would deny randomness a useful role. The general view seems to be that an agent's integrity is threatened if randomness plays a substantial role in her relevant decisions. After all, in which sense are these decisions hers, when they are due to chance? Many well-known arguments against a libertarian (indeterminism-based) position in the free will debate can be traced back to this worry.

Against this general view, we hold that genuine randomness is not only a tolerable way of breaking ties, but can in fact play a much more fundamental role as a resource for agency. In this paper, we develop a corresponding philosophical position, which we call *stochastic libertarianism.* According to stochastic libertarianism, a rational, learning agent reaches a decision through an underlying stochastic process in her memory that is extended in time. In that process, randomness plays a constructive role: options and considerations unfold stochastically and culminate in action. Stochastic libertarianism thus embraces a positive approach to randomness as an empirical fact of our world. It is there, and it is useful: We argue that randomness is not just something that agents can employ in highly specific situations or otherwise tolerate to a limited extent, but rather constitutes a generic force that drives their agency.

Our paper is structured as follows. In §2, we systematize our observations about randomness and tie-breaking that we made in the above ice-cream story, delineating those cases in which the outcome of a decision doesn't matter. In §3 we consider the more typical case in which the outcome does matter, and comment on the proper role of randomness in such cases via a discussion of Van Inwagen's replay argument. In §4 we explain a concrete model of stochastic libertarianism, called projective simulation, that has recently been developed in the context of artificial intelligence. After that, in §5, we return to the discussion of the replay argument and reconsider whether it is really fair to attribute the outcome of a stochastic process to an agent, as an action that is properly hers. We argue that given specific aspects of the process, attribution is indeed warranted. We wrap up in §6.

2 Some things don't matter

As we said above, some decisions are just a matter of tie-breaking, and their outcome doesn't matter. An abstract formulation of our ice cream story as a case of mere tie-breaking would be the following: Our agent Bob (i) is facing a limited set of options and (ii) has no preference at all one way or the other; furthermore, which option is picked, has (iii) no consequences for his future and (iv) no meaningful connections with his past, but (v) one of the options has to be picked. In such a case, when there is nothing at all that can tip the balance, *any* way of tipping it must be appropriate. This appears to us to be a conceptual point: If you argue against a specific way of tie-breaking (e.g., against the employment of randomness), you thereby indicate that what you were dealing with wasn't really a case of mere tie-breaking after all.

Let us belabor this point a bit, because it is of crucial importance. All of the characteristics (i)–(v) of mere tie-breaking listed above are relevant. In tie-breaking, there must be a limited set of options that are tied, not a completely open setting with no clearly delineated options. Only in such a setting do the other characteristics make sense: There must be a tie with respect to the agent's preferences among that limited set of options, and each of the options must, as it were, be isolated from the agent's past and future. In a completely open setting, in which the agent just has to "do something", the notion of a tie doesn't make sense. And if one of the available options has a meaningful connection with the agent's past or future, even though current preferences even out, that would show that the options aren't tied after all. Furthermore, it must be that one of the options simply *has* to be chosen; remaining inactive must not be an extra option.[1]

These characteristics are quite demanding. And in the end it may well be that genuine situations of tie-breaking are relatively rare. Maybe almost all of our choices can acquire some meaning because they are made in a social context and before the background of limited resources. Hardly anything we do is really isolated from our past and our future. Our ice-cream picking example, however, seems to get quite close to the ideal: Sometimes it really doesn't matter whether it's chocolate or vanilla.[2] What is important here is the conceptual point: *Given* that a situation is one of mere tie-breaking as described, *any* way out is appropriate. Thus, if a choice situation involves mere tie-breaking, then randomness as a tie-breaker is okay. If a decision is completely isolated and no option is preferred, then this just means that any way of making the decision must be as good as any other. If the agent ends up with one of the options due to some random happening, this is just as well. An argument against the employment of randomness in such a case must amount to changing the example. If the decision context is such that it does not matter either way, then per definition, there can be no argument against randomness as a tie-breaker, whether internal or external to the agent.

Still, one might argue that an agent whose decisions are driven by random processes, will act erratically and, at least sometimes, do something wrong or nonsensical. This worry is important, because it points to another conceptual issue: We have to distinguish functioning under randomness from malfunctioning due to erratic processes. In our example described in the first section, Bob's choice was between vanilla and chocolate. The reason why randomness is appropriate in that case, is that the outcome of the choice does not matter, and the choice constitutes a case of mere tie-breaking. But randomness as a tie breaker has to be distinguished sharply from erratic behavior.

[1] This may be due to the dynamics of the situation, in which inactivity may be ruled out, or it may be that inactivity is also one of the options, but is tied with respect to the others.

[2] Ironically, it seems that among the best candidates for mere tie-breaking situations we find the experimental settings of neuroscientific free-will experiments such as Soon et al. (2008): Whether you click left or right in such an experiment, it really does not matter at all.

There are two types of erraticness. If we see that the Bob keeps making random choices even though chocolate, but not vanilla, consistently gives him a stomach ache, we might call his behavior erratic or irrational, and question his integrity as an agent. We assume that people learn from past mistakes. On our definition of mere tie-breaking, given the meaningful connection of the choice with Bob's past, this simply isn't a case of tie-breaking, and if Bob behaves as if it were, he fails to make the connection.

The other type of erraticness can occur in a context of true mere tie-breaking, viz., when an agent's behavior falls outside the tied options, perhaps randomly, but through malfunctioning. This would be the case, for example, if a human agent has a seizure or struggles with a tremor (neurological motor dysfunction) that, e.g., reduces her control of pointing at the flavor she decided for (so she may end up with strawberry, even though the choice was between chocolate and vanilla). It may well be that there is an element of randomness in the occurrence of such erratic behavior. But the important point is that the employment of randomness for tie-breaking (or, to anticipate the discussion of the following sections, for decision making quite generally) in a properly functioning system does not have to lead to such erraticness at all. The role of randomness in a properly functioning system can be confined such that only one of the limited number of sensible options is chosen. Randomness doesn't mean that anything goes.

3 The replay argument

We have argued that random choices are okay when the outcome doesn't matter. But what if it does? A choice matters if the agent has a preference, or if there is a relation between the options at hand and the agent's past experiences and future expectations. The ice cream example was, by assumption, disconnected from Bob's past and future. For choices for which this assumption doesn't hold, however, we need a historical perspective on an agent's development. We learn from experience. Thus, judgments about whether some behavior is meaningful or erratic, have to be made before the background of past experience, i.e., past choices and their consequences. Can randomness play a positive role in explaining meaningful choices in such a historical perspective?

Leaving cases of mere tie-breaking by the wayside, the important question therefore is whether randomness can play a constructive role in a meaningful decision process. As mentioned, there are strong arguments suggesting a negative answer. For concreteness, let us consider the example famously employed by Van Inwagen (2000) in his replay argument, which is supposed to show that indeterminism in a decision process threatens an agent's freedom, and specifically her integrity as the author of the decision. In Van Inwagen's example, an agent, Alice, is facing a difficult decision whether to lie or to tell the truth. (We may assume that Alice has been asked by Carl whether she has seen Bob of late. She knows that Carl will probably start his usual sermon about Bob's frivolous character if she admits to seeing him, and she'd rather avoid that—but is that worth telling a lie?) In fact she wrestles with the issue for a while and then decides, at time t_{tr}, to tell the truth. According to Van Inwagen, if you hypothetically roll back the situation to a time t_0 just before t_{tr}, when (as the libertarian has to assume) both lying and telling the truth were real possibilities, the individual rerun can turn out one way or the other, and from among 100 runs, there will be n with Alice lying, and $100 - n$ with her telling the truth.[3] What appears to be a problem is this: Given the statistics of the hypothetical reruns, Alice's individual decision appears to be a matter of chance, with a ground-floor probability for lying of (approximately) n per cent. So how can Alice be made responsible, or praised, for her actual decision to tell the truth, which is just a random happening? The fact that her decision involves an element of chance, seems to threaten her integrity as an agent.

We agree that the argument has quite some intuitive force. But we hold that a proper focus on the detailed dynamics of a decision process shows that the argument is faulty.

Let us look at the course of decision making in the first few of the 100 runs under consideration. By hypothesis, all of the runs start at t_0, a little before t_{tr}, with Alice and the environment in exactly the same state. Also by hypothesis, in the actual course of events (run 1), Alice decides to tell the truth at t_{tr}. It follows that truth-telling is a possibility for t_{tr} in any run. We can assume that run 2 is similar in that respect, and is also one in which Alice tells the truth at t_{tr}. But run 3, let us assume, is different—what does it look like? The situation at time t_0 is by assumption one in which Alice is

[3] We accept for the sake of the argument that any run in which Alice does not lie is one in which she tells the truth. Actually this is debatable, since there may be other options, such as walking away.—We are making a few small changes to the exact layout of Van Inwagen's argument in order to simplify exposition. None of this affects the point we are making.

wrestling with the available options, based on a variety of aspects of the situation at hand. These aspects can be remembered associations with similar situations in her past, normative and prudential considerations, conceived consequences, or emotions that come up in connection with the deliberation process. Any such association takes a little time—and this matters for the temporal fine structure of the runs. By assumption, in all runs, considerations favoring truth-telling were salient just before t_{tr} (exactly as in run 1). But in run 3, let us assume, another consideration (perhaps an idea about long-term consequences) comes up, and Alice is busy entertaining that consideration. Thus, at time t_{tr} in run 3, Alice is neither telling the truth nor lying, but still considering. A little time later, let us assume, Alice again tells the truth, but now based on a different consideration than the one that was salient just before t_{tr}. And run 4, we may assume, is similar to run 3 up to entertaining the idea about long-term consequences, but then differs in that yet another consideration (perhaps one about short-term consequences) comes to mind, and Alice a little later tells a lie. The various runs are differentiated not just by the outcome of Alice's decision process, but also by the course of considerations that were entertained during that process, and by the time at which Alice ends up telling the truth or lying.

This level of detail is missing in Van Inwagen's version of the replay story. But that detail is crucial. Consider a difficult choice you recently made yourself: If you are like one of us, this is a process of considerations going back and forth, some favoring one outcome and some favoring another, and others perhaps neutral—and what happens in the course of such a deliberation process is not limited to pertinent considerations coming to mind either. Lots of developments of a deliberation process are possible, and the time it takes can vary wildly. But does that make the choice a chance happening?

There are two perspectives one can take on Alice's decision. Van Inwagen gives a coarse-grained description in which runs are classified only as leading to Alice's telling the truth or lying. The suggestion is that these truth-tellings or lyings are possibilities for the exact same time—in the same way in which a toss of a coin makes for two alternative possibilities for the same time.[4] We agree that *if* the alternative to Alice's telling the truth at t_{tr} is

[4] If you don't like the example of coin tossing, let a photon pass a beam splitter.—Van Inwagen does not take an explicit stance on *when* Alice tells the truth or lies in the different runs. The way he sets up the argument, however, strongly suggests that he is assuming alternatives for the same time. See Van Inwagen (2000, 15f.), esp. p. 16: "[...] in each replay, Alice will *either* agent-cause cerebral events that, a second or so later, will result

lying *at* t_{tr}, something erratic must be going on. The choice matters, and becomes understandable by the considerations on which it follows—but just before t_{tr}, Alice's considerations favored truth-telling, so how could she lie based on these considerations? The real alternative *for time* t_{tr} is to continue deliberating, not making a different choice.[5] It is true that the series of runs still allows for Van Inwagen's coarse-grained description; we can count (given the hypothetical replay scenario) in which runs Alice tells the truth and in which she lies. (Note that there may also be some in which she just walks away.) But that is not all there is to the decision making process.

Attention to the temporal fine structure of the individual runs thus affords an improved view on the alternate possibilities that a libertarian has to assume. It is true that a libertarian position must assume that alternate possibilities are given in an indeterministic way—that is so by definition. But the libertarian does not have to accept an opponent's description of these possibilities without further argument. We claim that on the fine-grained perspective that we sketched, Alice's integrity as an agent, and her responsibility for her decision, is not affected if the process leading to her action is one in which the sequence of considerations, and the question whether a consideration leads to action, is fundamentally indeterministic. In fact, we claim that an indeterministic organization of the decision process, coupled with a proper view of Alice as a learning and developing agent, endows Alice with more flexibility in her decision making than a purely deterministic process ever could, and thus enhances her agential capabilities.

In a temporally extended decision process, randomness can play a constitutive role in making alternative considerations (possible reasons) salient. Seen in the larger context of the deliberation process, a random occurrence of a consideration is not an element of erraticness, but a constitutive part of the dynamics of association. Each of the runs in the hypothetical replay scenario has a rational structure that is explainable through the considerations that were salient before the action. Whatever Alice does, if it is the outcome of a temporally extended process of considering, it makes sense, and it can be properly attributed to her.

This verdict is strengthened if we take into account a long-term historical perspective as well. Given the state Alice is in just before t, we agreed

in bodily movements that constitute her telling the truth or agent-cause cerebral events that, a second or so later, will result in bodily movements that constitute her lying."

[5] This point is suggested, e.g., in Broad (1933, 240), Keil (2007, 115) and Steward (2012, 155ff.).

that the process by which one consideration is followed by another, or by an action, is fundamentally indeterministic. But it seems reasonable to assume that the probabilities for which consideration can follow which, are influenced by Alice's past experience. Alice is, by assumption, a normal adult human being, so she grew up from a child to become a responsible member of society. In that process, she made many choices, and these had many external consequences from which she learned. Even a simple reinforcement learning scheme can explain how based on feedback, Alice's association/action probabilities are influenced by her history, and thus constitute part of who she is.

In order to ground a notion of libertarian agency, therefore, both the short-term and the longer-term temporal history of a decision process needs to be taken into account.

4 A concrete model for stochastic libertarianism

So far, what we have offered to support our view of indeterministically driven decision making, are some broad-brush descriptions of an associative process. For those who remain skeptical of the tenability of our story, we can offer a little more detail. There is a formally well described and physically well motivated model for agency that exhibits just the elements of associative memory organization and random option selection that we have invoked above. The model, called *projective simulation*, features a memory structure in the form of a dynamic network of clips, which are units of episodic memory.[6] Stochastic transitions in such a memory are the basis of an explicit model of deliberation in a learning context.

The basic idea is as follows. Triggered by some perceptual input s, a specific memory clip, c, is activated (with probability $\mathcal{I}(c|s)$). Subsequently a random walk $c \to c^1 \to c^2 \to \cdots \to c^m$ through the clip network ensues, involving m transitions (where the number m itself is not determined beforehand), until activation is coupled out, triggering some motor action a (with probability $\mathcal{O}(a|c^m)$). The random walk through the clip network

[6] The model of projective simulation was introduced in Briegel and De las Cuevas (2012). A detailed discussion of this model from the perspective of philosophy of action can be found in Briegel and Müller (2013).

follows certain weights (probabilities), e.g., $p^{(t)}(c'|c)$ for a single clip-to-clip transition $c \to c'$ at time t. These weights are built up through the agent's learning history and thus encode her past experience connecting sensory input to action output, including these actions' consequences. In one run of projective simulation, a given input s (think: Alice's situation at t_0) can lead to any of a number of outputs a, a', a'', …, and output can be triggered at different times $t_0 + m\Delta t$ corresponding to a deliberation length of m transitions (each taking the time Δt). So we capture in that model the detailed structure of the space of possible reruns described above. In particular, the random exploration of different dynamic paths in the clip network mirrors the unfolding of reasons or considerations that precedes a specific action in our discussion of the replay argument.

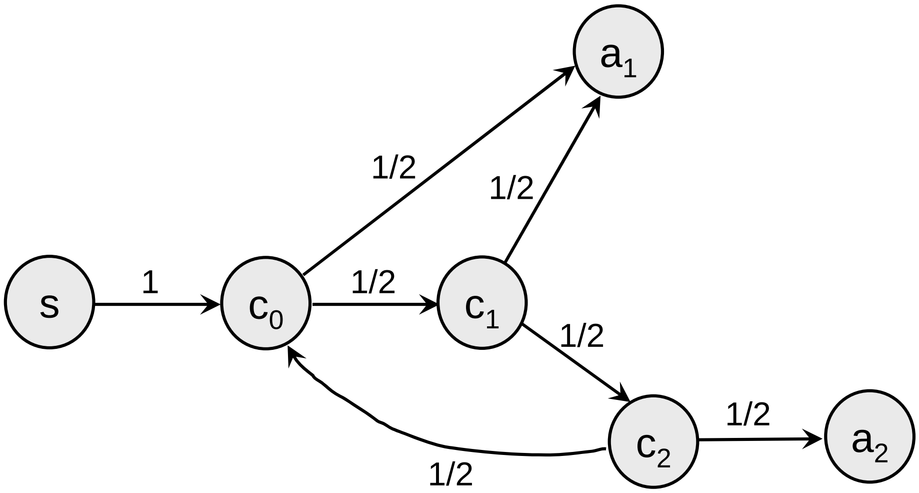

Figure 1: Clip network corresponding to the replay discussion. Arrows are labeled by the respective transition probabilities. Note that the probabilities of all transitions leaving any given clip add to unity.

To illustrate the mentioned aspects of the model, consider the following simple network (Fig. 1), which is meant to mimic the essential elements of our discussion of the replay argument.[7] Our agent, Alice, is confronted with a situation s at time t_0; this is the initial set-up of all the replays. In the model,

[7] The point of the following illustration is that we can show what the stochastic dynamics looks like in a simple case. We do not mean our network to give a realistic picture of an actual decision situation, which will of course be much more complex.

this situation corresponds to a perceptual input node s, which triggers a first clip, c_0. We can interpret this clip c_0 as an episodic consideration favoring truth-telling, perhaps Alice's thought that it's generally bad to lie, coupled with the emotional setting of situations in which she was told so. In the model, this clip can either trigger action a_1—truth-telling—or another clip, c_1. To match the story given in §3, we can interpret this clip c_1 as Alice's consideration of the long-term effects of lying in the specific case at hand (involving, for example, the fear of being detected). In the model, that clip can again trigger action a_1—truth-telling—, or lead to clip c_2. That clip we interpret as Alice's consideration of short-term effects of lying in the specific case at hand (involving, for example, the thought that this would get her out of the uncomfortable situation she is in). In the model, clip c_2 can either trigger action a_2—lying—or lead back to the first clip, c_0, i.e., Alice's general consideration about lying. While certainly not a realistic picture of what is really going on in a difficult moral choice such as the one Alice has to make, the model does represent the salient features of the description that we gave in the previous section.

Now, *calculemus.* In order to simplify the math, let us assume that all transitions have equal weight, such that, for example, both possible transitions from clip c_1—truth-telling or arriving at the next consideration c_2—have equal probability of $\frac{1}{2}$. An agent who has learned from experience, will of course have adjusted the transition weights; if Alice is rigoristic about never lying, her weight for the $c_0 \to a_1$ transition will be much higher than the weight for the transition $c_0 \to c_1$.

The shortest path through the network leading from percept s to some action is the following:

$$s \to c_0 \to a_1.$$

The probability of that path, given s, is $1 \cdot \frac{1}{2} = \frac{1}{2}$. In Table 1, we list a few more paths with their respective probabilities. (Note that due to the feedback structure of the model—the possible transition from c_2 to c_0—there is no limit to the length of possible paths, but of course the probability of paths becomes smaller the longer they are.)

Given these simple stochastic dynamics, the total probability that it will finally lead to truth-telling at some time after t_0, under the assumed weights, is the sum of the probabilities of all paths leading to truth-telling, which

Path	lenght	probability	truth/lie?
$s \to c_0 \to a_1$	2	0.5	truth
$s \to c_0 \to c_1 \to a_1$	3	0.25	truth
$s \to c_0 \to c_1 \to c_2 \to a_2$	4	0.125	lie
$s \to c_0 \to c_1 \to c_2 \to c_0 \to a_1$	5	0.0625	truth
$s \to c_0 \to c_1 \to c_2 \to c_0 \to c_1 \to a_1$	6	0.03125	truth
$s \to c_0 \to c_1 \to c_2 \to c_0 \to c_1 \to c_2 \to a_2$	7	0.015625	lie
...	...	...	...

Table 1: Possible paths in the network of Fig. 1.

comes down to $p_{\text{truth}} = 6/7$. The probability for lying is, accordingly, $p_{\text{lie}} = 1 - p_{\text{truth}} = 1/7$. Note that by adjusting the weights, we can arrange for any ratio of truth-telling vs. lying.[8]

In the given model, truth-telling can occur immediately, after two steps, at time $t_0 + 2\Delta t$, but also after a much longer deliberation time, e.g., after 12 steps, corresponding to time $t_0 + 12\Delta t$. Similarly, lying can occur after four, after seven, or after more steps.[9]

For simplicity's sake, we have ignored the aspect of learning in our discussion of the above toy model. In the fully specified framework of projective simulation (Briegel and De las Cuevas, 2012), learning proceeds in a reinforcement set-up, by adjusting the clip-to-clip transition weights. The projective simulation process is indeterministic, yet each single run makes sense before the background of the agent's learning history reflected in the transition weights (or, more generally, in its memory structure) at the given time. And each run adds to that learning history through the feedback (in the simplest reinforcement learning scheme, reward or punishment) that the

[8] Consider, for example, a network where the probabilities in all binary branchings are adjusted in such a way that the direct transitions to a_1 (i.e. those from clips c_0 and c_1) occur with probability p and the direct transition to a_2 (i.e. the one from clip c_2) with probability $1 - p$ (with $0 \leq p \leq 1$). The effective probability for truth-telling is then given by the expression $p_{\text{truth}} = \frac{p(2-p)}{1-p(1-p)^2}$. An effective 50:50 distribution among the two options is obtained for the configuration with $p \simeq 0.245$, which is the real-valued solution of the cubic equation $p^3 - 4p^2 + 5p - 1 = 0$.

[9] The fact that in our model, for any given number of deliberation steps, the action is fixed, is due to the simple structure of the model. If you add an extra transition, e.g., from c_2 to a_1, both lying and truth-telling are possible after four steps. To repeat, the model is only meant to illustrate our discussion, not to provide a realistic picture of an actual decision process in its full complexity.

resulting action receives in turn. Furthermore, as part of the process of projective simulation, new clips may be created out of already existing ones, which need not correspond to any factual experience in the agent's past. If the activation of such "fictitious" clips, as part of a random walk, leads to rewarded actions, their embedding into the clip network will be strengthened and they may become an integral part of the episodic clip network. This process of random clip creation, together with the propagation dynamics, will then lead to new options for action, as well as to new paths of considerations in the agent's memory.

Note that based on such an indeterministic decision process with feedback leading to learning, deterministic reactions to specific stimuli can be learned. Indeterministic agents need not be unreliable or haphazard. In some cases they can exhibit behavior typical of hard-coded routines (even starting from the weights of Fig. 1, given proper reinforcement, Alice can learn to become a rigorist about not lying), but in other cases, such agents will react completely randomly (Bob can stick to complete randomness in his ice-cream choice if it makes no difference). And this flexibility makes good sense.

Within this model, therefore, we use indeterminism as a central resource. This does not mean to say that we abandon an otherwise deteministic agent to random processes. In our view, there is no reasonable deterministic notion of our deliberating agent to begin with, which could then be "randomized". The random processes we are referring to here are not something external that would randomize the agent's actions (as if the agent was given independently and beforehand). On the contrary, the random processes form a constitutive element of the agent's memory and the very process of decision finding.

It should be pointed out that the model of projective simulation, including its rules for transitions and compositions in clip space, represents a specific model of reinforcement learning in a physically inspired approach to (quantum) artificial intelligence. It is meant to be a simple model for artificial agents that can learn and show intelligent behavior, without the claim to give account of any deeper or more advanced aspects of human agency. Nevertheless, for our purposes it serves as a formal model of agency, where the process of decision finding that precedes action in a given (learning) environment can be mapped out in detail. For further treatment, we refer the interested reader to Briegel and De las Cuevas (2012) and to Briegel and Müller (2013).

5 But is it fair?

Returning to our agent, Alice, who is struggling to make up her mind, we can be assured that as long as the stochastic process leading to her decision functions properly and is not disturbed, it will follow a detailed course of associative deliberation steps that can be properly and meaningfully attributed to her. The fact that a process that constitutes part of her identity as an agent, makes use of indeterministic randomness in its dynamics, does not threaten her integrity as an agent.

But is it fair to hold Alice accountable for the concrete outcome of a random process going on in her memory? Well, yes, because it is *her* memory, which belongs to, or even defines, her identity. As long as we can identify the randomness of that process as a constitutive element of her ability to deliberate and to learn, rather than an erratic event that jeopardizes her learning process, the outcome is properly Alice's.

Some people will still not be convinced. Suppose Alice tells the truth and we want to praise her for making that tough choice in the given circumstances. If it is random, how can it be her choice? Again, details matter. Let us suppose that the 100 reruns under consideration, each of which gives a detailed real possibility for Alice starting at just before time t_{tr}, are in fact divided as 50 cases of lying and 50 cases of truth-telling.[10] Let us furthermore assume that this 50:50 division reflects an objective indeterministic probability in the given situation.[11] Then the statistics of what Alice is doing are like the statistics of throwing a fair coin. But unlike the case of vanilla vs. chocolate considered above, throwing a coin in Alice's situation would be an inadequate means of arriving at a decision, and would not do justice to the situation. What the indeterministic decision process does, however, is not just to provide an outcome satisfying certain statistics, but in each and every run to provide a path of considerations that is shaped by Alice's past experience. A coin toss would not do that, and is therefore inadequate in a case in which it matters what Alice does. In her case, decision making is not mere tie-breaking. It is true that a decision one way or the other has to be reached (by assumption), but the considerations employed in that process matter a lot. They do not just matter for the justification of Alice's decision and our moral evaluation of her choice in the single case at hand—they

[10] As mentioned above, this is probably inadequate, since other options may become salient in the deliberation process.

[11] See note 8 above for the respective weights for the clip network of Fig. 1.

also provide a foothold for learning through feedback. And Alice doesn't just learn based on the outcome, but also on the path taken. If it turns out that in Alice's decision process, the wrong considerations were salient, and feedback is negative, these considerations can be suppressed in the future.

To illustrate, suppose that there are two cases in which Alice tells a lie with bad consequences.[12] In case one, let us assume, her lying was based on the consideration that this course of action will spare her from having to hear Carl's sermon about Bob again, but as it turns out, he delivers it anyway. In that case, she will learn that lying is not the simple way out that she thought it was. In the other case, let us assume, her lying was based on the consideration that she simply didn't want Carl to know that she met Bob, while it turns out that Carl in fact had just met Bob, who was flattered by the attention he had received through Alice's ice cream book-keeping. In that case, she can learn that she should not deny her social contacts.

Finally, suppose Alice herself has access to the replay statistics, and knows that of 100 runs, she ends up telling the truth in exactly half the cases. It would still not be adequate for her to simply toss a coin and thereby simulate the statistics. That would not be satisfying, since it would amount to giving up the attempt to connect her action to reasons—which is what she did in each of the runs, each of which was based on her own considerations. If in the case of a meaningful decision, the link between action and reasons is given up, this threatens the agent's integrity. The mere fact that the deliberation process is indeterministic, on the other hand, does not.

6 Conclusion: Stochastic libertarianism

In this paper, we have proposed a novel approach to libertarianism. Contrary to most published attempts at establishing free agency under indeterminism, we choose a "let's face it" approach that posits indeterminism at the very heart of an agent's deliberation process. We do not invoke any special theory of agent-causal powers (even though we would be happy if our approach could be interpreted as a contribution to the understanding of agent causation), nor any metaphysically extravagant assumptions. The core of our theory of stochastic libertarianism is that a random process of deliberation leading to action can in principle ground the rational and attributable agency of

[12] The model of Fig. 1 is too simple to represent the following story; minimally, there would have to be an extra clip coupled to action a_2.

a learning agent. We do not claim that our theory is psychologically or neurologically adequate, though we are interested in attempts to link the theory to the concrete material basis of our own agency. In this paper, our aim has been to establish that the position of stochastic libertarianism is conceptually tenable, thus marking a possible stance in the free will debate. We have substantiated this claim by pointing out that there exists a formally well specified and physically well motivated model, projective simulation, that shows an instance of the stochastic dynamics that we postulate in our theory. Even if it turns out that the material basis of our own agency is completely different from that model, we claim that projective simulation provides a framework for the conceptual discussion, and perhaps also for the technological implementation, of stochastic libertarianism in embodied agents.

We are aware that in a single paper we cannot connect to the whole of the vast discussion of libertarianism. A number of issues—consciousness, moral responsibility, and arguments against compatibilism, to mention just a few—have been left out completely. However, if we have succeeded in establishing just the mere tenability of stochastic libertarianism, we have already achieved a number of important things. To list some of the most salient issues:

- We have pointed out a flaw in Van Inwagen's much-discussed replay argument and shown how indeterminism need not threaten an agent's integrity. This is important because that argument, or other arguments structurally like it, is seen as one of the main conceptual difficulties for libertarianism.

- We have presented a theory that shows how in our understanding of agency, we need not be afraid of indeterminism, but can view it as a resource. A number of technological advances, e.g., in cryptography, and conceptual advances, e.g., in quantum information, rely on that attitude of viewing indeterminism as a resource. We can hope to interact with these discussions in future research.

- Through a focus on learning and the historical development of an agent's identity, we have provided a framework in which we can make a clear, normative distinction between indeterminism-based rational agency and erratic behavior. Such a distinction is desparately needed if one wants to establish libertarianism, since most arguments against that view boil down to conflating the two. Stochastic libertarianism shows how to separate the functional employment of randomness from haphazard chanciness.

- While indeterministic models have been considered in philosophy of action, often these invoke an unclear capacity of the agent to make a current reason active. That further active moment in agency threatens to become a homunculus, thereby just passing the problem of the initiation of action on to the next level. In our approach, no such extra entity is needed. Pure natural randomness, such as provided for by quantum mechanics, can do the job.

- Stochastic libertarianism, by putting randomness at the core of agency, claims that indeterminism is not just something that can be tolerated to some extent, but something that is needed to begin with. Embarking on this approach, we can provide a theory that may be judged by its explanatory merits vis-à-vis compatibilist contenders. Certainly a lot of modeling, simulation, and further conceptual work is needed to deepen our understanding of the theory, but a start has been made.

To sum up: Stochastic libertarianism shows that integrity in action does not need determinism.

References

Briegel, H. J. and De las Cuevas, G. (2012). Projective simulation for artificial intelligence. *Scientific Reports*, 2(400). doi:10.1038/srep00400.

Briegel, H. J. and Müller, T. (2013). A chance for attributable agency. *submitted.*

Broad, C. D. (1933). *Examination of McTaggart's Philosophy.* Cambridge University Press, Cambridge.

Keil, G. (2007). *Willensfreiheit.* De Gruyter, Berlin.

Soon, C. S., Brass, M., Heinze, H.-J., and Haynes, J.-D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, (11):543–545. doi:10.1038/nn.2112.

Steward, H. (2012). *A Metaphysics for Freedom.* Oxford University Press, Oxford.

Van Inwagen, P. (2000). Free will remains a mystery: The eighth Philosophical Perspectives lecture. *Philosophical Perspectives*, 14:1–19.