Causality in Physics and Effective Theories of Agency

Daniel A. Roberts
Facebook AI Research
New York, NY 10003
danr@fb.com

Max Kleiman-Weiner
Harvard University
Cambridge, MA 02138
maxkleimanweiner@fas.harvard.edu

Abstract

We propose to combine reinforcement learning and theoretical physics to describe effective theories of agency. This involves understanding the connection between the physics notion of causality and how intelligent agents can arise as a useful effective description within some environments. We discuss cases where such an effective theory of agency can break down and suggest a broader framework incorporating theory of mind for expanding the notion of agency in the presence of other agents that can predict actions. We comment on implications for superintelligence and whether physical bounds can be used to place limits on such predictors.

1 Causality in Physics

We start with a description of what is meant by causality in microscopic physics.

The modern notion of causality arises out of Einstein's principle of relativity [1], as consequence of the fact that mass or energy cannot travel faster than the speed of light c.[1] Causality shows up when trying to determine the influence of a particle or *degree of freedom* at a particular spatial point $\mathbf{x}$ and time t on another degree of freedom at a different spatial point $\mathbf{y}$ and time t'. As a consequence of the fact that the universal speed limit is c, if the distance squared separating any two degrees of freedom is greater than the distance squared that light can travel in the time interval separating their occurrence, so that $||\mathbf{x}-\mathbf{y}||^2 > c^2|t-t'|^2$, then the two degrees of freedom are too separated in space *and* time to have been able to communicate with each other. Such degrees of freedom are called *spacelike* separated in spacetime, and there's no notion of whether one is in the past or the future of the other. In other words, they are *causally disconnected* from each other.

Since experimentally we find that our universe satisfies the principle of relativity, any framework that describes microscopic physics must be relativistic in nature.The microscopic laws of physics are expressed in a framework known as quantum field theory. As quantum field theory obeys the principle of relativity, in rigorous formulations microscopic causality is encoded as an axiom [2]. Since quantum field theory is also a *quantum* theory, the precise statement replaces vague terms like "communicate" and "influence" with the effect of "measurements" on "observables"—i.e. things that can be measured.

In particular, a question about causal influence is the question of whether the effect of a measurement of some observable W at point x and time t can affect later measurements of a some observable V at point y and time t'. The statement of causality is that such measurements cannot affect each other if the particles are spacelike separated. The precise statement is that the observables must commute as operators

$$[W(\mathbf{x},t),V(\mathbf{y},t')]=0, \qquad ||\mathbf{x}-\mathbf{y}||^2 > c^2|t-t'|^2, \tag{1}$$

where the expression $[W(\mathbf{x},t),V(\mathbf{y},t')] \equiv W(\mathbf{x},t)\,V(\mathbf{y},t') - V(\mathbf{x},t)\,W(\mathbf{y},t')$ is a commutator, which measures the degree to which the operators fail to satisfy a commutativity property. In physics

[1] Violating this notion of relativity could be used to create logical paradoxes, e.g. the kind that occurs if you go back in time and kill your grandfather before he could conceive your mother.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

terms, this is a mathematical object representing the effect of measurements of V at spacetime point $(\mathbf{y}, t')$ and then W at spacetime point $(\mathbf{x}, t)$ compared to measuring W first at $(\mathbf{x}, t)$ followed by measuring V at $(\mathbf{y}, t')$. The statement of causality (1) is that this must vanish when the points $(\mathbf{x}, t)$ and $(\mathbf{y}, t')$ are spacelike separated. Not only can there be no causal relationship between these degrees of freedom, they also cannot correlate.

This makes intuitive sense, if the particles are so separated that they cannot communicate or exchange energy, then certainly whatever disturbance a measurement causes at $\mathbf{x}$ at time t can't reach $\mathbf{y}$ by time t' to have any influence on the outcome of the future measurement. Since we can also think of "measurement" as the outcome of experiments, this also maps onto our intuitive notion of causality: performing an experiment and making a measurement is a type of "intervention" and this constraint tells us when such experiments are allowed to have a causal relationship and when they are not. In other words, if counterfactually one makes a measurement that one wouldn't have otherwise been performed, this can only have any effect on the environment in regions that are not spacelike separated from the measurement. Note that the idea of a measurement is sufficiently general to encompass most changes that can be made to the environment at a particular place in space and time.

What if the observables are not spacelike separated so that $||\mathbf{x} - \mathbf{y}||^2 < c^2|t - t'|^2$, and they can influence each other? In physics, we'd say that they are "in causal contact" and equations like (1) are used to quantify the degree of influence. In order to actually get a number out (1), we need a way of comparing W and V at the same time.[2] This means we need to know how observables change in time, which would be expressed as some kind of function that takes W at time t to time t'. If we had such a function, that would tell only us how W correlates with time.[3] What we want is an understanding of how interactions with other observables, other measurements, other degrees of freedom can influence the way W changes with time. For this, we have to specify a particular quantum field theory within the framework.[4]

The reason we described quantum field theory as a framework is because it allows the expression of any causal model that is consistent with the constraints of relativity and quantum physics. In theoretical physics, models are specified by enumerating which particles are allowed to interact and how strong their interaction is. This is usually done in terms of a Lagrangian, and the principle of least action gives a partial differential equation describing the evolution of the observables, in expectation.[5] A key point is that such a differential equation encodes the dynamics of the observables. If the differential equations for different observables are coupled, then they can influence each other causally and build up nontrivial correlations. Such differential equations *are* the causal model, and quantum field theory is a framework for building such models subject to physical constraints that prevent logical contradictions.

In Appendix A, we give an example of how Newton's second law is a framework for expressing causal models in the context of classical non-relativistic physics and further discuss how the coupling of equations relates to causal influence. In the main body the paper, we shift our focus to understanding how this rigid framework could allow for intelligent agents to emerge.

2 Effective Theories of Agency

All quantum field theories that we use to describe our universe are effective (field) theories [6], including the standard model [3–5], offering only a phenomenological (but causal) model of microscopic physics. Effective field theories try to relate bottom-up knowledge to top-down knowledge. Such bottom-up theories allow for much simpler descriptions that the top-down descriptions (i.e., if we had to describe humans but we're only allowed to discuss the dynamics of the fundamental particles that make us and our environment up) by focusing only on the *relevant* degrees of freedom at the scale of interest [7, 8]. The project of physics might be completed when top down completely

[2] Technically, here we mean the same time slice.

[3] Note that we didn't actually need a differential equation to determine whether W, V are allowed to be in causal contact in (1). In microscopic physics, there are no latents variables. Any time variables are correlated, there is necessarily a causal relationship between them. However, to understand the mechanism of that relationship—how they are related—we need something more.

[4] The particular quantum field theory that describes our universe (at low energies and without gravity) is known as the Standard Model [3–5].

[5] The full distribution is given by Feynman's path integral, but the distinction isn't important for our purposes.

intersects bottom up; when we have simple descriptions of basic phenomena and understand how to connect them to some ultimate fundamental theory. However, the bottom-up effective descriptions can breakdown, and the framework of effective field theory tells you when you need to revert to the more complicated but fundamental description.

In this vein, it is sometimes useful to introduce the idea of an agent that can make independent decisions as a way of keeping track of the macroscopic behavior of some systems. Let's consider this in the framework of reinforcement learning [9].

Central to reinforcement learning is the distinction between agents and the environment. Formally, a Markov Decision Process (MDP) consists of a set of states $\mathcal{S}$, actions $\mathcal{A}$, rewards $\mathcal{R}$, with the transition probability of going from state s to s' conditioned on taking action a given $p(s'|s,a)$, for $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$. In general, we assume each state $s \in \mathcal{S}$ allows an agent to take some subset of the actions $\mathcal{A}(s) \subseteq \mathcal{A}$. In this framework, an agent takes actions that can affect the environment. In particular, an agent in a certain state selects from a set of actions that often lead to different future states in turn. The environment provides a reward in some states. In this formalism, agents attempt to learn a policy that specifies how to select actions in a particular state to maximize overall reward.

In the MDP framework, the agent is considered to be independent of the environment. This can be seen formally by noting that the process by which an agent selects one action or another is outside the specification of the particular state. Formally, an agent acts according to a policy $\pi(a|s)$, which specifies the probability in state s of taking action $a \in \mathcal{A}(s)$. The information in $\pi(a|s)$ is not part of the MDP, but is something external. In state s, the agent is completely *free* (as in *will*) to take any of the actions in $\mathcal{A}(s)$.

For example, in a game such as chess, the board configuration and the rules describing how the pieces can move are considered to be part of the definition of the game itself. However, a particular player's strategy $\pi(a|s)$ is external to the game. Thus players are free to make any of the allowable moves they'd like. Perfect knowledge of the state s is not sufficient to predict what action a player will take.[6] An agent may also build models of its environment by attempting to learn the transition rules between states $p(s'|s,a)$. By experimenting on the environment such a scientist-agent may build causal models of its world, just as scientist use these principles to learn the fundamental laws of physics as we discussed in the previous section [10, 11].

As has been pointed out by numerous philosophers and scientists and college students throughout the ages (see [12] and references therein), this notion of an agent is in some tension with fundamental physics laws (differential equations) that rigidly and deterministically specify the dynamics given initial conditions. In particular, no distinction was made between degrees of freedom in the environment (i.e., making up the states in $\mathcal{S}$) and degrees of freedom representing the agent (and storing the policy $\pi(a|s)$ and generating decisions a).

In microscopic physics, there's only states $\mathcal{S}$. The transitions between states is dictated by the time evolution operator U, which determines how the state changes from time t to time $t + \delta t$

$$s_0 \xrightarrow{U} s_1 \xrightarrow{U} \ldots \xrightarrow{U} s_t \xrightarrow{U} \ldots \xrightarrow{U} s_\infty. \tag{2}$$

However, encapsulated in this simple Markov Process (since clearly there are no decisions here) is a rich enough structure to describe intelligent agents that think of themselves as free enough to do science, build causal models, and even work on reinforcement learning building intelligent agents that participate in seemingly more complicated MDPs. Explaining how such intelligent agents interface with the fundamental laws is known as the "observer problem" in physics, and is central to much confusion about interpretations of quantum mechanics.

The process (2) can describe any MDP if we imagine defining a new set of states $\mathcal{H}$ that not only specify the environment ($\mathcal{S}$ and $\mathcal{R}$), but also specify the function $\pi(a|s)$ and the random seed (if the policy is deterministic). The state $h \in \mathcal{H}$ would contain information about the function approximation of $\pi(a|s)$, the learning rule given rewards, and thus all the choices and therefore states would be determined in advance. If (2) specified a chess game played by humans, it would have to include the mental states of both players as part of the state of the game. Clearly, this is a cumbersome way to talk about chess, since we don't get to make use of what's universal between chess games (the rules) and what's variable (the players and their strategies).

[6] Though in principle it is sufficient to predict the optimal action.

So, agency is often a useful abstraction. If the true state approximately factorizes into a part that is "environment-like" and a part that's "agent-like" then it might be useful to use MDP formalism to think about the dynamics. The idea of having free will to make such choices is a phenomenological description—it's an *effective theory*, a useful abstraction with a certain range of validity. The effective theory gives an agent a choice, say between actions a and a', when in state s. Ultimately, we know that when we include the agent as part of the state specification, that the agent only has the illusion of choice, but is predestined to take action a in state s. If they were predestined to take action a', then they would have really been in some other state s'. The effective description of an agent is useful if they factorize from the environment such that the states s, s' are approximately the same such that other agents cannot realistically distinguish between them.

2.1 Effective Theories can break down

Importantly, effective field theories in physics let you know when they break down and the description is no longer valid. Does the same kind of reasoning apply to effective theories of agency? We've belabored the point that effective theories are useful, but the verb "use" requires an object specifying to whom the theory is useful for. Effective theories are useful to particular observers doing experiments, or, in this case, agents making decisions. So an effective theory of agency will be a useful way for either an agent to think about itself or about other agents.

One hypothesis is that people use an effective theory of agency to understand their own behavior i.e., we commonly perceive our own behavior as if we are agents able to make free choices in the sense described above. Since this is a perception, it can be shattered. To be concrete, imagine that Alice is asked to make k binary decisions in response to k questions. We also imagine that Alice is told these questions sequentially and doesn't know what decision she will make until hearing the question and then thinking about how to answer it. If another agent, Bob, were to write down the answer to all k questions ahead of time (which has an exponentially small probability of happening by chance), the effective description of Alice as an agent has broken down. Bob must know how Alice will answer these k questions before Alice even knows what the k questions are.

From Bob's perspective, Alice is no different than some part of the environment. In the language of Dennett [13], Bob uses the *physical stance* to describe Alice. She is no different from a simple physical system, without any control over how she answers the k questions. Even if Alice's behavior includes some stochasticity, such a policy plus the hidden seed used to draw random samples could simply be included in Bob's model of the states and their transition rules. There's no confusion for Bob, since he doesn't distinguish between an agent like Alice and the environment. To him, it's all environment.

What would it feel like to lose one's perception of oneself as an agent distinct from the environment? Is a powerful predictor such as Bob necessary to experience this loss of agency? These questions are highlighted by studying constructed situations, such as Newcomb's problem [14], where Bob can exploit Alice if she doesn't drop her (in this case naive) effective notion of agency.[7] Since the naive effective theory of Alice can break down in the presence of Bob, Alice should change her behavior otherwise she be exploited.

To know whether or not to fall back to this effective theory, Alice need a model how Bob models her. This recursive theory of mind perspective is necessary, otherwise she can be exploited. In future work, we intend to explore this directly in some toy multi-agent models to better understand how these effective theories can work.

3 Limits in the Future

How practical is it for an agent to have a perfect model of an another? Questions like this are of interest to researchers interested in existential risk from interacting superhuman-level AI or superintelligence (see e.g., [22]). To answer this question, we can turn back to physics and the way any effective theory

[7] This problem has received a lot of attention in the decision theory community, the philosophy community and the AI-safety community [14–16, 16–21]. However, it hasn't really received any attention from the physics community. In the literature, treatment of this problem often makes use of "backwards-in-time causation" that seems to violate causality. Instead, we suggest that the difficulty is not with causation but instead is related to Alice using the wrong effective theory of agency to describe herself.

of agency must be grounded in the fundamental laws of physics. In a relativistic universe (such as our own), a major problem for Bob occurs if Alice starts conditioning her decisions on the outcome of events that are outside of Bob's lightcone, such as quasars or patches of the cosmic microwave background.[8] Such information is "causally disconnected" from Bob and completely unknowable to him. Even if Bob contains a perfect simulation of Alice such that Alice can't tell whether she is the simulated or "true" version (as in Aaronson's resolution of Newcomb's problem [17]), the simulations will begin to diverge as soon as Alice uses information that's causally disconnected from Bob to make her decisions, rendering simulated Alice not predictive of real Alice.

Of course, we don't tend to condition our decisions on cosmic events,[9] so this would have to be an active change in the way we make decisions. (Note that one can use this more practically than just generating random and unpredictable outcomes. For instance, if Alice wanted to hide in a location unknown to Bob, she could use a fixed algorithm with input from cosmic signals outside Bob's lightcone to generate the location. Of course, Bob would know Alice is predetermined or precommitted to behaving in this manner and choose a different set of actions so as to not allow Alice to do this.) From Alice's perspective, a theory of mind understanding of different effective theories of agency would be useful in knowing when to implement such strategies. (Which, adversarially, would incentivize Bob not to use his predictive power over Alice to exploit her, less she learn of his abilities and switch to a different effective theory of agency.)

More generally, Alice's interactions with the environment constantly creates causal influences that can be moving away from her at the speed of light. If Bob doesn't have a high-level effective model of Alice but instead is using the physical stance to model all the particles that make her up, he might run into severe computational difficulties simulating her. Such a simulation is very computationally difficult due to chaos and the butterfly effect; small changes of the system (Alice and her interactions with the environment) lead to exponentially growing changes in the system's (i.e. Alice's plus the environment's) behavior. The laws of physics limit how much information can be stored in a region [29–33], how fast information can spread in space [34–38], how fast chaotic perturbations can grow [39], and perhaps limit how complex a system can become [40, 41].

This manuscript is a first attempt at using these sorts of ideas from physics in conjunction with models of multi-agent scenarios. In future work we will formally estimate the difficulty of simulating even a simple RL agent and to what extent the laws of physics constrain the behavior and power of a superintelligence's ability to model it.

Acknowledgments

We would like to thank Robbie Kubala for participating in an initial collaboration on this material and Voldemort for bringing it to our attention. We are additionally grateful to Léon Bottou and Josh Tenenbaum for comments and discussions. DR is grateful for the hospitality of the MIT Computational Cognitive Science Group during the completion of this work. This extended abstract was brought to you by the renormalization scale Λ and grew to encompass the lightcone with Lyapunov exponent λ.

A Frameworks for Causal Models

To discuss how such frameworks work, we don't need to introduce the actual machinery of quantum field theory. Instead, we can talk about a simpler (though ultimately less general) framework that is much more familiar, namely Newton's second law.

In the epilogue of his seminal work [42], Pearl brings up an argument that Bertrand Russel made in 1913 that causality doesn't show up in physics. Russel points out that the laws of physics are time reversal invariant, and yet physicists still talk about *cause* and *effect* as if they're included in the laws. (We directly addressed the way in which causality appears in microscopic physics in §1.)

[8] A similar mechanism was proposed in [23] and experimentally tested in [24–26] to close a "free will" loophole in tests of Bell's inequality [27, 28] that rule out local hidden variable theories for quantum mechanics.

[9] Though see [17] for an fanciful idea on how such a mechanism could be used to put free will back in the laws of physics, essentially hiding it in initial conditions

Pearl brings up Newton's second law as

$$F = ma, \tag{3}$$

commenting that we can algebraically rearrange this expression in many different forms, e.g.

$$m = a/F, \qquad m = F/a. \tag{4}$$

Despite this, he continues, we say colloquially that force causes acceleration and not the other way around. And we certainly never say that the ration F/a causes mass. He echoes Russell wondering where this notion of *cause* comes from, since clearly it doesn't come from the equations themselves. He suggests that we need something additional to distinguish the fact that humans understand that it's the force that "causes" motion.

The confusion arises because F, m, and a are just seen as arbitrary mathematical symbols. The resolution comes from understanding what (3) is actually supposed to represent. In particular, Newton's second law is a second order differential equation for the position x of a particle as a function of time t. Note that in (3), neither x or t even appear!

The more correct way of writing Newton's second law is as

$$F(x) = m\frac{d^2x(t)}{dt^2}, \tag{5}$$

which relates a function of $x(t)$, the force $F(x)$ to the second derivative of $x(t)$ with respect to time. The solution of this equation for a specified $F(x)$ plus the particle's initial position and initial velocity will allow you to predict the location and velocity of the particle at any later (and also any earlier) time. The mass m is just there for dimensionality and could even be absorbed into a new function $f(x) \equiv F(x)/m$ that represents force per unit mass.

The force is an arbitrary function of x. Another word for it is a "source" for the dynamics of x. We see that x is the primary variable, and the form of the function $F(x)$ controls how x behaves since it's proportional to its second derivative. This form (5) makes it very clear that it's the force that causes motion. The fact that this holds generally for all choices of $F(x)$ is what makes Newton's second law so powerful.[10]

The confusion arises because (3) is not actually an algebraic equation, and there are actual and essential asymmetries between the variables. What Newton's second law is, in its proper form (5), is a model. In fact, it's a precisely a general causal model of the dynamics. The input to the model is always $F(x)$ no matter how you move it around algebraically. The model is a differential equation that has to be solved.

Once you solve the model, you get a function $x(t)$; the position is a function of time. At this point, without any additional information, it's nonsense to say whether time causes the position or position causes time; they are just observed to correlated in a specific way. It's the underlying model or theory that we understand that tells us the cause.

As an aside, we should compare this with an actual algebraic relation. Following [42], it's natural to describe the two equivalent equations

$$2x + y = 0, \qquad y = -2x \tag{6}$$

in a very different way. We might say that the former has no structure at all, whereas the rearrangement to give the latter relation $y = -2x$ tells us that y is determined by x. Since the rules of algebra tell us that both forms are actually equivalent without some additional information. (Such information might include telling us whether the symbol '$=$' has to do with *assignment* or *equality*. In some fields, this confusion is sometimes alleviated by introducing a different symbol e.g. '$\equiv$' to denote assignment.)

A solution to Newton's second law, for example $x = -(1/2)gt^2$ for a particle starting at a position $x = 0$ dropped from rest in a constant gravitational field with strength g, is exactly like the algebraic equation $y = -2x$. We could also have written this as $t = \sqrt{-2x/g}$, telling us the time we expect to see the particle at position x. Note that mass and force and acceleration appear nowhere in this

[10] This is true so long as m isn't too small or too large or $F(x)$ isn't too large, in which case quantum and relativistic effects have to be taken into effect. This just means we replace Newton's Law with a different differential equation.

"solution" which just tells us that x and t have some kind of relationship. But in order to reason counterfactually and in order to understand causality, you can't just have $x = -(1/2)gt^2$. You need the model, $F = ma$, with $F \equiv -mg$ and $a \equiv \frac{d^2x}{dt^2}$. In fact, this is readily seen in human causal judgment. To model and explain human causal judgments and counterfactual simulations of causal scenes requires intervening on the underlying dynamic model not just the particular trajectories [43–45].

A.1 Causality and interactions

Considered the coupled set of differential equations

$$f(x,y) = m_x \frac{d^2x}{dt^2}, \qquad g(x,y) = m_y \frac{d^2y}{dt^2}, \tag{7}$$

which have the interpretation of two classical particles interacting. The particles are labeled x and y, with masses m_x and m_y, respectively. The dynamics of x is determined by the force on x, $f(x,y)$, which is a function of the positions of x and y. Similarly, the dynamics of y depend on x and y through the function $g(x,y)$. This is really just a slight generalization of our previous discussion since we could have also written this system so that it looks like "$F = ma$"

$$F(X) = M\frac{d^2X}{dt^2} \iff \begin{bmatrix} f(x,y) \\ g(x,y) \end{bmatrix} = \begin{bmatrix} m_x \\ m_y \end{bmatrix} \frac{d^2}{dt^2} \begin{bmatrix} x \\ y \end{bmatrix} \tag{8}$$

if X is the vector $X \equiv (x,y)^T$, M is the vector $M \equiv (m_x, m_y)^T$, and $F \equiv (f,g)^T$. In other words, it's now a matrix equation. Therefore, as per our previous discussion physicist would be justified in saying that the force $f(x,y)$ causes the dynamics of x, and $g(x,y)$ causes the dynamics of y. On the other hand, the forces themselves are determined by the instantaneous positions of the particles. *But since the force is a function and not a differential equation, it an algebraic statement, like* (6)*. Thus, it does not make sense to say that the particles' positions* cause *the force. Instead, we might say that the force is determined by them.*

As before, solutions to this model will have the form $x(t)$ and $y(t)$. In a regime where $y(t)$ is invertible, we can find $t(y)$ and then express x as a function of y as $x(y) \equiv x(t(y))$. This makes it completely clear that functional or algebraic relationships naturally express correlations; $x(y)$ is a (deterministic) statement about how x and y correlate, but contains no causal information. The causal information is contained in the model of their dynamics, (7), which tells us how forces on the particles determine their motion. Once we have the solution to the differential equation, we're only left with correlations and are free to rearrange them $x(t)$, $y(t)$, $t(y)$, $t(x)$, $x(y)$, $y(x)$, etc., as we like.

One point made by Mach [46] is that the concept of a force may be superfluous. Consider the case of Newtonian gravity

$$f(x,y) = g(x,y) = -G_N \frac{m_x m_y}{(x-y)^2}, \qquad \text{(Newtonian gravity)}, \tag{9}$$

where G_N is Newton's gravitational constant, and where for this to be sensible x, y are now taken to be radial coordinates for two particles that live in three spatial dimensions. Mach argues that we should just interpret (7) as two dynamical equations that describe the motion without any need to invoke the concept of a force. Our point is that precisely because $F = ma$ is a framework for describing causal models, for which universal gravitation is one specific model, necessitates the introduction of the concept of force.

An related comment is that in (9), the cause of motion of the particles is the gravitational interaction of the particles, which itself is caused by their relative positions, which changes due to gravity.... For coupled equations, there is some notions of cause and effects being tangled up. That is, the cause of a particle's motion at time t might be the effect of a different particle at time $t - \epsilon$ that was previously perturbed by the original particle at time $t - 2\epsilon$.

One interesting point about coupled equations is whether there's a change of basis that decouples the equations. If not subject to other outside forces, then this means that the degrees of freedom in the decoupled basis are not actually interacting. In other words, they do not have any causal relationship to each other. This is despite the fact that the particles themselves (as seen in the original basis) do seem to have a complicated relationship. There's a technical point, which is that we expect this to

happen for linear second order differential equations, such as Newton's second law with particularly simple choices for the forces $f(x, y)$ and $g(x, y)$.

More generally, the dynamical equations are nonlinear. In this case, we say that the particles are *interacting*. In this case, the causal relationship is nontrivial (though usually very hard to solve exactly). In the extended version of this work, we will explain this point carefully using some simple systems (such as coupled oscillators) and simple examples from particles physics.

As an aside, the discussion in this section has focused on deterministic classical physics. We don't expect that extending the discussion to the evolution of classical probability distributions should make too much of a difference. Similarly, since agents are classical, we don't think quantum mechanics should be too important in the present discussion either. (This is true even though the discussion in §1 of where causality ultimately comes from in physics was in the context of quantum field theory.)

References

[1] Albert Einstein. On the Electrodynamics of Moving Bodies. *Annalen der Physik*, 17:891–921, 1905.

[2] Raymond F Streater and Arthur S Wightman. *PCT, spin and statistics, and all that*. 1964.

[3] S. L. Glashow. Partial Symmetries of Weak Interactions. *Nucl. Phys.*, 22:579–588, 1961. doi: 10.1016/0029-5582(61)90469-2.

[4] Steven Weinberg. A Model of Leptons. *Phys. Rev. Lett.*, 19:1264–1266, 1967. doi: 10.1103/PhysRevLett.19.1264.

[5] Abdus Salam. Weak and Electromagnetic Interactions. *Conf. Proc.*, C680519:367–377, 1968.

[6] Steven Weinberg. Phenomenological Lagrangians. *Physica*, A96(1-2):327–340, 1979. doi: 10.1016/0378-4371(79)90223-1.

[7] Kenneth G. Wilson. Renormalization group and critical phenomena. 1. Renormalization group and the Kadanoff scaling picture. *Phys. Rev.*, B4:3174–3183, 1971. doi: 10.1103/PhysRevB.4.3174.

[8] Kenneth G. Wilson. Renormalization group and critical phenomena. 2. Phase space cell analysis of critical behavior. *Phys. Rev.*, B4:3184–3205, 1971. doi: 10.1103/PhysRevB.4.3184.

[9] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. A Bradford book. Bradford Book, 1998. ISBN 9780262193986.

[10] Michael R. Waldmann. *The Oxford Handbook of Causal Reasoning*. Oxford University Press, 05 2017.

[11] James Woodward. *Making Things Happen: A Theory of Causal Explanation*. Oxford Studies in the Philosophy of Science. Oxford University Press, 2005.

[12] Timothy O'Connor and Christopher Franklin. Free will. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition, 2018.

[13] Daniel Dennett. Intentional systems theory. In Brian P. McLaughlin Ansgar Beckermann and Sven Walter, editors, *The Oxford Handbook of Philosophy of Mind*. Oxford University Press, 2009. ISBN 9780199262618.

[14] Robert Nozick. Newcomb's problem and two principles of choice. In *Essays in honor of Carl G. Hempel*, pages 114–146. Springer, 1969.

[15] Nate Soares and Benja Fallenstein. Toward idealized decision theory. *arXiv preprint arXiv:1507.01986*, 2015.

[16] Eliezer Yudkowsky and Nate Soares. Functional decision theory: A new theory of instrumental rationality. 2017.

[17] Scott Aaronson. The Ghost in the Quantum Turing Machine. URL `https://arxiv.org/abs/1306.0159`.

[18] Robert C Stalnaker. Letter to david lewis. In *Ifs*, pages 151–152. Springer, 1980.

[19] David Lewis. Prisoners' dilemma is a newcomb problem. *Philosophy & Public Affairs*, pages 235–240, 1979.

[20] Allan Gibbard and William L Harper. Counterfactuals and two kinds of expected utility. In *Foundations and applications of decision theory*, pages 125–162. Springer, 1978.

[21] A. Ahmed. *Evidence, Decision and Causality*. Cambridge University Press, 2014. ISBN 9781107020894.

[22] N. Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014. ISBN 9780199678112.

[23] Jason Gallicchio, Andrew S. Friedman, and David I. Kaiser. Testing Bell's Inequality with Cosmic Photons: Closing the Setting-Independence Loophole. *Phys. Rev. Lett.*, 112(11):110405, 2014. doi: 10.1103/PhysRevLett.112.110405.

[24] Johannes Handsteiner et al. Cosmic Bell Test: Measurement Settings from Milky Way Stars. *Phys. Rev. Lett.*, 118(6):060401, 2017. doi: 10.1103/PhysRevLett.118.060401.

[25] Dominik Rauch et al. Cosmic Bell Test Using Random Measurement Settings from High-Redshift Quasars. *Phys. Rev. Lett.*, 121(8):080403, 2018. doi: 10.1103/PhysRevLett.121.080403.

[26] Ming-Han Li et al. Test of Local Realism into the Past without Detection and Locality Loopholes. *Phys. Rev. Lett.*, 121(8):080404, 2018. doi: 10.1103/PhysRevLett.121.080404.

[27] John S. Bell. On the Problem of Hidden Variables in Quantum Mechanics. *Rev. Mod. Phys.*, 38:447–452, 1966. doi: 10.1103/RevModPhys.38.447.

[28] John F. Clauser, Michael A. Horne, Abner Shimony, and Richard A. Holt. Proposed experiment to test local hidden variable theories. *Phys. Rev. Lett.*, 23:880–884, 1969. doi: 10.1103/PhysRevLett.23.880.

[29] J. D. Bekenstein. Black holes and the second law. *Lett. Nuovo Cim.*, 4:737–740, 1972. doi: 10.1007/BF02757029.

[30] Jacob D. Bekenstein. Black holes and entropy. *Phys. Rev.*, D7:2333–2346, 1973. doi: 10.1103/PhysRevD.7.2333.

[31] Jacob D. Bekenstein. A Universal Upper Bound on the Entropy to Energy Ratio for Bounded Systems. *Phys. Rev.*, D23:287, 1981. doi: 10.1103/PhysRevD.23.287.

[32] Gerard 't Hooft. Dimensional reduction in quantum gravity. In *Salamfest 1993:0284-296*, pages 0284–296, 1993.

[33] Leonard Susskind. The World as a hologram. *J. Math. Phys.*, 36:6377–6396, 1995. doi: 10.1063/1.531249.

[34] E. H. Lieb and D. W. Robinson. The finite group velocity of quantum spin systems. *Commun. Math. Phys.*, 28:251–257, 1972. doi: 10.1007/BF01645779.

[35] Matthew B. Hastings. Locality in quantum systems. 2010.

[36] Stephen H. Shenker and Douglas Stanford. Black holes and the butterfly effect. *JHEP*, 03:067, 2014. doi: 10.1007/JHEP03(2014)067.

[37] Daniel A. Roberts, Douglas Stanford, and Leonard Susskind. Localized shocks. *JHEP*, 03:051, 2015. doi: 10.1007/JHEP03(2015)051.

[38] Daniel A. Roberts and Brian Swingle. Lieb-Robinson and the butterfly effect. *Phys. Rev. Lett.*, 117(9):091602, 2016. doi: 10.1103/PhysRevLett.117.091602.

[39] Juan Maldacena, Stephen H. Shenker, and Douglas Stanford. A bound on chaos. 2015.

[40] Adam R. Brown, Daniel A. Roberts, Leonard Susskind, Brian Swingle, and Ying Zhao. Holographic Complexity Equals Bulk Action? *Phys. Rev. Lett.*, 116(19):191301, 2016. doi: 10.1103/PhysRevLett.116.191301.

[41] Adam R. Brown, Daniel A. Roberts, Leonard Susskind, Brian Swingle, and Ying Zhao. Complexity, action, and black holes. *Phys. Rev.*, D93(8):086006, 2016. doi: 10.1103/PhysRevD.93.086006.

[42] J. Pearl. *Causality*. Causality: Models, Reasoning, and Inference. Cambridge University Press, 2009. ISBN 9780521895606.

[43] Kevin A Smith and Edward Vul. Sources of uncertainty in intuitive physics. *Topics in cognitive science*, 5(1):185–199, 2013.

[44] Tomer D Ullman, Elizabeth Spelke, Peter Battaglia, and Joshua B Tenenbaum. Mind games: Game engines as an architecture for intuitive physics. *Trends in cognitive sciences*, 21(9): 649–665, 2017.

[45] Tobias Gerstenberg, Matthew F Peterson, Noah D Goodman, David A Lagnado, and Joshua B Tenenbaum. Eye-tracking causality. *Psychological science*, 28(12):1731–1744, 2017.

[46] Ernst Mach and T. J. McCormack. *The Science of Mechanics: A Critical and Historical Account of Its Development*. Open Court Publishing Company, 1988.