Learning on a Slippery Slope

Frank C. Krysiak
Department of Business and Economics, University of Basel, Switzerland

Lukas Tschabold*
Department of Business and Economics, University of Basel, Switzerland

Abstract

The effects of large-scale changes to the natural environmental are often uncertain. In many cases, it is only possible to learn these effects by allowing for some change. However, this can induce a risk of losing control over such an "experiment"; processes might be initiated that drive the system towards a new equilibrium, irrespective of whether this is desirable or not. In this paper, we analyze a simple model of experimenting with change that captures these points: the explorative argument (benefits of learning) as well as the "slippery slope" argument (losing control). Our study thereby extends the literature on (quasi-) option values by considering a setting where information is not gained by waiting but only by acting. We show that in such a setting, it is surprisingly often optimal either to abstain from initiating change or to only experiment on a small scale. Large-scale "experiments" with the natural environment also can be optimal, but under much more stringent conditions than a conventional analysis would suggest.

Keywords: Uncertainty, Environmental Change, Learning, Renewable Resources, Option Value

JEL: Q01, Q54, Q20, D81, D90

*Corresponding author, University of Basel, Department of Business and Economics, Peter Merian-Weg 6, CH-4002 Basel, Switzerland, Lukas.Tschabold@unibas.ch

1 Introduction

Since a long time, mankind has caused substantial changes to the natural and cultural environment in which it exists. Even in the fifth century B.C., Herodotus comments on large scale changes to the natural environment caused by human activity. More recently, the french revolution, industrialization, and advances in ICT have altered the conditions of human existence profoundly. Climate change, the introduction of GMOs in agricultural systems, or European integration provide examples of ongoing experiments with large-scale changes.

Many of the above changes have turned out to be beneficial. However, a common trait of all these examples is that at the time where change has been initiated, its eventual effects have not been predictable. The people starting the french revolution could not possibly have known how they would fare under the changed societal conditions, nor can we foresee today what the eventual effects of an unmitigated climate change might be. For such large-scale changes, it is even fair to argue that the effects can only be learned by experience. We cannot possibly predict how ecosystems and human societies will respond to a drastically changed climate; if we want to know for sure, we have to experience climate change at least to some extent.

However, experimenting with large-scale changes involves substantial risks. If climate change turns out to have drastic or even catastrophic consequences, it might have been better not to conduct such a global experiment. Furthermore, it might well be that, by allowing for some change to happen, we might lose control of the experiment. Positive feedback effects in the climate system might accelerate the change to an extent, where it becomes impossible to go back to the present state of our natural environment even on a very long time-scale. Similarly, GMOs introduced into the natural environment might spread in a way that makes it impossible to limit their geographic range.

In a social or cultural context, the latter problem is often referred to as stepping onto a slippery slope; we might initiate a process that we cannot stop and that leads to a state where we do not want to be. In fact, the "slippery slope argument" is pervasive in discussions on major social or environmental changes. Conservatives frequently argue that it is better not to experiment with such changes, because even a small step might entrap us in a detrimental but unstoppable process; we risk skidding down a slippery slope to an unpleasant end.

Surprisingly, economists have not fully taken up these points in their analysis of environmental change. The essence of both arguments, the explorative view of change and the danger of losing control, is not captured in most of economic studies of this problem. Since the seminal contribution of Arrow and Fisher (1974), uncertainty has received considerable attention in the analysis of environmental change. But most studies assume that uncertainty is resolved over time, irrespective of whether change is initiated or not. This is a "become wise by waiting" perspective that offers no incentives for experimenting with change. Furthermore, once better information is available, change can be stopped (albeit not always re-

versed), which does not allow for a "slippery slope" type of argument. Thus the trade-off between learning from an "experiment" and the risk that this experiment might run out of control is not covered.

Seen from this perspective, existing studies that use the Arrow and Fisher setting are useful for understanding and assessing the effects of uncertainty and irreversibility in small- and medium-scale projects. In such a setting, the consequences of comparable projects can be observed, so that uncertainty is indeed reduced by waiting. However, they are not applicable to fundamental or global changes whose novelty of scope exclude such a "learning by waiting" approach. But with the advent of global environmental changes, such as climate change, the question whether risky experiments should be undertaken has become increasingly relevant.

In this paper, we study this question in a simple and highly stylized setting that is strongly inspired by Arrow and Fisher (1974). We consider a setup where a social planner has to choose between an existing and well-known equilibrium state and a second feasible equilibrium that has unknown properties. To gain information about the second equilibrium, the social planner can initiate a change towards this equilibrium. However, this carries a risk of not being able to return to the original state, if this information indicates that a change to the second equilibrium is detrimental to social welfare.

We show that it will often not be optimal to experiment with such changes or to keep the experiment as small as possible, thereby reducing the costs of returning to the original state. Only if the second equilibrium is expected to have very attractive properties, large-scale experiments are worthwhile. In such cases it will usually be optimal to move directly to the second equilibrium, foregoing any chances of returning.

As we assume risk-neutrality, as in Arrow and Fisher (1974), these results are probably biased towards risking change. Thus our results suggest that experimenting with global changes will often not be a bright idea.

In the following, we briefly discuss how our work relates to existing studies. Then we present our model and deduce its implications. Finally, we relate our results to the well-known (quasi-) option value of avoiding irreversibility.

2 Review of the Literature

Our paper most closely relates to the ideas and results discussed in the classical quasi-option value literature. The origin of this branch is usually associated with Arrow and Fisher (1974). Other prominent authors in the field are Henry (1974) and Dixit and Pindyck (1994). A comprehensive overview of concepts can be found in Mäler and Fisher (2005).

The model presented by Arrow and Fisher (1974) discusses the decision problem of making an investment at the cost of irreversibly giving away another asset. While the current valuations of the asset and the investment are known, the future valuations are uncertain. The authors then show that an agent with linear utility

function, from today's point of view, turns to exhibit "risk-averse" preferences regarding this sort of investment. Hence, the agent shows more conservatism than within an equivalent problem where irreversibility is absent.

Henry (1974) analyzes a similar setting with more than two periods. Analogously to Arrow and Fisher (1974), he shows that if the social planer simply replaces random variables by their expected values in a dynamic optimization problem, she gets more inclined to take irreversible decisions. The reason is that future (optimal) reactions to newly arriving information are not fully accounted for. Both observations are two sides of the same coin and for them the term "irreversibility effect" became prominent in the literature.

The two "classical" contributions mentioned above have a pronounced focus on investment decisions in a public good and environmental context. On the other hand Dixit and Pindyck (1994) focus on investments made in a private sector. They use a continuous time setup, so that their investment problem slightly differs from Arrow and Fisher (1974); Henry (1974) but sticks to the same basic idea.

The "value of waiting", "real option", or "quasi-option" value, which is deduced in all these models, is equivalent to the value of newly arriving information as shown by Conrad (1980). It only exists when flexibility is preserved. Hence it promotes conservatism. As is shown in Mensink and Requate (2003), there are, however, conceptual differences in the definitions of the quasi-option between Arrow, Fisher, Henry and Hanemann (also referring to Hanemann (1989)) and Dixit and Pindyck (1994).

Related to this literature, Epstein (1980) advances a definition of "informativeness". In Epstein's model the agent can wait for a "scientific" signal and will update his beliefs about the true state of the world following Bayesian rules. Epstein then shows that when a certain functional assumption is met and the distribution over all possible scientific signals is regarded as more informative, the agent is more inclined to take decisions today that leave more flexibility tomorrow. Simply put, if more information is expected to be gained through waiting, more waiting is promoted. However, in comparison to Arrow and Fisher (1974) and our work the utility function is assumed to be non-linear. Connecting to Epstein (1980), Jones and Ostroy (1984), Gollier et al. (2000) as well as Salanié and Treich (2009) analyze the relation between expected information and flexibility.

What is common to these studies is that, *ultima ratio*, all new future information will be received independently of the agents decisions. Decisions only influence the way one can react to new information. In that sense there is no need to "try to see" in these models. Our model differs in that the only way to learn is to conduct an experiment. Additionally, the expected amount of newly gained information depends on today's decision.

Furthermore, in the above studies, the degree of irreversibility of a certain action is deterministic. In contrast, our model treats irreversibility as not *a priori*. It sets in as soon as the agent steps on the slippery slope. The probability for this event is endogenous: Being too curious can be dangerous.

Especially in the context of climate change, some authors stress the existence

of "endogenous risk" such as Chichilnisky and Heal (1993); Fisher and Narain (2003). They argue that the probabilities for catastrophic events are often falsely considered as exogenous. Fisher and Narain (2003) provides a model incorporating endogenous catastrophic risks but not endogenous information.

A conceptually different approach is used in Nævdal (2006) and Nævdal and Vislie (2012). Here, the controlling agent is faced with the risk of a catastrophic and irreversible environmental or social change. The change sets in immediately when the amount of an agent-controlled stock crosses an uncertain threshold. In Nævdal and Vislie (2012) a "precautionary tax on current resource extraction" is derived which can be levied upon the stock pollutant generating economic activity. This tax prices the externality generated by nearing the stochastic bound and thereby spurring the risk of a catastrophe and associated social costs. At the same time, it is optimal to build up more reversible capital to insure against possible future losses. These models incorporate high endogeneity with respect to the probability of catastrophic events. But they also do not include the "try to see" component.

From the newer literature the work of Attanasi and Montesano (2011) comes most close to our approach. They analyze a model where development increases the probability for disclosure of the true state of the world. Hence, there both exists exogenous and endogenous information. Therefore, development promises an additional "testing value" to the agent. The agent is assumed to be risk-neutral as in Arrow and Fisher (1974) and our model. One result is that the presence of endogenous information not necessarily spurs development but can also lead to more conservatism. In our paper all information is endogenous. However, our work differs from the contribution of Attanasi and Montesano (2011) by the slippery slope argument.

3 The Model

We use a highly stylized model to describe the decision whether to experiment with change when there is a danger of losing control. Assume that there is a society, represented by a social planner, that lives in a given setting of natural and cultural conditions. This setting is assumed to be an equilibrium, that is, it does not change without action being taken by this society. There is a second possible equilibrium, which the society can reach via taking some action. However, it is unknown whether this second equilibrium is better or worse in terms of achievable well-being than the current one. To get an idea about the quality of the second equilibrium, the society has to move towards this equilibrium. Thereby, it will learn the more about this equilibrium, the closer it moves towards it.

However, once a change towards the second equilibrium is initiated, there is an inherent dynamic that drives the system towards this second equilibrium. If these dynamic forces are not too strong, society can overcome them and return to the original equilibrium. But as the strength of these forces are also unknown in the beginning, there is a risk that once a change is initiated, it will lead to the second

equilibrium regardless of whether this is desired or not.

To keep the model tractable, we use a very simple temporal setup to describe this situation. There are three periods. In period 0, society is in its current equilibrium and decides whether to initiate a change or not. The quality of the existing equilibrium is known, but that of the second equilibrium as well as the strength of the inherent dynamics of the system are unknown. In the second period, society learns the strength of the dynamics and gets some information regarding the quality of the second equilibrium, but only if change has been initiated. Otherwise, no information is gained. In this second period, society also decides whether to move towards the new equilibrium, return to the old one (if this is feasible), or stay in the current disequilibrium (at some costs, as the system has to be kept artificially in this state). In the third period, the consequences of this decision are played out; no further action can be taken.

This is a simple but convenient depiction of the problem of learning on a slippery slope. Information can be gained only by experimenting with change, but such an experiment always includes the danger of running out of control; once initiated, a change might turn out to be both highly undesirable and unstoppable. The setting is a stylized description of current problems of social or environmental change. For example, we do not know whether a world with widely distributed GMOs is better or worse than our present world. To learn this, we have to introduce GMOs into natural systems but this includes the risk that they spread beyond our control. Similarly, we do not know the consequences of climate change for human well-being and cannot learn these consequences without experiencing climate change at least to some extent. However, climate change might be irreversible. Finally, social changes, such as enhanced political and economic integration in Europe, have uncertain consequences that cannot be predicted. However, starting those changes might lead to dynamics (such as change of trade patterns and migration) that cannot be simply reversed.

In line with Arrow and Fisher (1974), we use a very simple formal structure to model and analyze this setting. In particular, we also use a linear model and very simple "dynamics" to highlight the effects of learning by experimenting and of the risk of losing control. Assume that human well-being depends both on an aggregate "state of nature" x and actions taken that change this state h. In each of the three periods, temporal well-being is given by a weighted sum of these variables:

$$w_t = \alpha\, x_t + h_t. \tag{1}$$

Thus we model a risk-neutral society that receives an instantaneous reward from altering its conditions (via h_t) with the consequences being accounted for in a later period. We normalize our setting by assuming $x_0 = 0$. If the society decides not to initiate change, its well-being equals zero in all periods.

The second equilibrium is characterized by

$$x^{eq} = \beta - \eta, \tag{2}$$

with $\beta > 0$ being known and $\eta \geq 0$ being a random variable that is unbounded above and has the expected value $\hat{\eta} > 0$. This describes a situation, where the effects of changing to the new equilibrium have a known upper limit but no lower limit. It captures the idea that we are already in a rather good situation, so that maximal possible losses are far greater than maximal possible gains. Note, however, that we do not impose a probability distribution on η, so that large losses may or may not be rather unlikely.

A change to the second equilibrium is initiated by choosing an $h_0 \in [\varepsilon, 1]$, with $0 < \varepsilon \ll 1$, whereas $h_0 = 0$ implies that the society stays in the current equilibrium. In the latter case, the process is finished; society remains at x_0 for all periods.[1] In contrast, if an $h_0 \geq \varepsilon$ is chosen, the system moves to an intermediate state. The parameter ε describes the minimal change that is necessary to learn something about the second state.

The intermediate state is characterized by

$$x_1 = x_0 + h_0 \ (\beta - \eta_0) \,. \tag{3}$$

This intermediate state thus represents an interpolation between the original and the second equilibrium. The larger the initial change h_0 is, the closer this intermediate state is to the second equilibrium. For $h_0 = 1$, the new equilibrium is reached immediately.

When moving to this intermediate state, society observes the random variable η_0 which provides information both regarding the quality of the second equilibrium and regarding the strength of the inherent dynamics. We assume that η_0 has the same properties as η, in particular, we have $\eta_0 \geq 0$ and η_0 is unbounded above.

However, for $h_0 < 1$, the information regarding the quality of the second equilibrium is imperfect; some uncertainty remains. To model this, we assume that the second equilibrium is characterized by

$$x_2^{eq} = x_0 + \beta - (\eta_0 \, h_0 + \eta_1 \, (1 - h_0)) \,, \tag{4}$$

where η_1 is a random variable with the same distribution as η_0. Thus the η in Eq. (2) is a convex combination of two random variables with the same distribution. The more change society risks, the more it gets to know about the second equilibrium.

But more change implies a higher risk of losing control. We assume that the inherent dynamics are given $\eta_0 \, h_0$, that is, if society wants to stay at x_1, it has to choose $h_1 = -\eta_0 \, h_0$. If it wants to return to the original equilibrium, it has also to undo the original change, that is, to set $h_1 = -(1 + \eta_0) \ h_0$. To model a loss of control, we assume that $h_1 \in [-1, (1 - h_0)]$. Thus society can at most affect an accumulated change of 1 (which suffices to get into the new equilibrium). Furthermore, its ability to reverse the initial change is limited. If η_0 turns out to be

[1] This is a necessary assumption in a model with finite number of periods. It ensures that if a change is initiated, all costs of the change occur within the modeled periods.

large, it can be impossible to return to the initial equilibrium or even to stay at the intermediate state. This models the slippery slope.[2]

To close the model, we need some additional assumptions. First, we assume that if society wants to stay at the intermediate state, it has to set not only $h_1 = -\eta_0\, h_0$ but also $h_2 = -\eta_0\, h_0$. This captures the idea that it is costly to prevent the system from moving according to its inherent dynamics. In our above examples, these costs might represent the costs of limiting the spread of GMOs to a certain number of locations or the costs of climate engineering, that is, of technically removing GHGs from the atmosphere. In contrast, if society moves to either equilibrium, we have $h_2 = 0$.

Second, we assume that whenever $h_1 > -\eta_0\, h_0$, the system reaches the second equilibrium in period 2. This is a strong simplification, as we neglect the duration of the adjustment process. However, all that is lost content-wise is the option to delay an inescapable decline to an undesired state of nature. Similarly, we neglect the option to move the system to a state in-between the original and the intermediate state; if society wants to return to the original state but cannot achieve this, it can only choose between artificially keeping the system in the intermediate state and going forward to the new equilibrium. Again, this is a strong simplification, but it does not remove substantial content from the analysis.[3]

Furthermore, we assume that the social planner uses a discount factor δ when aggregating the expected welfare over periods. We assume $0 < \delta < 1$ and $\alpha\,\delta > 1$. The first condition is standard. The second one excludes cases where a change is solely undertaken because of the direct positive effect of the change on current welfare (h) and not because society sees a chance to switch to a better state. It is thus a helpful condition to focus the model on the essence of the problem.

Finally, we assume that ε is sufficiently small, so that the risk of losing control if negligible, if $h_0 = \varepsilon$ is chosen. This implies that $\text{Prob}(\eta \geq (1-\varepsilon)/\varepsilon) \approx 0$.

Altogether, our model has five parameters: The parameter $\alpha > 0$ describes the welfare attached to the state of nature relative to the immediate benefits of changes. A higher α thus places more weight on the outcome of a change than on the benefits acquired during the transition process. The discount factor δ describes the weight given to future periods. Finally, β and $\hat{\eta}$ capture the current knowledge regarding the quality of the second equilibrium. Note that, by our assumptions, we have $\hat{\eta} = \hat{\eta}_0 = \hat{\eta}_1 > 0$. Finally, ε describes the minimal amount of change that is necessary to gain any information about the quality of the unknown equilibrium.

[2] In addition, we assume that it is impossible to move back further than to the original state.

[3] In fact, if we assume that all $h_1 < -\eta_0\, h_0$ directly lead to the original equilibrium, which would be the other extreme assumption that is feasible in a three-period setup, we get the same qualitative results; only the boundaries reported in the next section shift somewhat.

4 Learning on a Slippery Slope

Although the model appears to be simple, it admits rather complex optimal behavior. The reason is that optimal behavior in period 1 depends both on the information received in period 1 and on what has been done in period 0. In fact, there are five different possible actions in period 1, if change has been initiated: (i) Society might decide to move on to the new equilibrium; (ii) it may stay at the intermediate state; (iii) it may return to the original equilibrium; (iv) it might want to return but cannot; (v) it might want to stay at the intermediate state but cannot. This will lead to three outcomes, namely moving to the second equilibrium, staying at the intermediate state, or returning to the original equilibrium. We will refer to these outcomes as moving on, staying, and returning.

Before we formally analyze the model, it is helpful to discuss some of the incentives for choosing among these different outcomes. It is obvious that if η is small in comparison to β, society will want to move on to the second equilibrium. Thus depending on the relative size of $\hat{\eta}$ and β, we can expect different patterns of optimal behavior.

However, expecting a small η and observing a small η_0 in the first period can have very different effects. If we observe a small η_0 without moving far, this might turn out to be misleading information about the second equilibrium (cf. Eq. (4)). Thus it might be a good idea to stay where we are, despite the inherent costs of staying in a disequilibrium described in the preceding section.

Finally, the observed η_0 provides not only information about the second equilibrium but also about the costs of staying or returning. Thus depending on the h_0 that has been initially chosen, it might be a good idea to respond to a high η_0 by returning (if h_0 is small, so that the costs of returning are small) or to move on to the second equilibrium (for a high h_0, as the costs of returning or staying might be too high).

Due to the action-dependent revelation of information, none of the standard solution concepts of dynamic optimization can be directly applied. Rather, we have to solve the model via backward-induction. We thus first characterize the choices in the first period, taking into account their implications for the second period. Then we combine this information to see under which conditions it is optimal to experiment with change and under which conditions conservatism is rational.

Assume that society has chosen an $h_0 \geq \varepsilon$ and thus observed η_0. Then, from the perspective of and with the information available at period 1, expected discounted welfare (periods 1 and 2) equals

$$\mathcal{E}\left(w_1^{on}\right) = \alpha\,\delta(\beta - \hat{\eta}) + h_0\left(\alpha\left(\beta - (1+\delta)\,\eta_0 + \delta\,\hat{\eta}\right) - 1\right) + 1, \qquad (5)$$

if society moves on;

$$\mathcal{E}\left(w_1^{stay}\right) = h_0\,(1+\delta)\;\left(\alpha\,\beta - (1+\alpha)\,\eta_0\right), \qquad (6)$$

if society stays;

$$\mathcal{E}\left(w_1^{ret}\right) = h_0\ \left(\alpha\,\beta - 1 - (1+\alpha)\,\eta_0\right), \tag{7}$$

if society returns to the original equilibrium.

Comparing these welfare levels shows which behavior is optimal in period 1. Due to the complexity of possible settings, we will do this comparison separately for four intervals of β. We will start out with the case where the new equilibrium is expected to be worse than the current one. All proofs are provided in the appendix.

Lemma 1. *Assume* $0 \leq \beta \leq \hat{\eta} - \frac{1}{\alpha} - \frac{2}{\alpha\,\delta}$. *Under this condition, there are two intervals of* h_0 *that differ regarding the optimal choices in period* 1*:*

(a) For $\varepsilon < h0 \leq \frac{(1+\alpha)\,\delta}{1+\delta\,(1+\alpha\,\beta+\alpha)}$, *it is optimal to*

(1) stay, if $\eta_0 \leq \frac{1+\alpha\,\delta\,\beta}{(1+\alpha)\,\delta}$,

(2) return, if $\frac{1+\alpha\,\delta\,\beta}{(1+\alpha)\,\delta} < \eta_0 \leq \frac{1-h_0}{h_0}$,

(3) stay, if $\frac{1-h_0}{h_0} < \eta_0 \leq \min[\frac{1}{h_0}, \frac{(1+\alpha\,\delta\,(\beta-\hat{\eta}))\,(h_0-1)}{h_0\,(1+\delta)}]$,

(4) and to move on in all other cases.

(b) For $h_0 > \frac{(1+\alpha)\,\delta}{1+\delta\,(1+\alpha\,\beta+\alpha)}$, *it is optimal to*

(1) stay, if $0 < \eta_0 \leq \min[\frac{1}{h_0}, \frac{(1+\alpha\,\delta\,(\beta-\hat{\eta}))\,(h_0-1)}{h_0\,(1+\delta)}]$,

(2) and to move on otherwise.

The setting of Lemma 1 is depicted in Figure 1. The figure shows the two intervals for h_0 and the corresponding regions where it is optimal to move on, stay, or return, as well as the regions where there is a loss of control.

With Lemma 2 we describe optimal period 1 behavior in a world where β is slightly bigger than in the world of Lemma 1. The prospects about the final equilibrium have become a little bit more promising, but still are not exceptionally good.

Lemma 2. *Assume* $0 \leq \hat{\eta} - \frac{1}{\alpha} - \frac{2}{\alpha\,\delta} < \beta \leq \hat{\eta} - \frac{1}{\alpha\,\delta}$ *(or* $\hat{\eta} - \frac{1}{\alpha} - \frac{2}{\alpha\,\delta} \leq 0 \leq \beta \leq \hat{\eta} - \frac{1}{\alpha\,\delta}$*). Under this condition, there are three intervals of* h_0 *that differ regarding the optimal choices in period* 1*:*

(a) For $\varepsilon < h_0 \leq \frac{(1+\alpha)\,\delta\,(1+\alpha\,\delta\,(\beta-\hat{\eta}))}{\alpha\,\delta\,(1+\beta\,(\alpha\,\delta-1)-(1+\alpha)\,\delta\,\hat{\eta})-1}$, *it is optimal to*

(1) stay, if $\eta_0 \leq \frac{1+\alpha\,\delta\,\beta}{(1+\alpha)\,\delta}$,

(2) return, if $\frac{1+\alpha\,\delta\,\beta}{(1+\alpha)\,\delta} < \eta_0 \leq \frac{1-h_0}{h_0}$,

(3) and to move on in all other cases.

(b) For $\frac{(1+\alpha)\,\delta\,(1+\delta\,\alpha\,(\beta-\hat{\eta}))}{\alpha\,\delta\,(1+\beta\,(\alpha\,\delta-1)-(1+\alpha)\,\delta\,\hat{\eta})-1} < h_0 \leq 1 - \frac{1+\alpha\,\delta\,\beta}{\alpha\,\delta\,(1+\hat{\eta})-1}$, *it is optimal to*

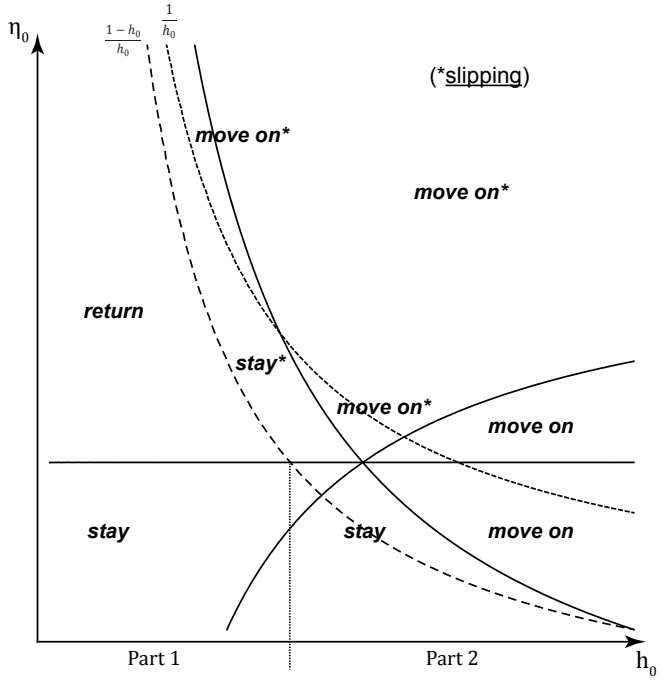


Figure 1: Case $0 \leq \beta \leq \hat{\eta} - \frac{1}{\alpha} - \frac{2}{\alpha\,\delta}$. Regions in which it is optimal to move on, return, or to stay, as well as regions in which control is lost. All regions depicted as a function of h_0 and η_0. Part 1 and 2 refer to the intervals of h_0 identified in Lemma 1. Here, for example, ***stay**** refers to a region where society wants to return, but cannot and, hence, stays.

(1) stay, if $\eta_0 \leq \frac{(1+\alpha\,\delta\,(\beta-\hat{\eta}))\,(h_0-1)}{h_0\,(1+\delta)}$,

(2) move on, if $\frac{(1+\alpha\,\delta\,(\beta-\hat{\eta}))\,(h_0-1)}{h_0\,(1+\delta)} < \eta_0 \leq \frac{1+\alpha\,\delta\,(\beta-\hat{\eta}\,(1-h_0))}{h_0\,(\alpha\,\delta-1)}$,

(3) return, if $\frac{1+\alpha\,\delta\,(\beta-\hat{\eta}\,(1-h_0))}{h_0\,(\alpha\,\delta-1)} < \eta_0 \leq \frac{1-h_0}{h_0}$,

(4) move on, if $\eta_0 > \frac{1-h_0}{h_0}$.

(c) For $h_0 > 1 - \frac{1+\alpha\,\delta\,\beta}{\alpha\,\delta\,(1+\hat{\eta})-1}$, *it is optimal to*

(1) stay, if $\eta_0 \leq \frac{(1+\alpha\,\delta\,(\beta-\hat{\eta}))\,(h_0-1)}{h_0\,(1+\delta)}$,

(2) and to move on otherwise.

The situation of Lemma 2 is depicted in Figure 2. Analogously, the intervals for h_0 given in the figure correspond to the intervals identified by Lemma 2.

The settings of Lemmata 1 and 2 are qualitatively very similar. In the setting of Lemma 1 society regards move on as optimal only in a very small h_0-η_0 region; staying or returning usually are considered as the better options. Therefrom this setting includes the risk of slipping in the sense of wanting but not being able to

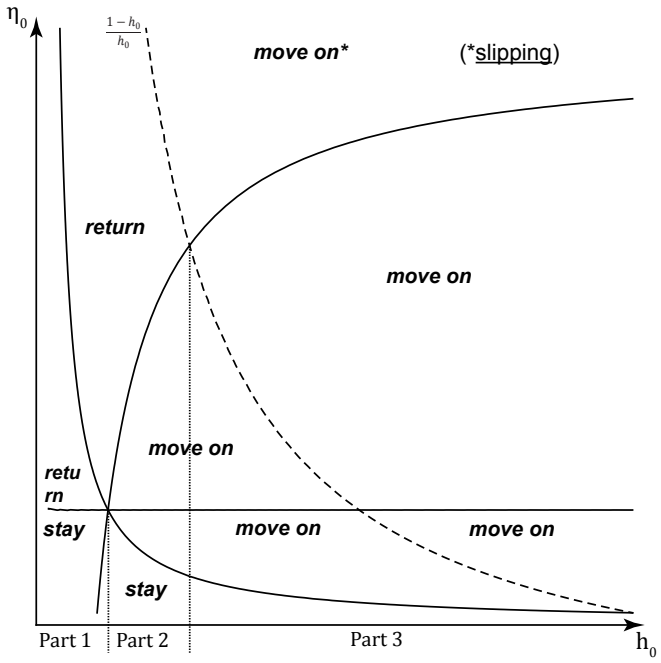


Figure 2: Case $\hat{\eta} - \frac{1}{\alpha} - \frac{2}{\alpha\,\delta} < \beta \leq \hat{\eta} - \frac{1}{\alpha\,\delta}$. Regions in which it is optimal to move on, return, or to stay, as well as regions in which control is lost. All regions depicted as a function of h_0 and η_0. Part 1,2,3 refer to the intervals of h_0 identified in Lemma 2. Here ***move on**** indicates an area where it would be optimal to return but, since this is not possible, move on becomes the preferred action.

stay. Lemma 2 on the other hand describes a world where this risk is not present by the fact that society is rather wants to finish a started experiment instead of staying. Hence the $1/h_0$ bound does not play a role.

Note that it is usually optimal to stay, if the observed η_0 is small. The reason is that in these worlds, where β tends to be small, society expects that another "draw" of the random variable will lead to an outcome that is so much worse than staying that it pays to endure the costs of staying in the intermediate state. However, the larger the initial h_0 has been, the smaller becomes the region in which it is optimal to stay, as the observed η_0 gets more predictive power regarding the quality of the second equilibrium. Thus a large-scale experiment (h_0 close to one), should either be totally abandoned (i.e., returning to the second equilibrium) or followed through completely.

In contrast, a small-scale experiment (h_0 close to zero) will never be finished voluntarily in the settings of Lemma 1 and 2; society will stay or return to the original equilibrium. It will only move on, if it loses control over the experiment. The reason is that a small-scale experiment has not much predictive power.

Next, we consider the case of a slightly larger β and thus a somewhat higher incentive to move to the second equilibrium.

Lemma 3. *Assume* $\hat{\eta} - 1/(\alpha\,\delta) < \beta \leq \hat{\eta} + 1 - 2/(\alpha\,\delta)$. *Now, there are two intervals of* h_0 *that differ regarding the optimal choices in period* 1*:*

(a) For $\varepsilon < h_0 \leq 1 - \frac{1+\alpha\,\delta\,\beta}{\alpha\,\delta\,(1+\hat{\eta})-1}$, *it is optimal to*

(1) move on, if $0 < \eta_0 \leq \frac{1+\alpha\,\delta\,(\beta-\hat{\eta}\,(1-h_0))}{h_0\,(\alpha\,\delta-1)}$,

(2) return, if $\frac{1+\alpha\,\delta\,(\beta-\hat{\eta}\,(1-h_0))}{h_0\,(\alpha\,\delta-1)} < \eta_0 \leq \frac{1-h_0}{h_0}$,

(3) move on, if $\eta_0 > \frac{1-h_0}{h_0}$.

(b) For $h_0 > 1 - \frac{1+\alpha\,\delta\,\beta}{\alpha\,\delta\,(1+\hat{\eta})-1}$, *it is always optimal to move on.*

Thus for larger values of β, there are less different outcomes in period 1. In particular, it is never optimal to stay. Compared to the settings of Lemmata 1 and 2, the second equilibrium is expected to have sufficient quality that, even if a small η_0 has been observed, it is never optimal to bear the costs of staying at the intermediate state. Consequently, the first cases of Lemma 1 or Lemma 2 disappear, as this cases have been based on staying/returning always dominating moving on.

Finally, we examine the case of a large β, which corresponds to a setting where there are strong incentives to move to the second equilibrium.

Lemma 4. *Assume that* $\beta > \hat{\eta} + 1 - 2/(\alpha\,\delta)$. *Then it is always optimal to move on.*

In this case, behavior in the second period is simple. Again, it is never optimal to stay. In addition, for this range of β, society only wants to return, if it cannot do so. Thus it always moves on, either because it wants to or because it has lost control.

Together, these four lemmata fully describe the optimal behavior in period 1. We now use this information to assess what is optimal in period 0, that is, before η_0 has been observed. Should society start an experiment where it might lose control or not? The following proposition partly answers this question.

Proposition 1. *Assume that* $\alpha \leq \frac{1+\hat{\eta}}{\delta\,\hat{\eta}}$. *For the following intervals of* β, *optimal behavior in period* 0 *can be characterized as follows:*

(a) For $\beta \in \left[0, \hat{\eta} - \frac{1}{\alpha\,\delta}\right]$, *it is never optimal to experiment, that is, it is optimal to set* $h_0 = 0$.

(b) For $\beta \in \left[\hat{\eta} - \frac{\delta\,(1-\varepsilon)+\varepsilon}{\alpha\,\delta^2+\alpha\,\delta\,\varepsilon}, \hat{\eta} + \frac{1}{\alpha} - \frac{1}{\alpha\,\delta}\right[$, *it is optimal to conduct an experiment* $\epsilon \leq h_0 \leq 1 - \frac{1+\alpha\,\delta\,\beta}{\alpha\,\delta\,(1+\hat{\eta})-1}$.

(c) For $\beta \geq \max\left[\hat{\eta} + 1 - \frac{2}{\alpha\,\delta}, \hat{\eta} + \frac{1}{\alpha} - \frac{1}{\alpha\,\delta}\right]$, *it is optimal to go directly to the unknown equilibrium, that is, to set* $h_0 = 1$.

It is important to note that these results hold for all distributions of η that conform to our assumptions; we have not used a particular distributional assumption. For this reason, Proposition 1 does not cover all possible cases in terms of β. There is a gap between Part 1 and Part 2 (which vanishes for $\varepsilon \to 0$) and, if $\alpha < 1 + 1/\delta$, there is a gap between Part 2 and Part 3. In these gaps, optimal behavior depends on the particular distribution of η.

When we drop the assumption of $\alpha \leq \frac{1+\hat{\eta}}{\delta\,\hat{\eta}}$, we cannot exclude interior optimal solutions under Part 1 of the proposition, and accordingly results become distribution dependent. A natural question then is: Is the condition $\alpha \leq \frac{1+\hat{\eta}}{\delta\,\hat{\eta}}$ in proposition 1 restrictive? The answer is: It depends. We would naturally assume that $0 < \delta \leq 1$. This says that in the most stringent case – namely when $\delta = 1$ – we restrict to $\alpha \leq 1 + \frac{1}{\hat{\eta}}$. In that case α is still allowed to be bigger than 1. Therefore, if we think that current generations probably are not that much affected by stocks but more by flows, this assumption is not particularly restrictive. But if we think otherwise, proposition 1 may not give us the answer we hoped for.

Furthermore, assume that we place more emphasis on the wellbeing enjoyed in the last period than in the second period. Meaning we introduce a different, smaller discount factor for the second period. Then the needed upper bound on α, to exclude interior solutions, will become more restrictive. Since we want to account for these possibilities will show some numerical examples in the next section.

Even without covering all possible cases, Proposition 1 provides substantial insight into the problem of experimenting with large-scale changes. To interpret this result, it is helpful to put it into perspective by comparing it to alternative settings.

As a first comparison, assume that learning is impossible before the new equilibrium is reached. As the model is linear, society will thus either not initiate change at all or switch completely to the new equilibrium. The latter is optimal if and only if

$$\beta \geq \hat{\eta} - \frac{1}{\alpha\,\delta\,(1+\delta)}. \tag{8}$$

This boundary (from which on change should be initiated in the comparison case) is strictly larger than the upper boundary of Part 1 of Proposition 1 and strictly smaller than the upper boundary of Part 2. Thus compared to a case where learning is impossible, learning on a slippery slope should be initiated in more cases.

Equation (8) is the boundary from which on a full change is better than no change at all, given the information available in period 0. That this boundary lies below the level where a full change is optimal in our setting, highlights the relevance of learning. Even if we expect a full change to be beneficial, it can be optimal to "test and see", that is, to induce a small change and base further change on the observed effects.

This effect goes into a similar direction as the "wait-and-see" effect derived in Arrow and Fisher (1974). It is thus instructive to compare our results to those derived from a setting, where society can learn the complete property of the second

equilibrium by waiting for one period. In this case, an immediate change (i.e., to forego learning by not waiting) is only optimal if

$$1 + \alpha\,\delta\,(1+\delta)\,(\beta - \hat{\eta}) > \alpha\,\delta^2\,\mathcal{E}\left(\max\left[0, \frac{1}{\alpha\,\delta} + \beta - \eta\right]\right). \tag{9}$$

The implied boundary is greater than that in Eq. (8) and (depending on the distribution of η) will often fall into the range of Part 3 of Proposition 1. Thus compared to a "wait-and-see" setting, our setup implies a higher propensity for experimenting with change and, in many cases, even a higher propensity to go directly to the unknown equilibrium.

Thus our result clearly shows the benefits of experimenting with change to gain knowledge. Compared both to a "wait-and-see" setting and a setting without the option of gathering information, change should be initiated more often. Although, its is optimal in some cases to keep change as small as possible while still allowing for learning.

However, our result also highlights that there are many cases, where experimenting with change is not a good idea. Given our assumptions, this is not trivial but rather a somewhat surprising result. We have assumed that the minimal experiment can always be undone and that, due the linearity of our setting, the net costs of revoking a change are rather small, if the change has been small. Furthermore, we have assumed risk neutrality. All these assumptions obviously favor experimenting with change. Indeed, one might expect that in our setting it will usually be optimal to conduct at least limited experiments.

However, this is not the case for two reasons. First, observing the effects of small changes provides only limited predictive power regarding the properties of the unknown equilibrium. If we want to gain substantial knowledge, we have to risk larger changes. But as Lemmata 1 and 2 show, larger changes will not be revoked in many cases, either because the costs are to high or because it turns out that reversing the change is impossible. As society anticipates that larger changes will often not be undone, even if they turn out to have detrimental effects, it will not initiate such changes in many cases.

This captures the idea of a typical slippery slope argument: We should not experiment with substantial changes, as we know that they might not be undone, even if they turn out not to be beneficial. Our results show that this argument has indeed merit. If the expected outcome of changes is not much better than the present state (which would be the case in Part 3 of Prop. 1), we should either not experiment at all or do an experiment that is small enough to avoid the risk of inducing irreversible dynamics.

In fact, the boundary set in Part 3 of Proposition 1 implies that an immediate all-out change is only optimal, if $\beta - \eta > 0$, that is, if the unknown equilibrium is expected to be strictly better than the current one. Given that the change by itself induces a welfare gain (via h_0), this indicates a surprisingly strong tendency to being conservative regarding large-scale changes.

Finally, our results provide one additional insight. Assume that society has started a change although this is not optimal. Then Lemmata 1–4 indicate that it is often not optimal to undo such a change. For instance, assume that actual conditions conform to the setting of Lemma 1 or Lemma 2. From Proposition 1 we know that it is not optimal to start experimenting in this case. However, once the experiment has been started, it will often be optimal either to stay or even to continue to the new equilibrium; only in a rather small set of conditions, it is optimal to undo the initial error.

5 Numerical Analysis

This section discusses some numerical simulations to our slippery slope model. We focus on giving some deeper intuition about the optimal choice of h_0 outside the restrictions of Proposition 1. The main results are presented in figure 3.

In increasing order of β, the figures 3a-3d show optimal period 0 behavior for different values of α and δ. Naturally, the maximum value considered for δ is unity. The simulations all are carried out with η being exponentially distributed and $\hat{\eta} = 1$. Accordingly, in the last figure 3d society takes decisions under the promising condition of $\beta > \hat{\eta}$. We consider the examined range of $0.79 \leq \beta \leq 1.01$ as sufficient to outline all relevant changes to the solution driven by β.

Generally, we present optimal behavior by demarcating the $\delta - \alpha$ plane into different zones where society optimally does not experiment, conducts a minimal experiment, $h_0 = \varepsilon$, conducts a "bigger than minimal" experiment, $\epsilon < h_0 < 1$, or directly goes to the new equilibrium.

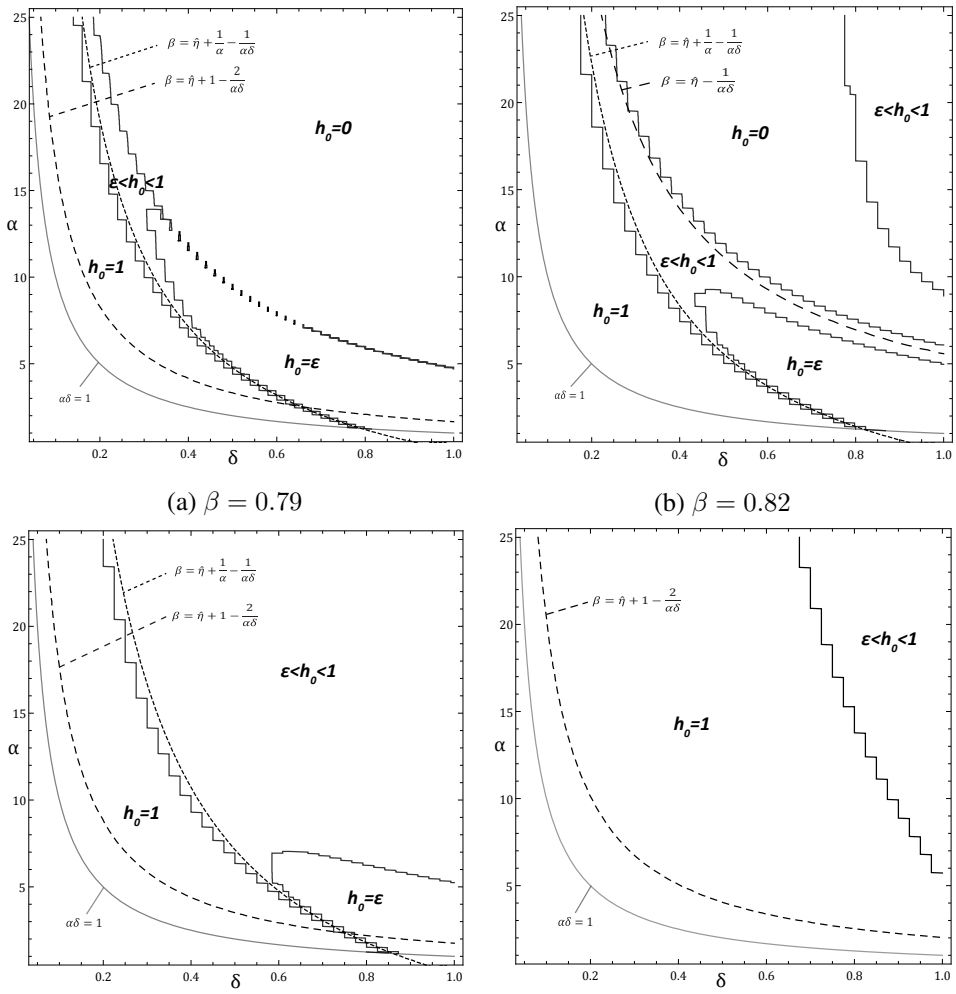
Irrespective of the size of β, we observe that for simultaneously low δ and α society always finds it optimal to directly move to the new equilibrium. In the identified area society obeys either Lemma 3 or 4.

The dashed lines as given in figures 3a and 3c accordingly correspond with part 3 of Proposition 1. For the cases we consider here, inside the zone of indeterminacy between part 2 and part 3 of the Proposition usually it is optimal to set $h_0 = 1$. Otherwise we observe $\epsilon < h_0 < 1$ as optimal.

Only for the cases where $\beta < \hat{\eta}$ the zone where $h_0 = \varepsilon$ is optimal appears. Here society always finds itself in Lemma 3. The zone forms a tongue which will always origin from the south-east corner where δ is big and α is low. The tongue bends into north-west direction and is enclosed by the the $h_0 = 1$ and the $\varepsilon < h_0 < 1$ zones. With increasing β the zone shrinks in size.

When we follow the changes to the solution over figures 3a-3c, we observe that there are two distinct zones wherein $\varepsilon < h_0 < 1$ is optimal. Originally these two zones are separated by a zone where $h_0 = 0$ is optimal. One of the two zones lies far out north-east. The other forms a belt lying closer to the origin and connecting to the $h_0 = 1$ zone. Then with increasing β these two zones merge.

We conjecture that the society's underlying motivation to conduct an experiment $\varepsilon < h_0 < 1$ is different between the two zones. The outer zone fully lies in



(a) $\beta = 0.79$ (b) $\beta = 0.82$

(c) $\beta = 0.86$ (d) $\beta = 1.01$

Figure 3: Numerical results under the exponential distribution $\eta \sim \lambda \exp[-\lambda\eta]$, $\lambda = 1$. The angular shaped lines demarcate different zones where $h_0 = 0$, $h_0 = 1$, $h_0 = \varepsilon$ or $\varepsilon < h_0 < 1$ may be optimal.

a $\delta - \alpha$ area where society's optimal second period behavior is described by Lemmata 1 and 2. The inner zone on the other hand primarily emerges under Lemma 3 and, very thin on its outer bound, under Lemma 2. Hence, if our conjecture is true, even after the fusion of the two zones there must occur a change in the main motivation to experiment, when we move out north-east in $\delta - \alpha$ space.

With $\beta > \hat{\eta}$, the area where minimal $h_0 = \varepsilon$ experiments are optimal disappears. Society will always either want to directly move to the new equilibrium or to conduct an intermediately sized experiment. Again it are Lemmata 1, 2 and 3 which underly the $\varepsilon < h_0 < 1$ zone. The $h_0 = 1$ zone includes cases of Lemmata 3 and 4.

For our interpretation we start with noting that in our model the motivation to set $h_0 > 0$ may come from three different factors. The first incentive stems from the state x_t and sets in when society hopes that the newly reached equilibrium is actually better than the current one. Obviously, this factor also can work on the negative side, when the new equilibrium is expected to be inferior. The second factor is h_t itself and is just the benefit from current actions. This may concern actions undertaken today as well as undertaken by future generations. The third factor is the learning effect. This effect is important if society expects to extract benefit from new knowledge about the true shape of the veiled equilibrium. For this last factor it is particularly important how second period society is expected to react to newly gained information.

Within the different $\delta - \alpha$ zones we identified so far, these three factors take different influence on the decision about h_0.

Start out with the $h_0 = 1$ zone. Here, with low δ and low α, the main incentive must stem from current actions h_t. This is especially true when $\beta < \hat{\eta}$ since not much improvement through x_t is to be expected. Finally, the agent of period 0, since he cares so little for future generations and the state, does not want to procrastinate any enjoyment of welfare from h_t. When $\beta > \hat{\eta}$ the decision to set $h_0 = 1$ is motivated by both h_t and the prospect of reaching a favorable equilibrium.

In the $h_0 = \varepsilon$ zone the interaction of forces is more interesting: Remember here that society will only conduct a minimal experiment under, first, the bad prospect of $\beta < \hat{\eta}$ and, second, Lemma 3. Observe as well that a minimal experiment will not provide any new information to which the next generation could react. In particular, the probability that in the second period society will simply move on and finish the experiment is close to one since, referring to Lemma 3, $\lim_{h_0 \to 0} \frac{1+\alpha\,\delta\,(\beta-\hat{\eta}\,(1-h_0))}{h_0\,(\alpha\,\delta-1)} = \infty$. Hence, we cannot properly term $h_0 = \varepsilon$ an "experiment" and, equivalently, learning cannot be society's objective.

Therefore, we interpret the decision to set $h_0 = \varepsilon$ as a forwarding of enjoyment through h_t to the second generation. The intention to do this emerges from the fact that δ is big and α is low; even though society does not care exceedingly for the state and prioritizes welfare coming from current actions, it does not want expected damage which is running through the state to be carried by two consecutive generations. This is a very non-standard form of showing affection towards future generations. Mostly it is an artifact which stems from the three period world we created.

With increasing β the $h_0 = \varepsilon$ zone accordingly shrinks since the expected damage from the state variable becomes lower. Also, with increasing α it cannot remain optimal to conduct a minimal experiment. The damage that will be, almost surely, carried by the third generation gets more weight with a larger α. Hence, society may want to never experiment at all, meaning also to prevent future generations from acting. Or it may want to conduct an experiment which actually will generate new knowledge to which the second generation can react. For very low β we observe that indeed $h_0 = 0$ becomes optimal with increasing α. For the values

of β presented here, it usually becomes optimal to conduct an intermediately sized experiment.

This last observation directly leads us to an interpretation for the inner belt-like $\varepsilon < h_0 < 1$ zone, as presented in figure 3b, where δ and α are both at an intermediate level: This is the zone where the main incentive to set $h_0 > 0$ stems from enjoyment through h_t in period 1, but where society additionally wants to insure against very bad final equilibria which may finally hit the third generation. Hence, mainly emerging under Lemma 3, the agent, besides the possibility to finish the experiment, wants to bequest knowledge to the second generation so that it could decide to return, if necessary. Note that the low number of points emerging from Lemma 2 may be motivated by the fact that for β very close to $\hat{\eta} - 1/(\alpha\,\delta)$ the behavioral structure of Lemma 2 gets very close to that of Lemma 3.

For even bigger α society would never experiment when β is not to large. The expected damage to third period society, which happens in the case of slipping, deters period 0 society from any experiment.

What then is behind the ominous $\varepsilon < h_0 < 1$ zone lying far north-east where both δ and α are big? We conjecture that here behavior is mainly influenced by considerations about x_t *and learning*. Note that we must be in Lemmata 1 or 2 and, hence, for low realizations of η_0 the second and every following generation will want to stay at an intermediate state. For low h_0 finishing of the experiment mostly will only happen involuntarily through slipping. Now, by the high emphasis put on the state and wellbeing of future generations the consideration becomes to bequeath a "better world". Therefore, by setting $h_0 > 0$ period 0 society gives future generations the possibility to maintain an intermediate state between original and finial equilibrium in the case η_0 should turn out to be low. Note that when $\beta > \hat{\eta}$ we can never be in Lemmata 1 or 2.

Observe that the simulations show that indeed any positive expected benefit must lie in the η_0 range where the second generation decides to either stay or to, in the case of Lemma 2, *voluntarily* move on[4]. Additionally, it can be shown that h_0 will never be set so enter part 2 of Lemma 1 or part 3 of Lemma 2, a result which also can be shown analytically.

The interpretation for the last observation is that only small to medium sized experiments truly can generate useful information. Very large scaled experiments only take away decision power from the second generation, mainly through slipping.

The main result from this section then exactly is the fact that there are two different motives for experimentation in a slippery slope world. One motive is to gain new knowledge, which then at a later stage justifies the conducting of a certain, potentially harmful action.

The other motive is the hope for a "good change" resulting from the experiment.

[4]It is actually easy to show that $\mathcal{E}\left(w_1^{ret}\right)$ is always negative under Lemmata 1 and 2.

6 Conclusions

In this paper, we have analyzed the question whether it is good idea to experiment with a large-scale change of our natural or cultural conditions if, on the one hand, this is the only way to gain knowledge about the effects of such a change, but, on the other hand, we might lose control of the experiment. In contrast to the literature on assessing environmental change under uncertainty, we assume that information is not gained by waiting but only by enacting change. Furthermore, we assume that there is not only the danger of irreversibility (we cannot undo a change) but also the risk of skidding down a slippery slope; we might not be able to stop the change that we have initiated.

Our analysis thus reflects two arguments that often surface in the debate on social or environmental change. First, the exploratory argument that we cannot know whether a change is beneficial unless we have tested it. Second, the slippery slope argument that we might start a process that we cannot or will not reverse, even if it leads to a highly undesirable state.

Our results show that both arguments have merit. Compared to a setting where information cannot be gained and to a setting where information always becomes available over time, it is optimal to experiment with change in more cases in our setup. This highlights the benefits of experimenting. However, despite using assumptions that strongly favor experiments, our setting yields an optimal solution that is surprisingly conservative in prescribing no or very limited experiments under many conditions. The reason is that minor changes are often not worthwhile (as they do not yield much information), but major changes will not be undone in many cases, even if they turn out to be devastating (either because undoing them is too costly or because it is infeasible). Anticipating this, it is often better not to initiate a change at all. This is indeed a typical slippery slope argument.

Although our setting is highly stylized, it captures the essential aspects of many current environmental problems. Should we experiment with climate change (as we are currently doing)? Should we introduce GMOs into the natural environment? In both cases, we cannot know the costs or benefits of such a change without experiencing at least some of the change. Furthermore, in both cases there is a clear risk that a change cannot be stopped once it has been initiated.

Our results indicate that an assessment of such policies based on expected values is misleading, as it does not take into account the option to alter decisions once information is gained. It will therefore recommend a full change too often. However, an analysis in the tradition of the option value literature will be too conservative. It assumes that we will get wise by waiting and thus neglects the benefits of experimenting with change.

However, our results are based on very simple model and can thus only indicate which arguments might have merit but cannot assess the relative strengths of these arguments in specific settings. An adequate assessment of these arguments in the context of specific applications will require a much more detailed model of the change, its dynamics and effects, and more specific assumptions on the distri-

bution of the unknown parameters. Given that even our simple setting has resulted in a rather complex set of possible outcomes, such a detailed assessment will almost surely require a numerical approach. Our results simply show that capturing both the idea of gaining information via experimenting with change and the risk of losing control over such an experiment might be interesting and might lead to a substantial shift in conclusions.

A Proof of Lemma 1

First assume away the risk of slipping. Comparing the different expressions for expected welfare from Eqs. (5)–(7) pairwise shows that, under the parameter constraints set in the lemma, a small value of η_0 favors staying, whereas a high value of η_0 favors returning. Moving on can be optimal in between:

We have the boundaries

- $\mathcal{E}\left(w_1^{on}\right) = \mathcal{E}\left(w_1^{stay}\right)$: $\eta^{o/s}(h_0) = \frac{(1+\alpha\,\delta\,(\beta-\hat{\eta}))\,(h_0-1)}{h_0\,(1+\delta)}$.
- $\mathcal{E}\left(w_1^{on}\right) = \mathcal{E}\left(w_1^{return}\right)$: $\eta^{o/r}(h_0) = \frac{1+\alpha\,\delta\,(\beta-\hat{\eta}\,(1-h_0))}{h_0\,(\alpha\,\delta-1)}$.
- $\mathcal{E}\left(w_1^{stay}\right) = \mathcal{E}\left(w_1^{return}\right)$: $\eta^{s/r}(h_0) = \frac{1+\alpha\,\delta\,\beta}{(1+\alpha)\,\delta}$ $(= const.)$.

Given the assumption $\alpha\,\delta \geq 1$,

- for $\eta_0 \leq \eta^{o/s}(h_0)$: $\mathcal{E}\left(w_1^{stay}\right) \geq \mathcal{E}\left(w_1^{on}\right)$,
- for $\eta_0 \leq \eta^{o/r}(h_0)$: $\mathcal{E}\left(w_1^{on}\right) \geq \mathcal{E}\left(w_1^{return}\right)$ and
- for $\eta_0 \leq \eta^{s/r}(h_0)$: $\mathcal{E}\left(w_1^{stay}\right) \geq \mathcal{E}\left(w_1^{return}\right)$.

Accordingly, at the boundaries the domination relationships switch.

The three boundaries will all meet at $h_0 = \frac{(1+\alpha)\,\delta\,(1+\alpha\,\delta\,(\beta-\hat{\eta}))}{\alpha\,\delta\,(1+\beta\,(\alpha\,\delta-1)-(1+\alpha)\,\delta\,\hat{\eta})-1} \equiv \tilde{h}$. Given the assumptions of Lemma 1, $0 < \tilde{h} < 1$.

For $h_0 < \tilde{h}$: We have that $\eta^{o/r}(h_0) < \eta^{s/r}(h_0) < \eta^{o/s}(h_0)$ and moving on is always dominated either by staying or returning. Optimal actions are staying for $0 \leq \eta_0 < \eta^{s/r}(h_0)$ and returning for $\eta^{s/r}(h_0) \leq \eta_0$.

For $\tilde{h} \leq h_0$: We will have $\eta^{o/s}(h_0) < \eta^{s/r}(h_0) < \eta^{o/r}(h_0)$ and the intermediate region between $\eta^{o/s}(h_0)$ and $\eta^{o/r}(h_0)$ where moving on dominating staying and returning appears. Accordingly, optimal actions are staying for $0 \leq \eta_0 < \eta^{o/s}(h_0)$, moving on for $\eta^{o/s}(h_0) \leq \eta_0 < \eta^{o/r}(h_0)$ and returning for $\eta^{o/r}(h_0) \leq \eta_0$.

We bring now back the risk of slipping: For $\frac{1-h_0}{h_0} < \eta_0 \leq \frac{1}{h_0}$ staying and moving on is possible but returning is not. For $\frac{1}{h_0} < \eta_0$ only moving on is possible. These restrictions may interfere with the optimal – call it "first best" – actions identified so far and force the agent to choose "second best" or "third best".

The boundary $\eta^{o/s}(h_0)$ crosses $\frac{1-h_0}{h_0}$ at $h_0 = 1$. Given the assumptions of the Lemma $\eta^{o/s}(h_0)$ will always lie above $\frac{1-h_0}{h_0}$ for any $h_0 < 1$.

Furthermore, there is an intersection between the boundary $\frac{1-h_0}{h_0}$ and the boundary $\eta^{s/r}(h_0)$ at $\check{h} = \frac{(1+\alpha)\delta}{1+\delta(1+\alpha\beta+\alpha)}$, with $0 < \check{h} < \tilde{h} < 1$. For $\check{h} < h_0$ we have that $\frac{1-h_0}{h_0}$ must lie below $\eta^{s/r}(h_0)$ and hence every time the agent sees returning as "first best" he will be forced to choose some other action. On the other hand, when $h_0 \leq \check{h}$ there is a positive region where the agent wants to return and is able to do so.

Assuming $h_0 \leq \check{h}$: For values of η_0 above $\frac{1-h_0}{h_0}$ the agent will optimally stay — in the "second best" sense — as long as η_0 is still below $\eta^{o/s}(h_0)$. But he can only do so if also $\eta_0 \leq \frac{1}{h_0}$. For any bigger value of η_0 only the option to move on remains. This proves part 1 of the Lemma.

In the case of $\check{h} < h_0 \leq \tilde{h}$: For values of η_0 above $\eta^{s/r}(h_0)$ returning is not possible and, again, the agent will take the "second best" action staying if η_0 is below $\eta^{o/s}(h_0)$ and simultaneously below $\frac{1}{h_0}$. Correspondingly, for any bigger value of η_0 it must be optimal to move on.

In the case $\tilde{h} < h_0$: For any η_0 which lies above the boundary $\eta^{o/r}(h_0)$ the option to return has already died because $\eta_0 > \frac{1-h_0}{h_0}$. Hence the agent will optimally stay if η_0 is simultaneously smaller than the boundary $\eta^{o/s}(h_0)$ and $\frac{1}{h_0}$ and for any η_0 value above moving on must be optimal.

The last two observations prove part 2 of the Lemma. □

B Proof of Lemma 2

Proceeds in the same fashion as the proof of Lemma 1. The major difference in the setting is that the boundary $\eta^{o/s}(h_0)$ will always lie below $\frac{1-h_0}{h_0}$ for $h_0 < 1$ and, hence, it will always lie below $\frac{1}{h_0}$. Therefore the event of not being able to stay is not important. □

C Proof of Lemma 3

The proof proceeds in the same way as that of Lemma 1. However for $\hat{\eta} - 1/(\alpha\delta) < \beta \leq \hat{\eta} + 1 - 2/(\alpha\,\delta)$, the first case of Lemma 1 ceases to exist (there is always some value of η_0 for which it is optimal to move on) and the option to stay is always dominated by the option to move on. □

D Proof of Lemma 4

Again, the proof proceeds in the same way as that of Lemma 1. However now, the first and second case of Lemma 1 cease to exist and it is never optimal to stay. □

E Proof of Proposition 1

We first have to calculate expected welfare from the perspective of period 0, taking into account the optimal response derived above to the observation of η_0. What we get is an extensive list:

(a) For the cases where it is always optimal to move on (Lemma 4 and Lemma 3, Part 2) we get

$$\mathcal{E}\left(w_0^a\right) = \alpha\,\delta^2\left(\beta-\hat{\eta}\right)+\delta+h_0\left(\delta\left(\alpha\left(\beta-\hat{\eta}\right)-1\right)+1\right).$$

(b) Lemma 3, Part 1:

$$\begin{aligned}\mathcal{E}\left(w_0^b\right) \;=\; & h_0+\delta\left(1+h_0\left(\alpha\left(\beta-\hat{\eta}\right)-1\right)+\alpha\,\delta\left(\beta-\hat{\eta}\right)\right.\\ & +h_0\left(\alpha\,\delta-1\right)\mathcal{E}\left(\max\left[\eta_0-\frac{1+\alpha\,\delta\left(\beta-\hat{\eta}\left(1-h_0\right)\right)}{h_0\left(\alpha\,\delta-1\right)},0\right]\right)\\ & +\mathcal{E}^*\left(\frac{h_0\left(1+h_0\,\eta_0+\alpha\,\delta\left(\beta-h_0\,\eta_0+\left(h_0-1\right)\hat{\eta}\right)\right)}{h_0\left(1+\eta_0\right)-1}\right.\\ & \left.\left.\cdot\max\left[\eta_0-\frac{1+h_0}{h_0},0\right]\right)\right)\end{aligned}$$

(c) Lemma 2, Part 3, and Lemma 1, Part 2 (when $\min\left[\frac{1}{h_0},\frac{\left(1+\alpha\,\delta\left(\beta-\hat{\eta}\right)\right)\left(h_0-1\right)}{h_0\left(1+\delta\right)}\right]=\frac{\left(1+\alpha\,\delta\left(\beta-\hat{\eta}\right)\right)\left(h_0-1\right)}{h_0\left(1+\delta\right)}$):

$$\begin{aligned}\mathcal{E}\left(w_0^c\right) \;=\; & h_0+\delta\left(h_0\left(\alpha\,\beta-\left(1+\alpha\right)\hat{\eta}\right)\left(1+\delta\right)\right.\\ & \left.+h_0\left(1+\delta\right)\mathcal{E}\left(\max\left[\eta_0-\frac{\left(1+\alpha\,\delta\left(\beta-\hat{\eta}\right)\right)\left(h_0-1\right)}{h_0\left(1+\delta\right)},0\right]\right)\right)\end{aligned}$$

(d) Lemma 1, Part 2 (when $\min\left[\frac{1}{h_0},\frac{\left(1+\alpha\,\delta\left(\beta-\hat{\eta}\right)\right)\left(h_0-1\right)}{h_0\left(1+\delta\right)}\right]=\frac{1}{h_0}$):

$$\begin{aligned}\mathcal{E}\left(w_0^d\right) \;=\; & h_0+\delta\left(h_0\left(\alpha\,\beta-\left(1+\alpha\right)\hat{\eta}\right)\left(1+\delta\right)\right.\\ & +\mathcal{E}^*\left(\frac{h_0^2\left(1+\delta\right)}{1-h_0\,\eta_0}\right.\\ & \left.\cdot\max\left[\eta_0-\frac{1}{h_0},0\right]\cdot\max\left[\frac{\left(1+\alpha\,\delta\left(\beta-\hat{\eta}\right)\right)\left(h_0-1\right)}{h_0\left(1+\delta\right)}-\eta_0,0\right]\right)\\ & \left.+h_0\left(1+\delta\right)\mathcal{E}\left(\max\left[\eta_0-\frac{\left(1+\alpha\,\delta\left(\beta-\hat{\eta}\right)\right)\left(h_0-1\right)}{h_0\left(1+\delta\right)},0\right]\right)\right)\end{aligned}$$

(e) Lemma 2, Part 2:

$$\begin{aligned}\mathcal{E}\left(w_0^e\right) \;=\; & h_0 + \delta\left(h_0\left(\alpha\,\beta - (1+\alpha)\,\hat{\eta}\right)(1+\delta)\right.\\ & + h_0\,(1+\delta)\,\mathcal{E}\left(\max\left[\eta_0 - \frac{(1+\alpha\,\delta\,(\beta-\hat{\eta}))\,(h_0-1)}{h_0\,(1+\delta)}, 0\right]\right)\\ & + h_0\,(\alpha\,\delta - 1)\,\mathcal{E}\left(\max\left[\eta_0 - \frac{1+\alpha\,\delta\;(\beta-\hat{\eta}\,(1-h_0))}{h_0\,(\alpha\,\delta-1)}, 0\right]\right)\\ & + \mathcal{E}^*\left(\frac{h_0\,(1+h_0\,\eta_0+\alpha\,\delta\,(\beta-h_0\,\eta_0+(h_0-1)\,\hat{\eta}))}{h_0(1+\eta_0)-1}\right.\\ & \left.\left.\cdot\max\left[\eta_0 - \frac{1+h_0}{h_0}, 0\right]\right)\right)\end{aligned}$$

(f) Lemma 2, Part 1:

$$\begin{aligned}\mathcal{E}\left(w_0^f\right) \;=\; & h_0 + \delta\left(h_0\left(\alpha\,\beta - 1 - (1+\alpha)\,\hat{\eta}\right)\right.\\ & + h_0\,\delta\,(1+\alpha)\,\mathcal{E}\left(\max\left[\frac{1+\alpha\,\delta\,\beta}{(1+\alpha)\,\delta} - \eta_0, 0\right]\right)\\ & + \mathcal{E}^*\left(\frac{h_0\,(1+h_0\,\eta_0+\alpha\,\delta\,(\beta-h_0\,\eta_0+(h_0-1)\,\hat{\eta}))}{h_0\,(1+\eta_0)-1}\right.\\ & \left.\left.\cdot\max\left[\eta_0 - \frac{1+h_0}{h_0}, 0\right]\right)\right)\end{aligned}$$

(g) Lemma 1, Part1 ($\min\left[\frac{1}{h_0}, \frac{(1+\alpha\,\delta\,(\beta-\hat{\eta}))\,(h_0-1)}{h_0\,(1+\delta)}\right] = \frac{(1+\alpha\,\delta\,(\beta-\hat{\eta}))\,(h_0-1)}{h_0\,(1+\delta)}$):

$$\begin{aligned}\mathcal{E}\left(w_0^h\right) \;=\; & h_0 + \delta\left(h_0(\alpha\,\beta - 1 - (1+\alpha)\,\hat{\eta})\right.\\ & + h_0\,\delta\,(1+\alpha)\,\mathcal{E}\left(\max\left[\frac{1+\alpha\,\delta\,\beta}{(1+\alpha)\delta} - \eta_0, 0\right]\right)\\ & + \mathcal{E}^*\left(\frac{h_0^2\,(1+\delta\,(\alpha\,\beta-(1+\alpha)\,\eta_0))}{h_0\,(1+\eta_0)-1}\max\left[\eta_0 - \frac{1-h_0}{h_0}, 0\right]\right.\\ & \left.\left.+ h_0\,(1+\delta)\;\max\left[\eta_0 - \frac{(1+\alpha\,\delta\,(\beta-\hat{\eta}))\,(h_0-1)}{h_0\,(1+\delta)}, 0\right]\right)\right)\end{aligned}$$

(h) Lemma 1, Part 1 ($\min\left[\frac{1}{h_0}, \frac{(1+\alpha\,\delta\,(\beta-\hat{\eta}))\,(h_0-1)}{h_0\,(1+\delta)}\right] = \frac{1}{h_0}$):

$$\begin{aligned}\mathcal{E}\left(w_0^g\right) \;=\; & h_0 + \delta\left(h_0\left(\alpha\,\beta - 1 - (1+\alpha)\,\hat{\eta}\right)\right. \\ & + h_0\,\delta(1+\alpha)\mathcal{E}\left(\max\left[\frac{1+\alpha\,\delta\,\beta}{(1+\alpha)\delta} - \eta_0, 0\right]\right) \\ & + \mathcal{E}^*\left(\frac{h_0^2\left(1+\delta\left(\alpha\,\beta-(1+\alpha)\,\eta_0\right)\right)}{h_0\,(1+\eta_0)-1}\max\left[\eta_0 - \frac{1-h_0}{h_0}, 0\right]\right) \\ & + \mathcal{E}^*\left(\frac{h_0\left(1+\alpha\,\delta\,(\beta-\hat{\eta}) + h_0\left(\eta_0 - 1 + \delta\left(\eta_0 + \alpha\,(\hat{\eta}-\beta)\right)\right)\right)}{h_0\,\eta_0 - 1}\right. \\ & \left.\left.\cdot\max\left[\eta_0 - \frac{1}{h_0}, 0\right]\right)\right)\end{aligned}$$

Observe that above all expectation operators work on at least one maximum operator. Expected value terms which are associated with η_0-regions where society slips are indicated with an asterisk. These terms must be negative.

E.1 Part 1 of Proposition 1

We start with Part 1 of the proposition and consider the case $\beta \leq \hat{\eta} - 1/(\alpha\,\delta)$. We want to show that, given the parameter restrictions, society never experiments. Since we are either in Lemma 1 or Lemma 2 we accordingly need to examine the corresponding expected welfare functions. This is (c)-(h) above.

The procedure of the proof is the following: Construct upper bounds for the expected welfare functions (c)-(h) by, (i), replacing all expectation operators which refer to slipping with zero. (ii), replace the expectation operator $\mathcal{E}\left(\max\left[\frac{1+\alpha\,\delta\,\beta}{(1+\alpha)\,\delta} - \eta_0, 0\right]\right)$ with $\frac{1+\alpha\,\delta\,\beta}{(1+\alpha)\,\delta}$ and, (iii), replace any other expectation operator with $\hat{\eta}$.

These upper bounds are all linear in h_0. It is then easy to show that under the assumptions made they induce $h_0 = 0$ as optimal.

E.2 Part 2 of Proposition 1

Part 2 of the Proposition concerns either Lemma 3 or Lemma 4.

Consider Lemma 3, Part 1. A lower bound to $\mathcal{E}\left(w_0^b\right)$ for arbitrary values of h_0 is given by $\alpha\,\delta^2\,(\beta-\hat{\eta}) + \delta + h_0\,(\delta\,(\alpha\,(\beta-\hat{\eta})-1)+1) = \mathcal{E}\left(w_0^a\right)$. This is true since the agent, as a baseline option, could always move on in period 1. But then he has the possibility of optimize welfare by choosing to return for some intermediate η_0 values.

Using this lower bound, if we set $h_0 = \varepsilon$ for Lemma 3, Part1 this yields a minimal expected welfare of $\alpha\,\delta^2\,(\beta-\hat{\eta}) + \varepsilon\,(\delta\,(\alpha\,(\beta-\hat{\eta})-1)+1) + \delta$. For $\beta \geq \hat{\eta} - \frac{\delta\,(1-\varepsilon)+\varepsilon}{\alpha\,\delta^2+\alpha\,\delta\,\varepsilon}$, this expression is positive. Hence, in this case, it is better to

conduct a minimal experiment than to not experiment at all. This, obviously, must also be true for Lemma 4.

Observe that for $\beta < \hat{\eta} + \frac{1}{\alpha} - \frac{1}{\alpha\,\delta}$ the value function $\mathcal{E}\left(w_0^a\right)$ is declining in h_0.

On the other hand an upper bound for $\mathcal{E}\left(w_0^b\right)$ is given by setting the slip term to zero and replace the other expectation operator with $\hat{\eta}$. For the given parameter restrictions we cannot exclude that this bound is an increasing function of h_0.

Hence, interior solutions may be optimal under Lemma 3, Part 1. But by the fact that, as shown, $\mathcal{E}\left(w_0^a\right)$ is declining in h_0 we can exclude that it is optimal to enter Part 2 of Lemma 3.

E.3 Part 3 of Proposition 1

Obvious. As already shown, the derivative of $\mathcal{E}\left(w_0^a\right)$ with respect to h_0 must be nonnegative when $\beta \geq \hat{\eta} + \frac{1}{\alpha} - \frac{1}{\alpha\,\delta}$. By $\beta \geq \max\left[\hat{\eta} + 1 - \frac{2}{\alpha\,\delta}, \hat{\eta} + \frac{1}{\alpha} - \frac{1}{\alpha\,\delta}\right]$ it is assured that we are in Lemma 4 (observe that under Lemma 3 an interior solution under Part 1 of the Lemma may be optimal even for $\beta \geq \hat{\eta} + \frac{1}{\alpha} - \frac{1}{\alpha\,\delta}$). □

References

Arrow, K. and A. Fisher (1974). Environmental preservation, uncertainty, and irreversibility. *The Quarterly Journal of Economics 88*(2), 312–319.

Attanasi, G. and A. Montesano (2011). The value of endogenous information above exogenous information in irreversible environmental decisions. *Working Paper*, 1–39.

Chichilnisky, G. and G. Heal (1993). Global environmental risks. *Journal of Economic Perspectives 7*(4), 65–86.

Conrad, J. (1980). Quasi-option value and the expected value of information. *The Quarterly Journal of Economics 94*(4), 813–820.

Dixit, A. and R. Pindyck (1994). *Investment under Uncertainty* (1 ed.). Princeton University Press.

Epstein, L. (1980). Decision making and the temporal resolution of uncertainty. *International Economic Review 21*(2), 269–283.

Fisher, A. and U. Narain (2003). Global warming, endogenous risk, and irreversibility. *Environmental and Resource Economics 25*(4), 395–416.

Gollier, C., B. Jullien, and N. Treich (2000). Scientific progress and irreversibility: An economic interpretation of the precautionary principle. *Journal of Public Economics 75*(2), 229–253.

Hanemann, W. (1989). Information and the concept of option value. *Journal of Evironmental Economics and Management 16*(1), 23–37.

Henry, C. (1974). Investment decisions under uncertainty: The "irreversibility effect". *The American Economic Review 64*(6), 1006–1012.

Jones, R. and J. Ostroy (1984). Flexibility and uncertainty. *Review of Economic Studies 51*(1), 13–32.

Mensink, P. and T. Requate (2003). The dixit-pindyck and the arrow-fisher-hanemann-henry option values are not equivalent. *Christian-Albrecht-University of Kiel Economic Working Papers* , 1–13.

Mäler, K. and A. Fisher (2005). Environment, uncertainty, and option values. In *Handbook of Environmental Economics*, Volume 2, pp. 571–620. Elsevier.

Nævdal, E. (2006). Dynamic optimisation in the presence of threshold effects when the location of the threshold is uncertain – with an application to a possible disintigration of the western antarctic ice sheet. *Journal of Economic Dynamics and Control 30*(7), 1131–1158.

Nævdal, E. and J. Vislie (2012). Resource depletion and capital accumulation under catastrophic risk: The role of stochastic thresholds and stock pollution. *Oslo University, Department of Economics in its series "Memorandum"* (24), 1–31.

Salanié, F. and N. Treich (2009). Option value and flexibility: A general theorem with applications. *Toulouse School of Economics Working Paper Series* (09-002), 1–26.