Theory Dec. (2011) 70:195–213
DOI 10.1007/s11238-010-9235-2

Beneficial safety decreases

Till Grüne-Yanoff · Holger Rosencrantz

Published online: 3 November 2010

Abstract We construct a model of rational choice under risk with biased risk judgement. On its basis, we argue that sometimes, a regulator aiming at maximising social welfare should affect the environment in such a way that it becomes 'less safe' in common perception. More specifically, we introduce a bias into each agent's choice of optimal risk levels: consequently, in certain environments, agents choose a behaviour that realises higher risks than intended. Individuals incur a welfare loss through this bias. We show that by deteriorating the environment, the regulator can motivate individuals to choose behaviour that is less biased, and hence realises risk levels closer to what individuals intended. We formally investigate the conditions under which such a Beneficial Safety Decrease—i.e. a deteriorating intervention that has a positive welfare effect—exists. Finally, we discuss three applications of our model.

Keywords Philosophy of risk · Rational choice · Risk offsetting behaviour · Judgement bias · Safety regulation · Risk policy · Paternalism

1 Introduction

Risk is commonly conceived of in relation to the outcome of events. However, risk-reducing regulation is often directed at objects or contexts related to such events, not the outcome itself. Workplace safety regulations, for example, often target properties of machines; transport safety the qualities of cars and roads. In some instances,

T. Grüne-Yanoff
Helsinki Collegium of Advanced Studies, Helsinki, Finland
e-mail: till.grune@helsinki.fi

H. Rosencrantz (✉)
Royal Institute of Technology, Stockholm, Sweden
e-mail: hrz@kth.se

there is a direct relation between these material properties and the outcomes of events involving them.[1] Yet when the event in question concerns human action, regulations of these material properties may not have the intended effect. This is because people often adjust their choices to their perception of safety-related properties. For example, people work differently with kitchen knives they perceive as very sharp, they hike differently on trails they see as steep and slippery, and they drive differently on roads and in conditions they deem difficult.

This poses a problem to the paternalistic regulator. Paternalistic regulators seek to reduce the risk people impose on themselves. Yet when they do so by regulating *safety features*, it is not obvious what effects these regulations will have on the outcome of their actions.

Many take it as intuitively obvious that increasing the safety features of objects and contexts leads to an improvement of the outcome. In contrast, we argue that sometimes, the increase of safety features yields a worse outcome, and conversely the *deterioration* of safety features leads to an improvement. We call such counterintuitive situations Beneficial Safety Decreases, and show that they can obtain under plausible conditions.

After reviewing the relevant literature, our argument starts with a simple risk offsetting model of how individuals choose their optimal risk exposure (Sect. 3). We then propose that individuals are biased in the way they perceive risks (Sect. 4). This bias, we argue, sometimes prevents individuals from realising their optimal risk exposure in their actions. However, this bias is not constant across all risks. Rather, it is large for mid-level risks, and small for low- and high-level risks. By getting individuals to choose risk levels that are less prone to this bias, the policy maker can improve their lot. However, he cannot do so directly: individuals choose risk levels they think are optimal for the environment they perceive. Instead, we argue in Sect. 5, the policy maker can influence individuals' environment in order to make them choose actions closer to their optimal risk exposure. Decreasing safety features in this way can, therefore, yield a beneficial effect. Making use of our formal model, we determine the necessary and sufficient conditions for such beneficial interventions; in addition, we offer a proof by construction for the existence of such an effect in Sect. 6. In Sect. 7 we discuss three applications of our model. Section 8 concludes.

2 Literature review

Based on the ideas by Lave and Weber (1970), and more fully developed by Peltzman (1975), a consumer theoretic literature models the effect of safety regulations on individuals' risk-taking behaviour. In Peltzman's model, the car driver chooses time spent on a given ride and resources spent on safety precautions in order to maximise expected income for a given milage. This optimisation is constrained by negative relations between (i) time spent and accident probability, and (ii) safety precautions and loss from an accident. Peltzman shows that an exogenous rise in the safety level of the individual's environment has two opposing effects. First, it reduces the loss from

[1] To reduce the risk of fire, for example, the load of an electrical grid must be controlled; to reduce earthquake risks, houses must be constructed in certain ways.

accidents; but second, it induces individuals to reduce driving time and hence increase the probability of an accident. Whether the overall effect of safety regulations is beneficial thus depends on the magnitude of these opposing effects. Peltzman conducted an empirical investigation and concluded that in the case of the National Highway and Motor Vehicle Safety Act of 1966, the consequent decrease in risk of death by accident was completely offset by drivers taking greater risk.

A large body of empirical literature has followed Peltzman's analysis. Most of these studies investigate offsetting as a correlation of safety standards and highway fatalities, using highly aggregated data. Some articles, however, seek to estimate relationships between safety standards and driver behaviour (Traynor 1993; Evans and Graham 1991; Winston et al. 2006). These studies tend to find strong evidence in support of the offsetting effect. Behavioural studies support this result further. For example, cars outfitted with antilock brakes are driven faster, more carelessly, and closer to the car in front, braked more abruptly, and have no lower accident rate per hour of exposure than cars without these devices (Sagberg et al. 1997). Wider road design has been shown to increase people's speed and lane position, offsetting potential safety effects from changing road width (van Driel et al. 2004).

Yet beyond this offsetting effect, it is sometimes observed that safety increases have led to an *increase* in accident rates. For example, in a large-scale experiment experimental subjects were given either (i) extensive, state-of-the-art driving training, or (ii) a minimal course providing the skills necessary to pass the driving test, or (iii) no training, it being assumed that they were trained by their parents. Members of group (i) obtained their driver's licences sooner and had significantly *more* crashes than those who had received minimal training or no high school driver training at all (Lund et al. 1986).

Similar effects have been suggested with respect to skill improvement in children. A Swedish study showed that the more traffic safety education children in kindergarten and primary school had received, the higher their traffic injury rate (Johansson 1997).

Related observations have been made with respect to other domains of risk prevention. For example, the U.S Food and Drug Administration in 1972 mandated the introduction of so-called 'child-proof' safety caps. Their introduction was followed by a substantial increase in the per capita rate of fatal accidental poisonings in children. It was concluded that the impact of the regulation was counterproductive, 'leading to 3,500 additional (fatal plus non-fatal) poisonings of children under age 5 annually from analgesics' (Viscusi 1984, p. 327).

Within a formal model very close to Peltzman's original approach, Viscusi (1984) argues that such a phenomenon is possible only if the regulated indifference curve at the same level of expected loss is flatter than the unregulated one. He does not further analyse such a case, suggesting only that it is 'difficult to meet these requirements', but that it is conceivable, 'if individuals do not perceive accurately the accident probabilities' (Viscusi 1984, p.325).

Arnould and Grabowski (1981) introduced perception biases in models of safety regulation. They argue that the low utilisation of passive seatbelts is likely caused by people's underestimation of low probabilities, leading to non-optimal risk choices. However, they explicitly disregard any offsetting effects in their models. Salanié and Treich (2009) model individuals making faulty choices due to their biased beliefs,

and suggest that a regulator should take these biases into account when setting safety standards. However, in their model the regulator sets safety levels in order to influence people into choosing alternatives that are paternalistically determined as best; the offsetting effect in people's optimisation is disregarded.

Paternalistic regulation as a means to benefit irrational individuals has been studied in some relatively recent theoretical articles. For example, Camerer et al. (2003, p. 1212) use the term 'asymmetric paternalism' for regulations that create 'large benefits for those who make errors, while imposing little or no harm on those who are fully rational'. The case for paternalistic regulation is also explored by O'Donoghue and Rabin (2003), who note that such policies should pay attention to the fact that individuals may commit errors in many different ways. While a paternalistic policy may help some irrational individuals, it may hurt other individuals who commit different errors, while being harmless to fully rational individuals. It is also problematic to label individual beliefs as 'erroneous' or 'irrational' simply on grounds that they differ from the experts' risk assessment, as distrust against expert authorities is an actual and sometimes sound feature of a democratic society (Portney 1992; Pollack 1998). These articles focus more on the social aspects of regulation, rather than the individual decision-making process.

Viscusi (1995) connects a model of risk perception with a Peltzman-style model of optimal risk choice. His 'quasi-Bayesian' model of risk perception consists of an apriori component, a situation-dependent component that is influenced by safety efforts, and a government information component. This allows modelling misperceptions, in particular overestimates of low and underestimates of high probabilities. Viscusi explores the effect of risk perception on precaution; because of a linear composition of risk perception, however, the model does not allow for cases where safety levels and accident rates are positively related.

None of the above references, as well as a much wider literature surveyed by the authors, provide a general theoretical framework showing under which conditions an exogenous increase in the riskiness of an environment can make individuals better off. The model developed in this article seeks to fill this gap. The remainder of this article follows Peltzman in modelling optimal risk choice, but introduces risk perceptions into this model in a different way than Viscusi (1995), showing how risk perceptions can lead to cases where safety increases effect risk increases.

3 Rational choice of accident rate

We construct an idealised choice model, in which the agent chooses how risky she wants to behave. The option she chooses has one of three consequences. With probability p, she will be harmed (H). With probability q, she will obtain some advantage A facilitated by engaging in risky behaviour.[2] We abstract from differences in harms and differences in advantages. With probability $1 - p - q$, she will neither be harmed

[2] For example, driving faster, or driving at all, gets one more quickly to a desired destination; using a sharp knife yields better cutting results; using more volatile investment options offers higher gains.

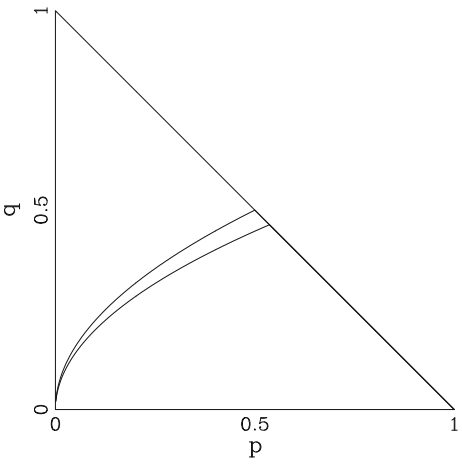

Fig. 1 The possibility frontier

nor obtain the advantage (i.e. $\neg(A \vee H)$). The agent prefers A to $\neg(A \vee H)$, and $\neg(A \vee H)$ to H. [3]

The agent's choice options can be described as lotteries over the three consequences $H, A, \neg(A \vee H)$. Graphically, it can be represented by a probability triangle, as depicted in Fig. 1. By convention, the horizontal axis measures the probability p of the harmful consequence H, increasing from left to right; the vertical axis measures the probability q of the advantageous consequence A, increasing from bottom to top. $\neg(A \vee H)$, the intermediate consequence, is located at the bottom left corner of the triangle. Any point in the triangle represents a probability mixture of the three consequences. The triangle as a whole represents the set of all possible lotteries of the form $\{p, H; q, A; 1 - p - q, \neg(A \vee H)\}$.

Not every locus of the triangle represents a possible choice option for the agent. Her choices are restricted by the following considerations. The agent can freely determine her chance of being harmed, but she cannot freely determine her chance of obtaining the advantageous consequence. This depends on the environment in which she acts, and her abilities to cope in such an environment. Further, the chance of obtaining the advantageous consequence also depends on how much risk the agent chooses to take. The agent's available options are, therefore, bound by a possibility frontier $f(p)$. Any lottery on or below this frontier is a possible choice option for the agent.

The following considerations constrain the form of such a frontier. First, it is implausible that an agent's ability to cope shifts suddenly when incrementally increasing her risk exposure p. Hence,

(i) f is continuous whenever $f(p) \leq 1 - p$.

[3] Peltzman's model commences with a technological complementarity between driving intensity and the probability of death of the driver (and so does Viscusi, as a complementarity of safety-related effort and expected loss). We model the same idea in a slightly different way, by constructing a skill-dependent possibility frontier in a probability triangle that represents points with a combined probability of beneficial and harmful consequences.

Second, for technical reasons, we also assume that it is continuously differentiable over this domain.

(ii) f is continuously differentiable whenever $f(p) \leq 1 - p$.

Third, it is plausible that the more risk exposure an agent dares, the higher an advantageous reward she will receive. We idealise this consideration by assuming that f is a monotonically increasing function of p. Hence,

(iii) $\frac{\partial f}{\partial p} > 0$ whenever $f(p) \leq 1 - p$.

Fourth, it is plausible that with each increase of risk exposure, the agent obtains lower increments in the probability of the advantageous option—the marginal gain from increasing risk exposure is diminishing. Thus, we assume that the possibility set demarcated by f is (strictly) convex.

(iv) $\frac{\partial^2 f}{\partial p^2} < 0$ whenever $f(p) \leq 1 - p$.

Finally, we assume that the agent must take some risk in order to make the benefit possible. Thus, the model is calibrated such that zero probability of harm results in zero probability of obtaining the benefit.

(v) $f(0) = 0$.

Figure 1 gives a graphical illustration of two functions satisfying these conditions.

The possibility frontiers illustrated in Fig. 1 represent two different combinations of external environment and the agent's own abilities or skills. Note that one of the depicted frontiers is strictly below the other. This represents an 'inferior' environment; compared to the other possibility frontier, the probability q of obtaining the benefit is lower for any value of p. Hence, a shift from a superior possibility frontier to an inferior possibility frontier would represent a deterioration of the agent's environment or skills. We will refer to such deterioration as a *safety decrease*.

Definition 1 A *safety decrease* is a pair of possibility frontiers $\langle f_i, f_j \rangle$ such that $f_j(p) < f_i(p)$ for all p such that $0 < p + f_j(p) \leq 1$.

It should be noted that a safety decrease, according to this definition, is not the same as an increased probability of harm. Rather, a safety decrease consists in the deterioration of the material conditions under which the agent balances the risk of harm against the chance of benefit.

Typically, the agent does not aim for a constant probability of obtaining the benefit. Given the standard assumptions of ordering plus continuity, the agent's preferences over different driving styles can be represented by a set of indifference curves with slope k. The additional standard assumption of independence of expected utility restricts the set of indifference curves to being upward sloping, linear and parallel.[4]

[4] Given the vNM axioms, the utility of a lottery $L = \{p, H; q, A; 1 - p - q, \neg(A \vee H)\}$ is equal to $u(L) = p \times u(H) + q \times u(A) + (1 - p - q) \times u(\neg(A \vee H))$. Consequently, we can write $q = \frac{u(\neg(H \vee A)) + p \times (u(H) - u(\neg(H \vee A))) - u(L)}{u(\neg(H \vee A)) - u(A)}$. As follows from the expected utility hypothesis, the values for $u(A), u(H)$, and $u(\neg(H \vee A))$ are unique up to a positive linear transformation. With $u(L)$ fixed for each indifference curve, it is then clear that the indifference curves are linear with the slope $k = \frac{\partial q}{\partial p} = \frac{u(H) - u(\neg(H \vee A))}{u(\neg(H \vee A)) - u(A)}$. Note that k is positive, since $u(A) > u(\neg(A \vee H)) > u(H)$.

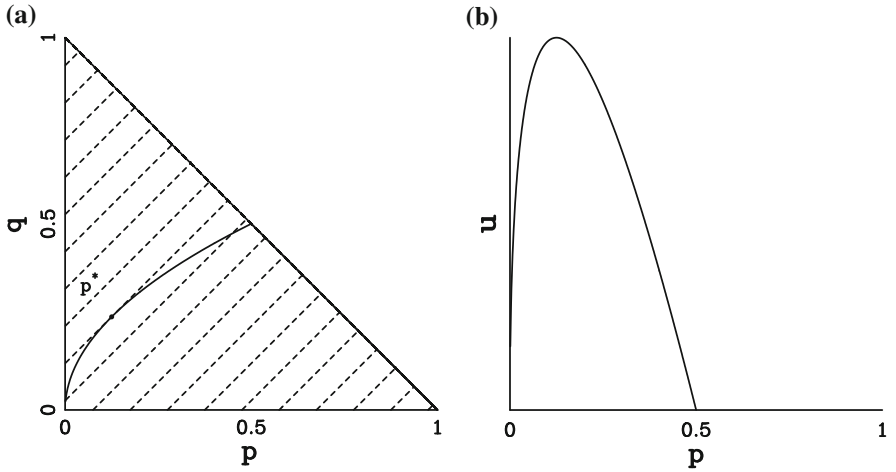

Fig. 2 The utility of risk level p: (a) intersection of possibility frontier f with indifference curves with slope $k = 1$, (b) utility function

An agent's risk preferences that satisfy these assumptions express the agent's evaluation of risk as the slope of these indifference curves, across the whole area of the probability triangle. Figure 2a depicts this situation. The point at which the possibility frontier is tangential to the indifference curves is the *optimal risk exposure* p^*, i.e., the solution $p = p^*$ to the equation $f'(p) = k$, where $f' = \partial f/\partial p$. Since possibility sets bounded by f are convex (as assumed in (iv)), the optimum is unique.

Definition 2 For any k and f, the *optimal risk exposure* $p^*_{k,f}$ *is the solution* $p = p^*$ *to the equation* $f'(p) = k$.

We now construct a utility function for any frontier point p. Note that while f is tangential to only one indifference curve, it intersects with other indifference curves at every point. These indifference curves $q(p) = u + k \times p$ have the same slope k, and only differ with respect to $q(0) = u$, i.e., the value at which the vertical line is intersected. By equating $f(p)$ with $q(p)$ we obtain $f(p) = u + k \times p$. Solving for u, this yields the utility function dependent on f, k, and p.

Definition 3 Let k be the slope of the agents indifference curves, and let $f(p)$ be a possibilities frontier. For any number p, the *frontier point utility* $u_{k,f}(p)$ is the number such that $u_{k,f}(p) = f(p) - k \times p$.

Figure 2b depicts this utility curve. A rational agent chooses that risk level which maximises $u_{k,f}$.

According to this model, agents are rational and know their risk coping skills. Therefore, they will compensate changes in the possibility frontier by changes in their risk exposure, in order to reach the risk-benefit tradeoff that is optimal according to their risk preferences. In other words, people choose the risk level that they are rationally willing to accept.

4 Risk judgement bias

Choosing a risk level requires that the agent knows which action will yield what risk for a given environment. Since information is scarce and agents fallible, agents' beliefs about risks may diverge from the actual risks they are facing. This consideration leads us to incorporate a risk perception component into our model.

Research in decision theory and psychology revealed that individuals perceive risks in ways that systematically differ from the standard decision-theoretic perspective. Consequently *risk perception* has been defined as 'the subjective assessment of the probability of a specified type of accident happening and how concerned we are with the consequences' (Sjöberg et al. 2004, p. 8). The term 'risk' thus has both an objective and a subjective interpretation. Whereas 'objective risk' typically refers to observed frequencies or statistical data, the term 'subjective risk' refers to risk assessed by individuals who may be influenced by different factors (Slovic 1987). Early psychometric literature from the 1970s took subjective risk to be individual estimations of objective risk. In the later psychometric literature, however, the level of objective risk is just one of several factors determining the level of perceived risk. Psychological, social, institutional, and cultural factors such as fear of new technology and lack of control or trust may affect risk perception (Sjöberg 2000). We say that risk judgement is *biased* if subjective and objective risks diverge.

Much of this risk perception literature has focussed upon explaining different levels of risk perception in terms of biases across a range of different activities: some activities are perceived as more dangerous than others, even if their objective risks are identical. In contrast, our model of risk judgement bias concern biases for different risk levels of one and the same activity. An individual may be biased more when confronted with low-level rather than high-level objective risks. These are the kind of biases we will focus on, and hence refer to bias effects where objective risk is the single determining factor. In this sense, the model to be outlined in this paper has more in common with the earlier works in the psychometric literature.

We augment the simple model of Sect. 3 with a risk bias function b. This function represents the risk perception of the agent in the model; b takes as its input the objective probability of harm, and it gives as output the subjective probability of harm. We assume that such a function exists and that it is a bijection, i.e., that there is a one-to-one correspondence between real and perceived probability of harm. Every possible objective probability of harm, therefore, is associated with a unique subjective probability (and vice versa). In modelling risk bias as a function, it is assumed that whatever effect we are seeing when risk is lowered is also at work when risk is increased.

Without determining b's functional form, we assume it to satisfy the following properties. First, the agent is capable of correctly perceiving a level of risk where the probability of harm is zero. Hence,

(vi) $b(0) = 0$.

Second, the agent is capable of correctly perceiving a level of risk where harm is certain. Hence,

(vii) $b(1) = 1$.

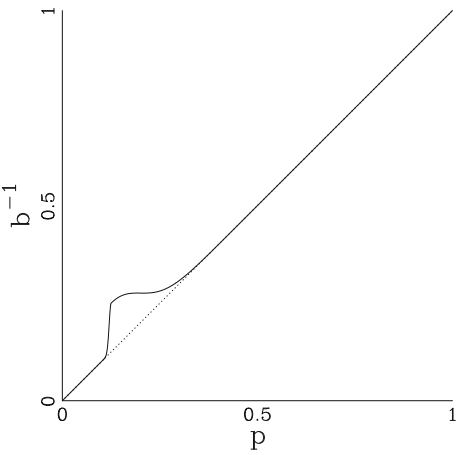

Fig. 3 The inverse bias function

Third, although the rate of incremental change is not necessarily uniform, an increased level of real risk always leads to an increased level of perceived risk. Hence,

(viii) $\frac{\partial b}{\partial p} > 0$.

Fourth, although the level of perceived risk may increase rapidly for some levels of real risk, there are no 'threshold levels' of real risk involving sudden shifts from one level of perceived risk to another. Hence,

(ix) b is continuous.

From these four assumptions, it follows that the corresponding properties also apply to the inverse of the bias function. A crucial assumption is that the agent in the model underestimates risk. In other words, there is a misperception represented by the difference $p - b(p)$.

(x) $b(p) < p$ for at least some values of p.

A particularly interesting case, which we will focus on in our argument, occurs when this misperception rapidly increases at about the same level as the optimal risk level.

For the purpose of providing a clear structural interpretation, we have so far presented and discussed the bias function b, which transforms real risk into perceived risk. What primarily is of interest in our model, however, is how the bias affects the choice of real risk. Choosing individuals, unaware of the real level of risk they face, optimise the risk they perceive, but face the real risk that they have chosen. Hence, we employ the inverse b^{-1} of the bias function. The inverse bias function (or 'de-biasing function') is a transformation from perceived risk to real risk.

As an illustration, take the function shown in Fig. 3; the vertical axis represents real risk, and the horizontal axis represents perceived risk. Along the dotted 45° line, these two probabilities are equal.

Ample empirical evidence supports assumptions (viii) and (x). It has been established that drivers overestimate their own perceptual motor skills and their safety skills in comparison to the average driver (Horswill et al. 2004). Studying different kinds

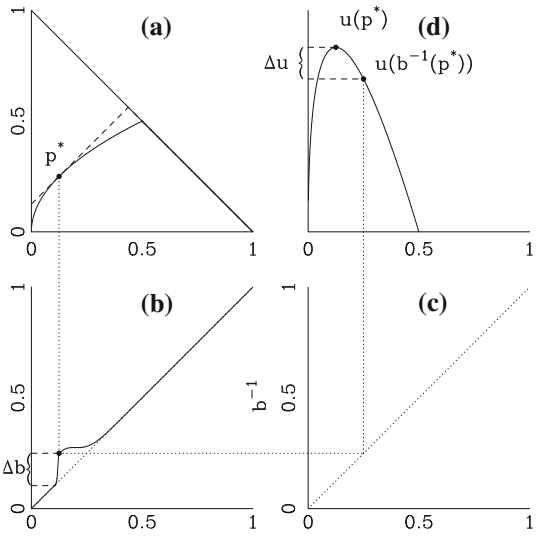

Fig. 4 Bias shifts utility result

of attitudes of these overoptimistic risk takers in the UK, Musselwhite (2006) found that those who took risk unintentionally formed the largest group. Thus, we model risk bias as taking *more* risk than was judged to be optimal.

A considerable number of studies show that overoptimism is not uniform across all risk levels of the same occupation. Rather, overoptimism increases with higher intended risk levels. This is exhibited by drivers who are compensating for safety or skill improvements. As they increase the riskiness of their actions in order to balance out the increased safety of their driving environment, they unintentionally *overcompensate*: they choose driving behaviour that leads to *more* accidents than in the status quo. For example, when lane markings were introduced to improve nighttime road visibility, drivers increased their speed to the extent that they habitually overdrive their headlamps, thus effectively worsening their safety (Cottrell 1988; Kallberg 1993; Rumar and Marsh 1998).

Moreover, Jorgensen and Pedersen (2002) argue in an analytical model that the relationship between the driver's subjective or perceived risk and the objective risk must be concave, which means that a rise in the objective probability has less influence on the driver's perceived accident rate as the real accident rate rises. This suggests too that it is more likely that drivers underestimate risk when it is high than when it is low.

Risk judgement bias leads an agent to choose a risk that differs from the objective optimal risk level. Furthermore, this bias has a negative welfare effect for the agent. We make this claim precise by combining the model of optimal risk choice from Sect. 3 with the model of risk perception bias from this section. The combined picture is provided in Fig. 4.

In the upper-left part (a), the agent chooses her optimal risk exposure by identifying the point at which the possibility frontier is tangential to her indifference curves. However, as the agent chooses her level of risk exposure according to how she perceives the situation, this is only the optimal *perceived* risk exposure. The *real* risk exposure is obtained after transformation (de-biasing) through the function b^{-1}, as illustrated

in the lower left part (b) of Fig. 4. Hence, when the agent perceives her risk-taking as optimal, the actual level of risk will be given by $b^{-1}(p^*)$, yielding a bias-induced behavioural deviation Δb.

Definition 4 For any optimal risk exposure $p_{k,f}$ and bias function b, the *bias-induced behavioural deviation* $\Delta b_{k,f}$ is equal to $|b^{-1}(p^*_{k,f}) - p^*_{k,f}|$.

In the case illustrated here, the real risk is higher than the optimal risk exposure. Consequently, as illustrated in the upper-right part (d) of Fig. 4, the utility will be lower than it would have been in the absence of perception bias. The utility of this action is $u(b^{-1}(p^*))$. The bias thus results in a welfare loss for the biased agent.

Definition 5 For any utility function $u_{k,f}$ and bias function b, the *bias-induced welfare loss* $\Delta u_{k,f,b}$ is equal to $u_{k,f}(p^*_{k,f}) - u_{k,f}(b^{-1}(p^*_{k,f}))$.

Note that a non-zero behavioural deviation Δb always yields a welfare loss, as the behavioural deviation drives the actual choice away from the unique optimal risk point p^*.

5 Beneficial safety decrease

A regulator who has control over the possibility frontier can reduce an agent's biased-induced behavioural deviation by decreasing the safety level of the agent's environment. Under certain circumstances, a reduction of behavioural deviation can have such a positive welfare effect that it outweighs the otherwise negative effect of a deterioration of the environment. In this section, we make this claim precise at the hand of our biased risk-choice model.

Recall the definition of a 'safety decrease' in Sect. 3: a pair of possibility frontiers $\langle f_i, f_j \rangle$ such that $f_j(p) < f_i(p)$ for all p such that $0 < p + f_j(p) \leq 1$. As discussed, a safety decrease may affect an agent's behaviour such that he acts more cautiously.

Definition 6 Given a slope k of indifference curves, a safety decrease $\langle f_i, f_j \rangle$ is *caution-inducing* (CI_k) if and only if $p^*_{k,f_j} < p^*_{k,f_i}$.

In general, we should expect such increased caution to reduce the agent's welfare, as p^*_j will be located on a lower indifference level that p^*_j. However, Fig. 5 illustrates that a risk increase can have a *positive* welfare effect.

We explain how such a positive welfare effect is possible by commenting on each part of the figure.

In the upper-left part (a) of the figure, there are two possibility frontiers: f_i and f_j. The first frontier is strictly above the second frontier. Consequently, the optimal risk exposure associated with each frontier—i.e., the point where the possibility frontier is tangential with the slope of the indifference curves—will differ; the optimal risk exposure $p^*_{f_i}$ associated with f_i is higher than the optimal risk exposure $p^*_{f_j}$ associated with f_j. The situation where the possibility frontier shifts from f_i to f_j, and consequently the optimal risk exposure shifts from $p^*_{f_i}$ to $p^*_{f_j}$, represents a deterioration of the environment. As the situation becomes more 'dangerous', the agent chooses

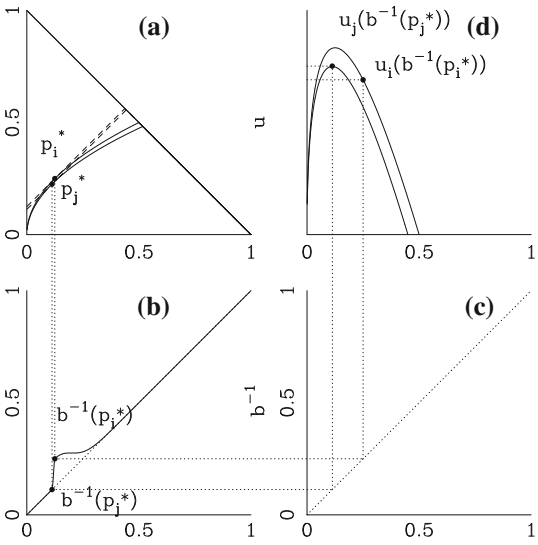

Fig. 5 Positive welfare effect of environmental deterioration

a lower level of risk exposure so as to maintain balance between the probability q of obtaining the benefit and the probability p of getting harmed. Since this puts his optimal risk exposure on a lower indifferent curve, it constitutes a *prima facie* welfare loss.

The upper-right part (d) shows this welfare loss more generally. The safety decrease $\langle f_i, f_j \rangle$ produced a decrease in the corresponding frontier point utility function. For all $p, u_i(p)$ is larger than $u_j(p)$. In particular, $u_i(p_i^*) > u_j(p_j^*)$, hence the *prima facie* welfare loss.

However, in the lower left part (b) of the figure, each of the two optimal risk exposures undergoes the transformation explained in Sect. 4. While at p_i^*, the agent is subject to a considerable behavioural deviation, the deviation at p_j^* is much smaller. Thus, $\Delta b_{k,f_i} > \Delta b_{k,f_j}$.

The lower-right part (c) just reflects these different deviations on the 45° line onto the two utility functions in the upper-right part (d) of Fig. 5. Now we see how the reduction of a behavioural deviation may outweigh the *prima facie* welfare loss of a safety decrease: although $b^{-1}(p_j^*)$ lies on the diminished utility function u_j, it is close enough to u_j's maximum to yield a higher utility than $b^{-1}(p_i^*)$ on u_i.

This leads us to the general consideration of when a safety decrease may be beneficial, which is the case if and only if the agent exposed to the lowered possibility frontier obtains a higher frontier point utility than under the previous possibility frontier.

Definition 7 Given a slope k of indifference curves, and a bias function b, a *Beneficial Safety Decrease* ($BSD_{k,b}$) is a safety decrease $\langle f_i, f_j \rangle$ such that $u_{f_j}(b^{-1}(p_{f_j}^*)) > u_{f_i}(b^{-1}(p_{f_i}^*))$.

Without making any further assumptions about the form of the involved functions, we arrive at the following results. First, a Beneficial Safety Increase consists of the trade-off of two effects on the utility function.

Lemma 1 *A Beneficial Safety Decrease ($BSD_{k,b}$) exists for some safety decrease $\langle f_i, f_j \rangle$ iff $\Delta u_{k,f_i,b} - \Delta u_{k,f_j,b} > u_{k,f_i}(p^*_{k,f_i}) - u_{k,f_j}(p^*_{k,f_j})$.*

Proof [5] Add $[u_i(p_i^*) - u_j(p_j^*)]$ to both sides of the inequality of $u_j(b^{-1}(p_j^*)) > u_i(b^{-1}(p_i^*))$. Rearrange, so that

$$[u_i(p_i^*) - u_i(b^{-1}(p_i^*))] - [u_j(p_j^*) - u_j(b^{-1}(p_j^*))] > u_i(p_i^*) - u_j(p_j^*) \quad (1)$$

Then by Definition 5, $\Delta u_i - \Delta u_j > u_i(p_i^*) - u_j(p_j^*)$. □

Thus, a BSD can be analysed into both the effect of a safety decrease on the shape of the utility function (right hand side of the inequality 1) as well as into the effect of the safety decrease on the choice of the biased optimum (left hand side of the inequality). If the latter yields a utility gain that outweighs the utility loss through the former, then a Beneficial Safety Decrease obtains. This result allows us to formulate a necessary condition for BSD:

Proposition 1 *A Beneficial Safety Decrease ($BSD_{k,b}$) exists for some safety decrease $\langle f_i, f_j \rangle$ only if $\Delta u_{k,f_i,b} - \Delta u_{k,f_j,b} > 0$.*

Proof By Definition 3, $u_{k,f}(p) = f(p) - k \times p$. Hence for any safety decrease $\langle f_i, f_j \rangle, u_j(p) < u_i(p)$ for any p. Thus, $u_i(p) - u_j(p) > 0$ for any p, and in particular for the optimal point p^*. Then by Lemma 1, $\Delta u_i - \Delta u_j > 0$. □

Thus, a Beneficial Safety Decrease is possible only if the safety decrease reduces the welfare loss, as the effect of the safety decrease on the utility function is always negative. Note that this is only a necessary, not sufficient condition for BSD.

We now turn to identifying a sufficient condition.

Proposition 2 *Let $\langle f_i, f_j \rangle$ be a caution-inducing (CI_k) safety decrease, and let b be such that $p^*_{k,f_j} < b^{-1}(p^*_{k,f_j})$. If*

$$\frac{\Delta u_{k,f_j,b}}{\Delta b_{k,f_j}} < \frac{u_{k,f_j}(p^*_{k,f_j}) - u_{k,f_i}(b^{-1}(p^*_{k,f_i}))}{b^{-1}(p^*_{k,f_i}) - p^*_{k,f_j}} \quad (2)$$

then $\langle f_i, f_j \rangle$ is a Beneficial Safety Decrease ($BSD_{k,b}$).

Proof Since $\langle f_i, f_j \rangle$ is a caution-inducing safety decrease, $b^{-1}(p_i^*) > b^{-1}(p_j^*)$ follows from condition (viii) of the bias function; consequently, $b^{-1}(p_i^*) - p_j^* > \Delta b_j$. Since $p_j^* < b^{-1}(p_j^*)$, we have $\Delta b_j^{-1} > 0$. We then have

$$\frac{\Delta u_j}{\Delta b_j} < \frac{u_j(p_j^*) - u_i(b^{-1}(p_i^*))}{b^{-1}(p_i^*) - p_j^*} < \frac{u_j(p_j^*) - u_i(b^{-1}(p_i^*))}{\Delta b_j}$$

[5] In this and in further proofs, indices k and b are omitted where convenient. Indices f_i and f_j are furthermore replaced by i and j, respectively.

Eliminating the denominator gives $\Delta u_j < u_j(p_j^*) - u_i(b^{-1}(p_i^*))$, which after subtracting $u_j(p_j^*)$ gives $-u_j(b^{-1}(p_j^*)) < -u_i(b^{-1}(p_i^*))$. Hence, $u_j(b^{-1}(p_j^*)) > u_i(b^{-1}(p_i^*))$. □

How should Proposition 2 be understood? We proceed by giving an interpretation to each of the constituent elements of Inequality 2. The expression on the left hand side of the inequality can be interpreted as the *bias power*: as a measure of how important it is for the agent to accurately perceive risks. The larger this fraction is, the more serious will the consequences of deviating from the optimal risk exposure be for the agent.

On the right hand side of Inequality 2, the expression in the numerator can be interpreted as the *welfare potential*: as a measure of the scope for improvement in the safety decrease. The larger this difference is, the larger is the potential positive effect of the safety decrease.

Finally, the denominator on the right hand side of Inequality 2 can be interpreted as the *error potential*: as a measure of the scope for behavioural deviation. The larger this difference is, the larger is the scope for deviation from the optimal risk exposure.

An informal interpretation of Inequality 2 may run as follows: 'If the bias power is small, if a large welfare gain can be made, and if the potential bias error is small, then a shift from a safe environment to an unsafe environment is beneficial.' This completes our general analysis of the conditions under which BSD exists.

6 Numerical example of BSD

We now show that a BSD exists for the specific and intuitively plausible functional form used in the illustrative graphs. The possibility frontier takes the form:

$$f_\alpha(p) = \alpha \times \sqrt{p} \tag{3}$$

We think this functional form is a plausible candidate for the possibility frontier. First, it satisfies properties (i)–(v). Second, increases of very low harm probabilities will lead to very high increases in the probability of advantageous consequences. This is often the case with risks encountered in everyday life. For example, driving 20 km/h on a country lane instead of 5 km/h, for example, is only a little more risky, but gets you home a lot quicker. Hence, the slope of f should be steep for p close to 0. Third, increases of higher risk levels lead to comparatively small increases in the probability of advantageous consequences. For example, using razor blades for cutting your vegetables may make the cutting a little easier, but seriously increases the probability of harm. Hence, the slope of f should be flat for p close to 1. Fourth, skills increasingly differentiate individual agents with increasing p. Most people are capable of coping with moderate dangers, e.g. driving at moderate speeds, using moderately dangerous tools, or using moderately difficult hiking paths. Yet when it comes to more dangerous environments, only few people are quite capable of achieving good results—think of race car drivers, chefs or mountain guides—while others lack the skills to cope in such environments, obtaining a comparatively low risk-benefit trade-off in these

unsafe conditions. Therefore, possibility frontiers based on different α should 'fan out' with increasing p, crossing the diagonal of the triangle at different heights.

Conveniently, possibility frontier 3 only depends on one parameter: α. Obviously, in this specific case, a pair $\langle f_{\alpha_i}, f_{\alpha_j} \rangle$ of possibility frontiers is a safety decrease if $0 < \alpha_j < \alpha_i$. The optimal risk exposure is given by setting the first derivative of f to k:

$$\frac{\partial u}{\partial p} = \frac{\partial f_\alpha}{\partial p} - k = 0 \tag{4}$$

$$\Rightarrow p^*_{k,f_\alpha} = \frac{\alpha^2}{4k^2} \tag{5}$$

However, the agent's risk judgement is biased. In the illustrative graphs, the bias function (from Fig. 3 and onwards) has the form:

$$b^{-1}(p) = \begin{cases} p + m \times \exp\left(-\frac{(p-m)^2}{\sigma_1^2}\right) & : \quad p < m \\ p + m \times \exp\left(-\frac{(p-m)^2}{\sigma_2^2}\right) & : \quad p \geq m \end{cases} \tag{6}$$

With $1 > \sigma_2 > \sigma_1 > 0$, the left hand side of the graph is steeper than the right. We think that this functional form is a plausible candidate for the bias function. It satisfies properties (vi)–(x). In particular, it is 'smooth' in the sense of being continuous and continuously differentiable, and it does not have a maximum.

We only need to consider the left side with $\sigma = \sigma_1$, as p^* always shifts to the left in a caution-inducing safety decrease. This yields the biased utility function u_{k,f_α}:

$$\begin{aligned} & u_{k,f_\alpha}(b^{-1}(p^*_{k,f_\alpha})) \\ & = f_\alpha(b^{-1}(p^*_{k,f_\alpha})) - k \times b^{-1}(p^*_{k,f_\alpha}) = \alpha \times \sqrt{b^{-1}\left(\frac{\alpha^2}{4k^2}\right)} - k b^{-1}\left(\frac{\alpha^2}{4k^2}\right) \\ & = \alpha \times \sqrt{\frac{\alpha^2}{4k^2} + m \times \exp\left(-\frac{\left(\frac{\alpha^2}{4k^2} - m\right)^2}{\sigma^2}\right)} \\ & \quad - \frac{\alpha^2}{4k} - k \times m \times \exp\left(-\frac{\left(\frac{\alpha^2}{4k^2} - m\right)^2}{\sigma^2}\right) \end{aligned}$$

In the illustrations, we use $k = 1$, $m = 0.125$, and $\sigma = 0.005$ for the parameters associated with the individual agent. These specific values can be inserted into the

above equation:

$$u_{1,f_\alpha}(b^{-1}(p^*_{1,f_\alpha})) = \alpha \times \sqrt{\frac{\alpha^2}{4} + 0.125 \exp\left(-\frac{\left(\frac{\alpha^2}{4} - 0.125\right)^2}{0.005^2}\right)} - \frac{\alpha^2}{4} - 0.125 \exp\left(-\frac{\left(\frac{\alpha^2}{4} - 0.125\right)^2}{0.005^2}\right)$$

Straightforward numeric calculation reveals that this expression is increasing with decreasing α for $0 < \alpha < \alpha_L \approx 0.688$ and for $\alpha > \alpha_H \approx 0.706$. In the interval $[\alpha_L, \alpha_H]$, however, the utility is increasing with decreasing α; a decrease in α—i.e., a safety decrease—would be beneficial in this interval. That is, any pair $\langle f_{\alpha_i}, f_{\alpha_j} \rangle$ such that $\alpha_L \leq \alpha_j < \alpha_i \leq \alpha_H$ is a beneficial safety decrease.

7 Discussion

Our model can be applied to three separate domains: first an explanation of anomalous safety-risk relations; second as a basis for critically assessing certain risk policies; and third as a conceptual basis for risk policy design.

First, when curved roads are straightened out, when slippery road sections are replaced by high-friction road surfaces, or when unmarked pedestrian crosswalks are equipped with zebra stripes, the intuition behind these changes is that they decrease the risks of those involved. Evidence referred to in this article shows that at least some of the time, such intended safety increases are offset by actors willingly and rationally choosing a riskier behaviour.

Yet beyond this offsetting effects, it is sometimes observed that safety *increases* have led to an increase in accident rates. The literature reviewed in Sect. 2 points to such averse effects from skill and environment improvements. Our model offers an explanation of these observations, by showing how safety increases, when combined with biased risk judgement, can yield a choice with increased risk levels.

Second, many risk-inducing behaviours are subject to regulations. Often, the regulations are justified as the prevention of harm to others. Yet apart from reducing such externalities, risk regulation is often justified as saving people from the risks they impose on themselves. Such paternalistic regulations may be based on the conviction that certain risk thresholds should never be crossed, irrespective of the reasons people may have for doing so. These kinds of hard paternalism are difficult to square with liberal positions common to current democratic societies. Alternatively, the regulator may intervene paternalistically because it is believed that people expose themselves to unintended risk levels. Limited capacities to perceive or process crucial information are possible reasons for such failures. The risk perception bias discussed in Sect. 4 is an example of such limited capacities: individuals choose the optimal perceived risk,

but by that expose themselves to an unintended level of real risk. This may lead to a welfare loss, which in turn may justify a paternalistic intervention. The regulator intervenes in order to help the individual make the choice she wants to make, but is incapable of making. It has been argued that these kinds of soft paternalism are compatible with liberal positions (e.g. Sunstein and Thaler 2003; Camerer et al. 2003). The use of these interventions has recently been widely advocated.

Our model works out a potential complication arising from such interventions. As we argued in Sect. 3, risk is not a property of objects or contexts, but the outcome of action in these contexts or involving these objects. Yet in many cases, risk regulation affects the properties of objects and environments, not people's actions themselves. A good part of the regulations concerning automobile safety, air traffic safety, building safety, children safety, sports safety and workplace safety are directed at the cars, air traffic control, building codes, toys, sports equipment and machines, not the people who use these objects or choose in these environments. In these cases, it is not immediately obvious what the effect of the regulator's intervention is with respect to the risk levels the affected people will ultimately choose. By presenting the possibility of a welfare-reducing safety increase, and characterising its sufficient conditions, our model thus serves as a basis to critically assess certain risk policies.

Third, our model also offers guidance for the design of paternalistic policies. If accident countermeasures sometimes may increase danger, rather than diminish it, then the possibility arises that lowering of safety levels could yield an improvement of the risk situation. That is, the regulator may tamper with the environment that agents adjust their optimal risk levels to, in such a way that the bias effect of the intended risk level is minimised. This idea has been raised before, but never been made precise. The economist Armen Alchian once suggested to fit each car's steering wheel with a spear directed at the heart of the driver, so as to make her acutely aware of the dangers involved in driving (quoted in Landsburg 1993, p. 5). Under the standard risk homeostasis model, such a proposal is nonsensical: all it does is to make the driver adjust her behaviour to the additional danger of even a light collision. This would create a welfare loss to the driver. Therefore, why could Alchian have proposed it? According to our model, the spike allows the driver to choose an action that realises her intended risk level more closely. In other words, the bias her action creates is *smaller* under the newly adjusted risk level than it was under the old, such that the decrease in bias (a welfare increasing effect) *offsets* the welfare loss through the increased danger.

Indeed, it seems that regulators have occasionally picked up on the idea of BSD. Driving through Skåne in southern Sweden last summer, one of the authors encountered a very narrow and unmanageable traffic roundabout. This seemed surprising, given the generally good quality of Swedish traffic regulation and concern for road safety. The surprise grew bigger when it became clear that the roundabout had been recently *reduced* in size, and hence *made* more narrow and less manageable. Assamin that smaller roundabouts increase driver caution, and that this was the intention behind reducing the size of the roundabout, this brief anecdote is a concrete example of an intervention with the purpose to alter agents behaviour. Similarly, the results presented by Kallberg (1993) and others have led to the removal of visual elements such

as road delineation in an effort to reduce speeding by providing less guidance. In an experiment in the Netherlands, edge-lines were completely removed to eliminate their guiding property. On-road evaluation data (De Waard et al. 1995) and accident data (Steyvers 1999) have shown that this measure works as intended. Our model provides a conceptual basis for arguments supporting policy measures of this kind.

8 Conclusion

In this article, we pursued three objectives. First, we proved the existence of a Beneficial Safety Decrease; we identified the conditions under which individuals are better off when the regulator deteriorates elements of the infrastructure.

Second, this article details the important influence of individuals' choices and choice biases on the results of safety measures. The policy maker must take this influence into account: what ultimately counts is the end result, measured in the individuals' own risk preferences, not the intention of the policymaker (e.g. 'maximise safety devices') or her values (e.g. 'none should take higher risks than threshold t'). As our article shows, even if policymakers aim to improve individuals' adherence to their own values, instead of imposing its own, there is urgent need for intervention. The actual case of the Skåne roundabout is, in our view, a successful example of such regulation. Many more areas of application can be thought of. Examples include zebra crossings, demarcated cycle paths on the road, safety ropes on alpine paths, safety devices for children, and many more. In each of these cases, introducing a 'safer' device may lead to an overcompensation that actually worsens the safety of the people using it. We hope that policymakers will take this result to heart when introducing new regulation in such areas.

Third, on a more philosophical note, the examples we discussed and the theoretical treatment we presented in this article show that risk is not a property inherent to the environment. Rather, risk is a feature that arises out of the choices people make in these environments. Many people deem a steep mountain ridge an inherently dangerous, while they consider a zebra crossing safe. But it may well be that a hiker on the ridge actually has a *lower* risk of death than a person crossing the road, because the ridge makes the hiker choose a very cautious behaviour, while the zebra crossing has no such effect on the pedestrian. This article, we hope, will therefore contribute to a rethinking of the notion of safety, and its connection to risk.

References

Arnould, R., & Grabowski, H. (1981). Auto safety regulation: An analysis of market failure. *Bell Journal of Economics, 12*, 27–48.

Camerer, C., Issacharoff, S., Loewenstein, G., O'Donoghue, T., & Rabin, M. (2003). Regulation for conservatives: Behavioral economics and the case for asymmetric paternalism. *University of Pennsylvania Law Review, 1151*(3), 1211–1254.

Cottrell, B. H. (1988). The effects of wide edgelines on two-lane rural roads. *Transportation Research Record, 1160*, 35–44.

De Waard, D., Jessurun, M., Steyvers, F. J. J. M., Raggatt, P. T. F., & Brookhuis, K. A. (1995). Effect of road layout and road environment on driving performance, drivers' physiology and road appreciation. *Ergonomics, 38*, 1395–1407.

Evans, W. N., & Graham, J. D. (1991). Risk reduction or risk compensation? The case of mandatory safety-belt use laws. *Journal of Risk and Uncertainty, 4*, 61–73.

Horswill, M. S., Waylen, A. E., & Tofield, M. I. (2004). Drivers' ratings of different components of their own driving skill: A greater illusion of superiority for skills that relate to accident involvement. *Journal of Applied Social Psychology, 34*(1), 177–195.

Johansson, B. S. (1997). Trafikträngade barn löper större olycksrisk' [Traffic-trained children run a larger accident risk]. *Väg-och Transportforskningsinstitutet Aktuellt* 4(June), 9.

Jorgensen, F., & Pedersen, P. A. (2002). Drivers' response to the installation of road lighting. An economic interpretation. *Accident Analysis and Prevention, 34*(5), 601–608.

Kallberg, V.-P. (1993). Reflector post-signs of danger? *Transportation Research Record, 1403*, 57–66.

Landsburg, S. (1993). *The armchair economist*. New York: Free Press.

Lave, L. B., & Weber, W. W. (1970). A benefit-cost analysis of auto safety features. *Applied Economics, 4*, 265–275.

Lund, A. K., Williams, A. F., & Zador, P. (1986). High school driver education further evaluation of the DeKalb County study. *Accident Analysis and Prevention, 18*, 349–357.

Musselwhite, C. (2006). Attitudes towards vehicle driving behaviour: Categorising and contextualising risk. *Accident Analysis and Prevention, 38*(2), 324–334.

O'Donoghue, T., & Rabin, M. (2003). Studying optimal paternalism, illustrated by a model of sin taxes. *American Economic Review, Papers and Proceedings, 93*, 186–191.

Peltzman, S. (1975). The effect of automobile safety regulation. *Journal of Political Economy, 83*, 677–726.

Pollack, R. A. (1998). Imagined risks and cost-benefit analysis. *American Economic Review, 88*, 376–380.

Portney, P. R. (1992). Trouble in Happyville. *Journal of Policy Analysis and Management, 11*, 131–132.

Rumar, K., & Marsh, D. K. (1998). *Lane markings in night driving: A review of past research and of the present situation*. Ann Arbor: UMTRI-98-50, University of Michigan Transportation Research Institute.

Sagberg, F., Fosser, S., & Sætermo, I. F. (1997). Investigation of behavioral adaptation to airbags and antilock brakes among taxi drivers. *Accident Analysis and Prevention, 29*, 293–302.

Salanié, F., & Treich, N. (2009). Regulation in Happyville. *Economic Journal, 119*, 665–679.

Sjöberg, L. (2000). Factors in risk perception. *Risk Analysis, 20*(1), 1–11.

Sjöberg, L., Moen, B.-E., & Rundmo, T. (2004). *Explaining risk perception. An evaluation of the psychometric paradigm in risk perception research*. Trondheim: Rotunde publikasjoner no. 84. Available online at www.svt.ntnu.no/psy/Torbjorn.Rundmo/Psychometric_paradigm.pdf (2009-11-16).

Slovic, P. (1987). Perception of risk. *Science, 236*, 280–285.

Steyvers, F. J. J. M. (1999). Increasing safety by removing visual cues—a contradiction?. In A. E. Gale, I. D. Brown, C. M. Haslegrave, & S. P. Taylor (Eds.), *Vision in vehicles VII* (pp. 301–310). Amsterdam: Elsevier.

Sunstein, C. R., & Thaler, R. H. (2003). Libertarian paternalism is not an oxymoron. *University of Chicago Law Review, 70*(4), 1159–1202.

Traynor , T. L. (1993). The Peltzman hypothesis revisited: An isolated evaluation of offsetting driver behavior. *Journal of Risk and Uncertainty, 7*, 237–247.

van Driel, C. J. G., Davidse, R. J., & van Maarseveen, M. F. A. M. (2004). The effects of an edgeline on speed and lateral position: A meta-analysis. *Accident Analysis and Prevention, 36*(4), 671–682.

Viscusi, W. K. (1984). The lulling effect: The impact of child-resistant packaging on aspirin and analgesic ingestions. *American Economic Review, 74*, 324–327.

Viscusi, W. K. (1995). Government action, biases in risk perception, and insurance decisions. *The Geneva Papers on Risk and Insurance Theory, 20*(1), 93–110.

Winston, C., Maheshri, V., & Mannering, F. (2006). An exploration of the offset hypothesis using disaggregate data:the case of airbags and antilock brakes. *Journal of Risk and Uncertainty, 32*, 83–99.