2001

Subrogation and Insolvency

Alan O. Sykes

SUBROGATION AND INSOLVENCY

*ALAN O. SYKES**

ABSTRACT

When tort judgments exceed the assets of tortfeasors and the tort victim has first-party insurance for a portion of the loss suffered, the question arises as to how the recovery from the tortfeasor should be divided between the tort victim on the one hand and the insurer via its rights of subrogation on the other. A common view among the courts and legal commentators is that the insured should be made whole before the insurer recovers subrogation. This paper employs simple models of optimal insurance contracts to show that the opposite rule will often be optimal. Accordingly, there is little basis for judicial interference with freedom of contract when the insurance agreement provides that the insurer must be made whole before the insured receives any portion of the recovery. The analysis also provides some support for allowing the insurer to take first as a default rule at common law.

"SUBROGATION" is "the substitution of one person in place of another with reference to a lawful claim."[1] For example, if Blue Cross/Blue Shield pays the medical expenses of a policyholder who has been injured by a negligent driver, Blue Cross/Blue Shield may then have the right to recover those expenses from the negligent party in a subrogation action. If the insured has already received money from the injurer, the insured may hold such funds in "constructive trust" for the insurer because of the insurer's subrogation rights. An action by the insurer against the insured to recover such funds is often termed an action for "reimbursement."

Subrogation rights are common in insurance relationships and may arise by contract or at common law. In most jurisdictions, the common law provides a subrogation right to insurers under property, liability, and some casualty policies. Although subrogation was not generally available at common law in health and medical policies, most such policies now include subrogation

* Frank and Bernice J. Greenberg Professor of Law, University of Chicago. I thank workshop participants at Chicago and the American Law and Economics Association annual meeting for their comments and gratefully acknowledge support from the Lynde and Harry Bradley Foundation and the Sarah Scaife Foundation.

[1] Black's Law Dictionary 1595 (4th ed. 1968).

[*Journal of Legal Studies*, vol. XXX (June 2001)]
 0047-2530/2001/3002-0004$01.50

clauses expressly.[2]

When injurers are fully solvent—able to pay damages judgments against them in full—subrogation (or reimbursement) prevents the injured party from achieving a double recovery (one from the insurer and one from the injurer). The public policy against such "windfalls" is an oft-stated rationale for subrogation.[3] An economist might rephrase the policy concern as one for excessive moral hazard that might arise if injured parties actually came out ahead after an accident, as well as one for suboptimal risk bearing that would arise if risk-averse insureds were more than made whole after covered losses. For these reasons, the insurer's right of subrogation is clearly part of the optimal insurance contract when injurers are fully solvent. Similarly, if all of the insured's losses are covered by first-party insurance, subrogation is also plainly optimal (whether or not the injurer is fully solvent) because the insured is made whole by first-party coverage. Any recovery by the insured from the injurer under these circumstances would again be a "double recovery."

The more difficult case—and the case that has been the subject of considerable litigation—arises when the injurer is not fully solvent and the insured's losses are not fully covered by first-party insurance. If the injurer has no assets at all, of course, the nature of subrogation rights is of no consequence. But suppose that an injurer has some assets, though less than the insured's total loss (I will term such injurers "partially solvent"), and that the insured has been made partially whole by first-party insurance benefits. How should the injurer's limited assets be divided between the insurer and the insured? To make it concrete, suppose that an insured has been injured by a negligent driver and a first-party insurer has paid medical bills in the amount of $200,000. The insured has no other applicable insurance coverage. The insured obtains a judgment against the negligent driver in the amount of $500,000, reflecting the $200,000 in medical bills, $100,000 in lost wages, and $200,000 in pain and suffering. The tortfeasor has assets of $200,000. Claiming subrogation rights, the insurer then insists that it is entitled to this $200,000. Should such a claim be allowed? At common law? If the insurance contract provides for it? To my knowledge, these issues have not been addressed in the law and economics literature.[4]

The courts have been forced to address them, however, and have come to

[2] See John F. Dobbyn, Insurance Law in a Nutshell 285–90 (3d ed. 1996). The common-law logic in providing or denying subrogation for different types of policies is interesting in its own right and raises puzzles that are beyond the scope of the analysis here. Life insurers are not entitled to subrogation, for example, on the rather artificial theory that they contract not to cover the economic losses of the insured but rather to pay a stipulated sum in the event of the insured's death—their contract is more of an investment than an insurance contract, it is often said. *Id.* at 285–86.

[3] *Id.* at 284.

[4] The issues are addressed in a fine student note, Jeffrey A. Greenblatt, Insurance and Subrogation: Where the Pie Isn't Big Enough, Who Eats Last? 64 U. Chi. L. Rev. 1337 (1997). Greenblatt does not employ the tools of insurance economics used here, however, and thus his analysis is quite different from the analysis in this paper.

no consensus. According to John Dobbyn, "In the absence of stipulations to the contrary in the insurance contract, the majority of courts direct that the insured is to be compensated first . . . to the extent to which his loss exceeds the insurance proceeds."[5] This rule is sometimes called the "made-whole doctrine"—the insured must be made whole before the insurer can recover anything. Dobbyn's statement of the made-whole doctrine suggests that parties to an insurance contract can readily agree to alternative arrangements, but in fact a number of courts insist that the insured be made whole before the insurer receives anything despite a seemingly clear provision to the contrary in the contract.[6] Other courts do allow the insurer to recover first if the contract so provides.[7] Still others have applied a sharing rule that divides the recovery in proportion to the losses borne by the insurer and the insured.[8]

Which of these approaches, if any, is the best? A naive response rests on the simple observation that the insured is risk averse and the insurer is approximately risk neutral. Therefore, one might suggest, optimal risk sharing requires that the insurer recover nothing until the insured has been made whole. Some commentators, including Kenneth Abraham in his well-known insurance treatise, argue from this intuition against any subrogation recovery by the insurer that would impair the insured's ability to obtain full indemnity.[9]

This line of thinking is mistaken, however, for it ignores the question of why the insured was underinsured in the first place. One reason for underinsurance is that risk-averse individuals do not desire insurance against all legally compensable losses—only those losses that cause the marginal utility of money to increase relative to other states of nature. Another reason for underinsurance relates to the fact that the price of insurance often is not "actuarially fair." Underinsurance may also arise because devices such as coinsurance and deductibles are valuable for controlling moral hazard. This paper explores these possibilities with the aid of simple formal models and sets out the optimal subrogation rule for each case.

The analysis suggests that it is optimal under a wide range of circumstances for insurers to be reimbursed in full before insureds receive anything from the partially solvent injurer, even though the consequence is that the insured is not made whole. The details depend on the reason why the insured lacks

[5] Dobbyn, *supra* note 2, at 293.

[6] See Garrity v. Rural Mutual Insurance Co., 77 Wis. 2d 537, 253 N.W.2d 512 (1977); Rimes v. State Farm Mutual Automobile Insurance Co., 106 Wis. 2d 263, 316 N.W.2d 348 (1982); Higginbotham v. Arkansas Blue Cross and Blue Shield, 312 Ark. 199, 849 S.W.2d 464 (1993); Powell v. Blue Cross and Blue Shield of Alabama, 581 S.2d 772 (Ala. 1990). In *Powell,* for example, the policy provided that "separate from and in addition to the Administrator's right of subrogation . . . [t]he right to reimbursement of the Administrator comes first even if a Member is not paid for all of his claim for damages." 581 S.2d at 774.

[7] See Gibson v. Country Mutual Insurance Co., 193 Ill. App. 3d 87, 549 N.E.2d 23 (1990).

[8] See Magsipoc v. Larsen, 639 S.2d 1038 (Fla. Dist. Ct. App. 1994).

[9] Kenneth S. Abraham, Distributing Risk: Insurance, Legal Theory and Public Policy 155 (1986).

full coverage for losses, but the general intuition may be expressed as follows: from the standpoint of optimal risk allocation, the fortuitous presence of an injurer who is liable for the insured's losses does not affect the risk that the insured should optimally bear. Thus, if it is optimal for the insured to bear the risk of uncovered losses associated with accidents in which no injurer is liable, the insured should bear the same risk in accidents in which an injurer is liable. Accordingly, no violence is done to optimal risk sharing if the insurer is reimbursed first by the partially insolvent injurer. The caveats all relate to the possibility that underinsurance is not optimal for the insured, presumably because of some market failure (at least relative to the first best).

The analysis is the same whether the case against the injurer is litigated to conclusion or is settled. I mention the latter group of cases only because they present a special problem for the made-whole doctrine: when an insured settles with a partially solvent injurer, the insured can credibly argue that the settlement amount is less than the amount that a court would have awarded at the conclusion of litigation (because the settlement negotiations took into account the injurer's limited assets) and thus that the settlement is not enough to make the insured whole. Yet because of the settlement, the amount required to make the insured whole will not be independently determined by a court as part of the action against the injurer. As a result, some courts considering reimbursement claims have required a "minitrial" under the made-whole doctrine to determine the insured's loss.[10] If the made-whole doctrine is a mistake, however, such actions are unnecessary.

I will proceed by considering in turn various explanations for the insured's lack of complete coverage against loss. A concluding section gathers the results and summarizes their implications for the law.

I. Nonpecuniary Losses

Insurance markets arise because it is valuable for individuals to shift money from "states of nature" in which the marginal utility of money is low (ordinary situations in which the insured pays premiums) to states in which the marginal utility of money is high (situations in which a covered loss has occurred and the insured receives a positive net payment from the insurance company). Conventional insurance economics further posits that the marginal utility of money is diminishing for individuals and, hence, that the states in which the marginal utility of money is high will ordinarily be the states in which the insured's wealth has suffered a significant decline.[11]

In personal injury cases, however, substantial damages are regularly awarded for nonpecuniary losses, such as pain and suffering. Such awards are defensible

[10] See Rimes v. State Farm Mutual Automobile Insurance Co.; Higginbotham v. Arkansas Blue Cross and Blue Shield.

[11] Classic treatments of insurance economics may be found in Kenneth J. Arrow, Essays in the Theory of Risk Bearing (1974); Karl H. Borch, Economics of Insurance (1990).

on optimal deterrence grounds, as the social cost of accidents assuredly includes these nonmonetary harms. But it does not follow that such harms should be or are insurable or that the rules of subrogation ought to shift to an insurer the risk that an insured may be unable to collect these awards from a partially solvent injurer.

Indeed, we do not observe pain and suffering insurance in the marketplace, for at least two reasons. First, the monetary value of pain and suffering is not readily ascertainable, and pain is often difficult to verify. Pain and suffering insurance might then create significant moral hazards, at least for certain types of accidents. Second, and perhaps more fundamental, it is hardly clear that the marginal utility of money is higher for people in states of the world in which they experience pain and suffering. If their economic losses (such as medical expenses) in those states are covered by insurance, the marginal utility of additional dollars may well be lower rather than higher in comparison to other states of the world because a person suffering pain may be less able to enjoy consumption expenditures.[12] And if that is the case, it will be irrational for them to reduce their wealth (by paying insurance premiums) in states in which the marginal utility of money is higher in order to receive money in states in which it is lower. A fortiori, the made-whole doctrine will be undesirable at least to the extent that it requires the insurer to forfeit subrogation rights until the insured has been fully compensated for pain and suffering.

This proposition can be illustrated and elaborated using a simple model. Consider an insured who maximizes expected utility. The insured's (Von Neumann–Morgenstern) utility function is $u_n(w)$ in the "no-accident" state of nature and $u_a(w)$ in an "accident" state, where w denotes wealth. The difference in utility across the no-accident and accident states reflects the nonpecuniary loss attributable to the accident. Thus, we imagine that $u_n(w) > u_a(w)$ at every wealth level.

I make the conventional assumption that utility is increasing in wealth but that marginal utility is decreasing in wealth (the utility functions are upward sloping and strictly concave). Let p denote the probability of an accident. If an accident occurs, the probability that an injurer will be held liable is p_1. The probability that no injurer will be held liable is p_2 (perhaps the injurer

[12] The point has been made by others, including Steven Shavell, Economic Analysis of Accident Law 233–34 (1987). Shavell also notes that there may be a divergence between the optimal damages award for insurance purposes and the optimal award for deterrence purposes. Where, for example, insureds would not desire insurance against nonpecuniary losses, the optimal award for insurance purposes would not include them. Yet because nonpecuniary losses are a real cost of accidents, optimal deterrence may require that injurers bear those costs. There may then arise a case for "decoupling" damages paid from damages received or for using fines paid to the state in lieu of some of the damages paid to victims, particularly if the liability rule is strict liability rather than negligence (the issue does not arise in negligence models for the most part because no one is ever rationally negligent in equilibrium). See *id.* at 247–54. The results here are very much related to Shavell's on these points.

cannot be found, the injurer is not negligent, or the accident is caused by the insured). Thus, $p = p_1 + p_2$.

In the event of an accident, the insured suffers a pecuniary loss equal to L. The initial wealth of the insured is z. The insured can purchase insurance coverage, and the amount of coverage can depend on whether an injurer is liable for the accident or not. Let C_1 be the amount of coverage when an injurer is liable for the accident, and let C_2 be the amount of coverage when no injurer is liable. In this model, I assume that the price of insurance is actuarially fair (that the expected value of the insurance policy net of premiums paid is zero) and will denote the insurance premium by π. I will further assume that the insured cannot purchase coverage in excess of the pecuniary loss L, an assumption that accords with the fact that we do not observe insurers offering "pain and suffering" insurance in practice.[13]

In the state in which an injurer is liable for an accident, the judgment against the injurer entered by a court is J. The judgment presumably exceeds the pecuniary loss L because it contains an element of compensation for the nonpecuniary loss. The injurer may be only partially solvent, however, so the amount collectable from the injurer is only $D \leq J$. Without loss of generality, I assume that these damages are collected by the insured[14] and that the insurer's "share" in accordance with its subrogation right appears as a reduction in C_1. Thus, we can interpret C_1 as net coverage, equal to C_2 less the amount received by the insurer in accordance with its subrogation rights when an injurer is liable. By allowing the insured to select C_1 and C_2 freely, I am implicitly assuming that freedom of contract prevails and that the parties to the insurance contract can choose the optimal amount for the insurer to receive through subrogation.

The insured's optimization problem is then to select coverage levels C_1 and C_2 to maximize expected utility, subject to several constraints. This problem can be written as

$$\max_{C_1, C_2} (1-p)u_n(z-\pi) + p_1 u_a(z-\pi-L+C_1+D) + p_2 u_a(z-\pi-L+C_2),$$

$$\text{subject to } \pi = p_1 C_1 + p_2 C_2,$$

$$C_1, C_2 \leq L,$$

$$C_1, C_2 \geq 0.$$

[13] For an extensive discussion of whether consumers desire to purchase pain and suffering insurance and whether the market is capable of supplying it, see Steven P. Croley & Jon D. Hanson, The Nonpecuniary Costs of Accidents: Pain-and-Suffering Damages in Tort Law, 108 Harv. L. Rev. 1785 (1995).

[14] Thus, my model can be seen as one concerning the optimal level of reimbursement rather than subrogation, but the two are analytically equivalent. The insured can typically collect from both the insurer and the injurer in practice, by the way, because of the "collateral source rule," which holds that an injured party's first-party insurance recovery does not reduce the liability of a tortfeasor.

The first constraint states that the insurance premium is actuarially fair (equal to expected payments by the insurer under the policy), so the insurer breaks even on the contract. The second set of constraints states that coverage cannot exceed the pecuniary loss. The third set of constraints states that coverage is nonnegative.

The problem is a nonlinear programming problem. I give the solution in some detail in the Appendix and simply provide a verbal summary of the key results here. The results turn on how the nonpecuniary loss affects the marginal utility of money. Although those effects might be complex in practice, I will consider three simple cases—where the marginal utility of money remains the same, where it falls uniformly, and where it increases uniformly. Optimal values of the choice variables are denoted with an asterisk.

1. Suppose first that $u_n'(w) \equiv u_a'(w)$. That is, suppose the nonpecuniary loss takes the form of some fixed disutility that does not affect the marginal utility of money. (This is true if we can write $u_n(w) = u_a(w) + K$, where K is a constant.)

In this case, the optimal insurance policy provides full coverage against the pecuniary loss ($C_2^* = L$). Although I assume that insurance in excess of this amount is unavailable, the insured would not wish to purchase it anyway because coverage equal to the pecuniary loss is sufficient to equalize the marginal utility of money across the no-accident and accident states of nature. Furthermore, the optimal policy requires that any money collected from the injurer be paid to the insurer as reimbursement up to the amount of the pecuniary loss. Any excess over that amount goes to the insured. This leaves the insured with the same marginal utility whether or not an injurer is liable for the accident, except where the amount collected from the injurer exceeds the pecuniary loss. In that event, the insured has a lower marginal utility of money when an injurer is liable, but such a situation is optimal given the nonnegativity constraint on coverage ($C_1 \geq 0$).[15]

2. Now suppose that the marginal utility of money in the no-accident state exceeds the marginal utility of money in the accident states for given wealth: $u_n'(w) > u_a'(w)$, for all w. Intuitively, we might imagine that the accident renders the insured less able to enjoy consumption expenditures (holding wealth constant).

In this case, optimal insurance coverage will be less than the full amount of the pecuniary loss and may be zero. The reason is that the reduction in

[15] Some readers may note a connection between this result and the considerable literature on the difference between damages that provide optimal deterrence and damages that provide optimal insurance. Perhaps the leading paper is Steven Shavell, On Liability and Insurance, 13 Bell J. Econ. 120 (1982). Where damages that are optimal for deterrence purposes must exceed damages that are optimal for insurance purposes, some "decoupling" between damages paid by injurers and damages received by victims may be desirable. Litigation costs also factor into such analyses. See, for example, A. Mitchell Polinsky & Yeon-Koo Che, Decoupling Liability: Optimal Incentives for Care and Litigation, 22 Rand J. Econ. 562 (1991).

the marginal utility of money caused by the accident makes it optimal for the insured to consume more wealth in the no-accident state. Where insurance coverage is positive, the optimal amount for the insurer to receive through subrogation is again equal to the collectable damages or the amount of coverage payable when no injurer is liable, whichever is less—this again equalizes the marginal utility of money across the accident states if possible (subject to the nonnegativity constraint). Thus, as in case 1, the optimal contract requires the insurer to be reimbursed fully for its outlays from the recovery against the injurer before the insured receives anything from the injurer. The made-whole doctrine, which imposes the opposite rule, is clearly suboptimal.

3. Finally, suppose that the nonpecuniary loss increases the marginal utility of money, holding wealth constant: $u'_n(w) < u'_a(w)$, for all w. Perhaps the accident leaves the insured less able to enjoy inexpensive activities such as walking and bicycling, and enjoyment must be found in more costly pursuits.

The increase in the marginal utility of money caused by the accident increases the amount of insurance coverage that the insured would like to carry to an amount in excess of the pecuniary loss. On the assumption that the market will not sell such coverage, the insured will do as well as possible by buying coverage equal to the pecuniary loss. In the accident state in which an injurer is liable, the optimal amount payable to the insurer through subrogation may now be less than before—in effect, if the contract allows the insured to keep some of the collectable damages before the insurer is reimbursed, it may partially overcome the problem caused by the refusal of insurers to sell the insured as much insurance as the insured wishes to purchase.

The difference between this case and the prior two is driven by my assumption that the market will not sell coverage exceeding pecuniary losses, even though such coverage is optimal in a first-best sense. Nothing in the model explains the market's resistance to selling such coverage, but factors outside the model may arise in practice to make coverage for nonpecuniary losses unavailable. Within the model, this problem can be partly overcome as noted by modifying the subrogation provision to allow the insured to obtain more coverage than would otherwise be available in the state in which an injurer is liable.

But this result should not be taken too seriously. First, it is unclear how often the phenomenon that necessitates it—an increase in the marginal utility of money because of an accident—will occur. Second, just as factors outside the model must explain why insurers are unwilling to sell coverage in excess of pecuniary losses, so might they argue against allowing the insured to obtain such coverage in a roundabout fashion through a modification of the subrogation rule. Certainly one can wonder whether a court can do better than the parties to the insurance contract in designing the optimal subrogation

arrangement for these cases. Notwithstanding the result here, therefore, I do not regard it as an argument for judicial interference with freedom of contract (more on this issue in the concluding section). The results as a whole for the nonpecuniary loss model suggest that a rule under which the insurer takes first from any damages collected from the injurer—the rule seemingly written into many insurance contracts in practice—may well be optimal much of the time.

II. Actuarially "Unfair" Premiums

When insurance is actuarially fair, a risk-averse insured will prefer to purchase full coverage against pecuniary losses. The reason relates to the fact that regardless of the amount of coverage purchased, the insured's expected wealth will be the same (the insurance policy has a net expected value of zero). Hence, the risk-averse insured can eliminate all risk, and maintain the same expected wealth, by purchasing full coverage—the insured will surely wish to do so because the elimination of risk, holding expected wealth constant, always raises expected utility for the risk averse.

In reality, of course, the expected value of insurance policies must be negative on average. Insurers have administrative expenses that must be covered, and their shareholders wish to earn a reasonable rate of return on investment. Insurance premiums may also exceed actuarially fair levels because insureds are not perfectly separated by risk category, so at times relatively low risk insureds will be pooled with relatively high risk insureds.

When insurance premiums exceed the expected value of payments to the insured under the policy, insureds may rationally purchase less than full coverage against pecuniary losses.[16] When such underinsurance is coupled with a partially solvent injurer, we again confront the question of whether the insurer or the insured should take first from the injurer's assets.

The formal treatment of this case is much the same as in the last section but simpler, and the notation will remain the same except for the following changes: we no longer need to consider nonpecuniary losses, so the utility function is simply $u(w)$. Furthermore, in the absence of nonpecuniary losses, it is unnecessary to distinguish the judgment from the size of the pecuniary loss. I thus assume that the judgment is now L, the value of the pecuniary loss, and that the amount of damages collectable from the injurer is $D \leq L$.

To capture the notion that insurance premiums exceed the actuarially fair level, I will assume that the premium for each dollar of coverage in each

[16] This point can be found in numerous standard microeconomics textbooks. See, for example, David M. Kreps, A Course on Microeconomic Theory 91–93 (1990); Hal R. Varian, Microeconomic Analysis 180–81 (3d ed. 1992).

state carries a "markup" of τ.[17] The constraints $C_1, C_2 \leq L$ can also be dropped, as they will be seen to be nonbinding at the optimum.

The insured's optimization problem is now

$$\max_{C_1, C_2} (1 - p)u(z - \pi) + p_1 u(z - \pi - L + C_1 + D) + p_2 u(z - \pi - L + C_2),$$

$$\text{subject to } \pi = p_1(1 + \tau)C_1 + p_2(1 + \tau)C_2,$$

$$C_1, C_2 \geq 0.$$

I again sketch the details of the solution in the Appendix. In words, the insured purchases less than full coverage against pecuniary loss in this model. Consequently, the insured will not equalize the marginal utility of wealth across all states of nature but will have a higher marginal utility of wealth in the accident states. The reason is that reductions in risk are accompanied by reductions in expected wealth, and the insured must strike a trade-off between the two. At the optimum, the insured will accept some reduction in expected wealth to reduce risk but does not eliminate risk altogether because it is too costly.

The optimal amount for the insurer to receive through subrogation will equal the damages collectable from the injurer or the total amount of first-party coverage payable for the accident, whichever is less (in other words, the optimal contract will call for the insurer to be made whole before the insured).[18] Intuitively, it makes no sense to purchase more net coverage when the injurer is liable than when no injurer is liable. Instead, subject to the constraint that net coverage can never be negative, it is best to equalize the marginal utility of money across the accident states by ensuring that the insured's net wealth is the same in each.

[17] Thus, I assume that insurance is actuarially unfair at the margin. By contrast, if actuarial unfairness arose entirely as a fixed cost per policy that did not vary with the amount of coverage purchased, one can readily show that rational insureds would still purchase full coverage if they bought insurance at all. Then, the underinsurance problem essential to the tension between subrogation and insolvency would not arise. The assumption that insurance premia are actuarially unfair at the margin is a realistic one, I believe. As coverage increases, various costs of claims processing seem likely to increase, such as the costs of investigating the validity of claimed casualty losses and valuing them, the costs of processing claims for medical care, and so on.

[18] A further implicit assumption of the model warrants brief discussion. Note that the markup τ is assumed to be the same for coverage in both states in which loss occurs. One could imagine that administrative costs for the insurance company might be higher in the state in which an injurer is liable, however, because of the costs of pursuing subrogation. The markup in that state, therefore, might be assumed to be higher. I do not treat this case in detail here, but it is worth noting that if the markup for coverage in the state in which an injurer is liable exceeds the markup in the other state, the result that the insurer should be made whole before the insured is strengthened. The intuition is that if the marginal price of coverage when an injurer is liable is greater than the marginal price when no injurer is liable, the amount of coverage that an insured will purchase for the state in which the injurer is liable declines relative to the amount the insured will purchase for the state in which no injurer is liable.

III. Moral Hazard

When insurers cannot observe the precautionary behavior of their insureds or when courts cannot verify precautionary behavior after the fact, it becomes impossible for insurers and insureds to contract for jointly optimal precautions against loss. And when insurance coverage cannot be conditioned on proper precautions against loss, an externality arises (from the insured's perspective) owing to the existence of insurance—the insured (typically) bears the costs of precautions against loss, but the benefits of those precautions are realized by the insurance company to the degree that the loss is insured. The result is an insufficient incentive for the insured to take precautions, known as "moral hazard."

Various responses to the moral hazard problem can be imagined. In some cases, the insurer could bear the cost of the precautions. In others, some observable and verifiable datum may exist (other than information on the loss itself) that is imperfectly correlated with precautionary behavior, on which coverage can be conditioned in whole or in part. Another response (perhaps in conjunction with those above) is to leave some of the loss on the insured to provide some incentive for the insured to take precautions against loss.[19] Where underinsurance arises for the purpose of reducing moral hazard, we again confront the question of whether the insurer or the insured should take first from the assets of a partially solvent injurer. A variant of the model developed above will allow us to address the issue.

I will assume that utility takes the form $v(w, x) = u(w) - m(x)$, where $u(w)$ is a strictly concave function of wealth as before, x is a nonmonetary precautionary action that is continuously measurable, and $m(x)$ is the utility equivalent of that action. The probability of an accident is now $p(x)$, a decreasing, convex function (additional precautions reduce the probability of an accident, but at a decreasing rate). The accident causes a pecuniary loss of L, as before. When an accident occurs, the probability that an injurer is liable is α, and the probability that no injurer is liable is $(1 - \alpha)$. The damages collectable from the injurer are again $D \leq L$. In this model, I assume that insurance premiums are actuarially fair.

The insured will select coverage levels C_1 and C_2 and the precaution level x to maximize expected utility. The "first-best" insurance contract would simply maximize expected utility subject to the constraints that premiums be actuarially fair and that coverages be nonnegative. This first-best optimum

[19] On the moral hazard problem generally and some of the devices for dealing with it, see generally Bengt Holmstrom, Moral Hazard and Observability, 10 Bell J. Econ. (1979); Mark V. Pauley, The Economics of Moral Hazard: Comment, 58 Am. Econ. Rev. 531 (1968); Stephen A. Ross, The Economic Theory of Agency: The Principal's Problem, 63 Am. Econ. Rev. 134 (Papers & Proc. 1973); Steven Shavell, On Moral Hazard and Insurance, 93 Q. J. Econ. 541 (1979); Richard Zeckhauser, Medical Insurance: A Case Study of the Tradeoff between Risk-Spreading and Appropriate Incentives, 2 J. Econ. Theory 10 (1970).

is unattainable in practice, however, because the insured cannot be required by contract to choose the first-best level of precautions. Instead, in choosing precautions, the insured will take the level of coverage and the associated insurance premium as parameters and choose a precaution level that is privately optimal. The fact that the insured will behave "selfishly" in this way adds an additional constraint to the problem, which may now be written

$$\max_{x,\, C_1,\, C_2} [1 - p(x)]u(z - \pi) + \alpha p(x)u(z - \pi - L + C_1 + D)$$

$$+ (1 - \alpha)p(x)u(z - \pi - L + C_2) - m(x),$$

$$\text{subject to } \pi = \alpha p(x)C_1 + (1 - \alpha)p(x)C_2,$$

$$C_1,\, C_2 \geq 0,$$

$$\{\partial Ev(w, x)/\partial x\}\big|_{\pi} = \alpha p'u(z - \pi - L + C_1 + D)$$

$$+ (1 - \alpha)p'u(z - \pi - L + C_2)$$

$$- p'u(z - \pi) - m' = 0.$$

The first constraint states that the insurance premium is actuarially fair. The second constraints are the familiar nonnegativity restrictions on the level of coverage. The third constraint states that the precaution level x must be "incentive compatible"; that is, it must maximize the insured's expected utility taking the level of coverage and the insurance premium as parameters. For simplicity, I assume that the solution entails an "interior" value of x, so additional constraints on its magnitude are unnecessary. I also omit the constraints $C_1,\, C_2 \leq L$ because they will be seen to be nonbinding at the optimum.

Although it is more complicated to derive (see the Appendix), the solution here is almost identical for present purposes to the solution for the model with actuarially unfair premiums. Because of moral hazard, full coverage against pecuniary loss is undesirable. Instead, the insured will bear some loss to induce greater care. The amount of the loss borne by the insured should be the same whether or not an injurer is liable (subject to the nonnegativity constraint on net coverage) to equalize the marginal utility of money to the insured in the accident states. Thus, except where the recovery from the injurer exceeds the amount paid out by the insurer, the amount recovered from the injurer should be paid to the insurer rather than the insured. In all events, the insurer will be made whole before the insured receives any portion of the recovery.

IV. Conclusion and Extensions

In each of the cases considered save one, the optimal insurance contract requires that the insurer be made whole from the assets of the partially solvent

injurer before the insured receives anything. Such a requirement, of course, is the precise opposite of the made-whole doctrine so prevalent in U.S. courts. All of the results are driven by the fact that the optimal amount of loss for the insured to bear does not vary according to whether an injurer happens to be liable for harms to the insured. Thus, if it were optimal for the insured to collect from the injurer and thereby receive more than the insured's first-party insurance benefits when an injurer is liable, it would also be optimal for the insured to have bought more insurance against the contingency that the same injury would occur under circumstances in which no injurer is liable (or has assets to satisfy a judgment).

The one exception considered in the formal analysis arose in the case in which nonpecuniary losses from the injury increased the marginal utility of money, thereby leading the insured to desire coverage in excess of pecuniary losses, but the insurance market was assumed to be unwilling to sell the insured an amount of coverage that exceeds pecuniary losses. Then, if the insured can collect from the partially solvent injurer before the insurer, the insured can in effect increase coverage in some states of nature toward the level that the insured desires but that the market will not supply.

This last result is generalizable to other circumstances in which the market will not supply the "first-best" amount of coverage. For example, imagine a small company with one or two workers. Perhaps the workers would like to obtain disability insurance against the possibility that an injury will leave them unable to work. Yet it is conceivable that disability insurers will not accept business from such small firms (I have no idea whether this is true in practice) because of the adverse-selection problem—they may be afraid that anyone who applies for insurance will represent an exceptionally high risk.[20] Here again, by allowing the insured to collect from the partially solvent injurer ahead of (say) the insured's health insurer, that insurer in effect sells some of the desired disability insurance that the market will not otherwise supply.[21]

In these types of cases, courts could in theory make insureds better off by allowing them to take first from the assets of a partially solvent injurer. Insurance companies would presumably charge them for the privilege through higher premiums, of course, but by hypothesis insureds are willing to pay a competitive price for such coverage and simply cannot obtain it. Furthermore, the problem is one that, by definition, is unlikely to be susceptible to a contractual solution. Insurers will no more make exceptions to their usual

[20] More precisely, the market may have "unraveled" because of insurers' prior experience, which led to losses, premium increases, and the further exodus of low-risk insureds. The classic exposition is Michael Rothschild & Joseph Stiglitz, Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information, 90 Q. J. Econ. 629 (1976). Of course, market solutions to the adverse-selection problem can and do emerge.

[21] This scenario is easy to formalize as well. I will supply the details to any interested readers.

subrogation provisions to make up for their unwillingness to supply more coverage than they will sell the desired coverage in the first place.

To say that courts can improve matters in theory, however, hardly establishes that they can do so in practice. Most fundamental, how can a court identify insureds who should be able to buy more coverage when insurers are unable to identify them? The same information problems that impede insurers must also afflict courts, and indeed courts may be at an added disadvantage because they lack the actuarial experience and savvy of insurers. If subtle subrogation rules might in theory be used to ameliorate a "market failure" elsewhere, therefore, that failure seems to be one that courts cannot hope to fix in practice.

Furthermore, with reference to the existing state of the law, the made-whole doctrine seems seriously at odds with sound policy. The analysis here suggests that the opposite rule—that the insurer should be made whole first—will be best for insureds except when the insured is unable to purchase efficient amounts of insurance coverage. I can imagine no basis for thinking that that situation is the rule rather than the exception.

These conclusions hold whether the subrogation rights in question arise by operation of contract or at common law. Accordingly, it provides support for a default rule at common law that allows insurers to be made whole first, as well as for the enforcement of express contractual provisions that provide for that arrangement.

APPENDIX

I. Nonpecuniary Losses

Substitute the first constraint into the objective function and formulate the Lagrangean:

$$\begin{aligned} L = {} & (1-p)u_n(z - p_1C_1 - p_2C_2) \\ & + p_1u_a(z - p_1C_1 - p_2C_2 - L + C_1 + D) \\ & + p_2u_a(z - p_1C_1 - p_2C_2 - L + C_2) \\ & + \lambda_1(L - C_1) + \lambda_2(L - C_2). \end{aligned}$$

The Kuhn-Tucker necessary (and sufficient[22] in this case) conditions for an optimum can then be written as

$$\begin{aligned} \partial L/\partial C_1 = {} & -p_1(1-p)u_n'(z-\pi) + p_1(1-p_1)u_a'(z-\pi-L+C_1+D) \\ & - p_1p_2u_a'(z-\pi-L+C_2) - \lambda_1 \le 0; \end{aligned} \tag{A1}$$

$$C_1 \ge 0, \qquad C_1(\partial L/\partial C_1) = 0; \tag{A2}$$

[22] The objective function is strictly concave in the choice variables, and the constraints are linear, so the problem satisfies the Kuhn-Tucker sufficiency theorem. See, for example, Alpha C. Chiang, Fundamental Methods of Mathematical Economics 722 (2d ed. 1974).

$$\partial L/\partial C_2 = -p_2(1-p)u_n'(z-\pi) - p_1p_2u_a'(z-\pi-L+C_1+D) + p_2(1-p_2)u_a'(z-\pi-L+C_2) - \lambda_2 \leq 0, \tag{A3}$$

$$C_2 \geq 0, \qquad C_2(\partial L/\partial C_2) = 0, \tag{A4}$$

$$L - C_i \geq 0, \qquad \lambda_i \geq 0, \qquad \lambda_i(L - C_i) = 0. \tag{A5}$$

Conjecture that the solution is interior. With both coverage levels positive, conditions (A1) and (A3) hold with equality. An interior solution implies further that the λ_i terms equal zero. Divide p_1 out of condition (A1) and p_2 out of condition (A3), then subtract (A3) from (A1) to establish that $u_a'(z-\pi-L+C_1+D) = u_a'(z-\pi-L+C_2)$. Substituting this result back into condition (A1) further implies

$$u_n'(z-\pi) = u_a'(z-\pi-L+C_1+D) = u_a'(z-\pi-L+C_2). \tag{A6}$$

In words, if the solution is interior, then the marginal utility of money must be equalized across all states of nature. Were it otherwise, the insured could profit by shifting money from a low marginal utility state to a high marginal utility state through adjustments in coverage levels.

Furthermore, condition (A6) establishes that $C_1^* = C_2^* - D$. In words, the optimal amount for the insurer to receive through subrogation is D, the amount of damages collected from the injurer. This will be true regardless of the value of D—that is, regardless of the degree to which the injurer is insolvent.

Although we have characterized an interior solution, we have not established that the interior solution is feasible (recall that any solution satisfying the necessary conditions is indeed an optimum because the necessary conditions are here sufficient). Its feasibility will depend on the relation between $u_n(w)$ and $u_a(w)$, as well as on the magnitude of D.

1. Suppose that $u_n'(w) \equiv u_a'(w)$. The optimal contract provides $C_2^* = L$ and $C_1^* = \max[L - D, 0]$.

Proof. Condition (A6) implies $C_2^* = L$ (full coverage against pecuniary loss) and $C_1^* = L - D$. For this solution to satisfy nonnegativity, we must in turn have $D \leq L$—the collectable damages are less than the pecuniary loss. When collectable damages exceed the pecuniary loss L, the optimum becomes $C_1^* = 0$ (this follows from the fact that condition (A1) is negative evaluated at $C_1 = 0$ and from the nonnegativity constraint). Q.E.D.

2. Suppose that $u_n'(w) > u_a'(w)$, for all w. The optimal contract here provides $C_2^* < L$ and $C_1^* = \max[C_2^* - D, 0]$. It is possible that $C_2^* = C_1^* = 0$.

Proof. Conjecture once again that an interior solution arises. Condition (A6) again implies that $C_1^* = C_2^* - D$. It further implies that $C_2^* < L$ because utility is strictly concave. If this solution is to be the optimum, the marginal utility of coverage C_2 must be positive at $C_2 = 0$: from condition (A3), one can deduce that $u_n'(z) > u_a'(z - L)$ is necessary. Were it otherwise, the purchase of insurance would transfer wealth from a high marginal utility state to a low marginal utility state. Assuming that $C_2^* > 0$ (subrogation is uninteresting unless some insurance is purchased), then a contract in which $C_1^* = C_2^* - D$ will be feasible if $D < C_2^*$. Otherwise, $C_1^* = 0$ for the same reason as before. Q.E.D.

3. Suppose that $u_n'(w) < u_a'(w)$, for all w. The optimal contract requires $C_2^* = L$ and $C_1^* \in [0, L]$.

Proof. Condition (A6) now requires $u_n'(z-\pi) = u_a'(z-\pi-L+C_2) \Rightarrow C_2 > L$, which is impermissible, so $C_2^* = L$ and $\lambda_2 > 0$. Conditions (A1) and (A3) now imply

$C_1^* > L - D$ or conceivably $C_1^* = L$, depending on the degree to which the marginal utility of money in the accident states exceeds the marginal utility of money in the no-accident state. But if the recovery from the injurer exceeds L by a sufficient amount, it is still possible to have $C_1^* = 0$. Q.E.D.

II. Actuarially Unfair Premiums

The optimal contract in this model will provide that $C_2^* < L$ and $C_1^* = \max[C_2^* - D, 0]$.

Proof (Sketch). The solution procedure is as before. Using the analogs to conditions (A1) and (A3) above (the necessary conditions are once again sufficient), one can derive that, for an interior solution, $u'(z - \pi - L + C_1 + D) = u'(z - \pi - L + C_2)$ or that $C_1^* = C_2^* - D$. This interior solution will be feasible as long as $C_2^* - D \geq 0$. Intuitively, the insured once again equalizes the marginal utility of money in the accident states if it is possible to do so—the optimality of doing so follows from the fact that the markup per dollar of coverage is the same in both states.

Substituting this result into the analog to condition (A1) and rearranging terms yields, for an interior solution,

$$u'(z - \pi - L + C_2)/u'(z - \pi) = (1 - p + \tau - p\tau)/(1 - p - p\tau) > 1.$$

From the concavity of $u(w)$, it follows that $C_2^* < L$ at the optimum. We already know that $C_1^* = C_2^* - D$ as long as the nonnegativity constraint is satisfied. It is trivial to show that $C_1^* = 0$ otherwise. Q.E.D.

III. Moral Hazard

The optimal contract in this model will provide that $C_2^* < L$ and $C_1^* = \max[C_2^* - D, 0]$.

Proof. The Kuhn-Tucker sufficiency theorem is not in general satisfied for this problem, but I will assume that the first-order conditions have a unique solution. Substitute the first constraint into the objective function and formulate the Lagrangean:

$$\begin{aligned}
L = {} & [1 - p(x)]u[z - \alpha p(x)C_1 - (1 - \alpha)p(x)C_2] \\
& + \alpha p(x)u[z - \alpha p(x)C_1 - (1 - \alpha)p(x)C_2 - L + C_1 + D] \\
& + (1 - \alpha)p(x)u[z - \alpha p(x)C_1 - (1 - \alpha)p(x)C_2 - L + C_1 + D] - m(x) \\
& + \lambda\{\alpha p'u[z - \alpha p(x)C_1 - (1 - \alpha)p(x)C_2 - L + C_1 + D] \\
& \quad + (1 - \alpha)p'u[z - \alpha p(x)C_1 - (1 - \alpha)p(x)C_2 - L + C_1 + D] \\
& \quad - p'u[z - \alpha p(x)C_1 - (1 - \alpha)p(x)C_2] - m'\}.
\end{aligned}$$

It will suffice to write the analogs to conditions (A1) and (A3) above (replacing the expression for π for notational simplicity):

$$\begin{aligned}
\partial L/\partial C_1 = {} & -(1 - p)\alpha p u'(z - \pi) + \alpha p(1 - \alpha p)u'(z - \pi - L + C_1 + D) \\
& - (1 - \alpha p)\alpha p u'(z - \pi - L + C_2) \\
& + \lambda[\alpha p'(1 - \alpha p)u'(z - \pi - L + C_1 + D) \qquad \text{(A1')} \\
& \quad - (1 - \alpha)p'(\alpha p)u'(z - \pi - L + C_2) \\
& \quad + p'\alpha p u'(z - \pi)] \leq 0
\end{aligned}$$

and

$$\begin{aligned}\partial L/\partial C_2 = &-[1-p](1-\alpha)pu'(z-\pi) - \alpha p(1-\alpha)pu'(z-\pi-L+C_1+D)\\ &+ (1-\alpha)p[1-(1-\alpha)p]u'(z-\pi-L+C_2)\\ &+ \lambda\{-\alpha p'(1-\alpha)pu'(z-\pi-L+C_1+D) \qquad \text{(A3')}\\ &\quad + (1-\alpha)p'[1-(1-\alpha)p]u'(z-\pi-L+C_2)\\ &\quad + p'(1-\alpha)pu'(z-\pi)\} \leq 0.\end{aligned}$$

Conjecture an interior solution. Divide (A1′) by αp and (A3′) by $(1-\alpha)p$, then subtract (A3′) from (A1′). After considerable algebra, one obtains

$$\{u'(z-\pi-L+C_1+D) - u'(z-\pi-L+C_2)\}[1+\lambda p'/p] = 0,$$

which implies $u'(z-\pi-L+C_1+D) = u'(z-\pi-L+C_2)$. Again, unless the nonnegativity constraint on C_1 prevents it, the insured will want to equalize marginal utility across the accident states, and $C_1^* = C_2^* - D$.

Substituting this result into (A1′) yields, again after considerable algebra,

$$\begin{aligned}&(1-p)[u'(z-\pi-L+C_1+D) - u'(z-\pi)]\\ &+ (\lambda p'/\alpha p)\{(1-p)[\alpha(1-\alpha)]u'(z-\pi-L+C_1+D)\\ &\qquad + \alpha pu'(z-\pi-L+C_2)\} = 0.\end{aligned}$$

The second term is negative, implying that the first term is positive and, hence,

$$u'(z-\pi-L+C_1+D) > u'(z-\pi).$$

In words, the marginal utility of money in the accident state in which an injurer is liable (equal to the marginal utility in the other accident state for an interior solution) must exceed the marginal utility of money in the no-accident state. Using the concavity of $u(w)$, we can deduce that the wealth of the insured is lower in the accident states. Thus, for the interior solution, we have $C_1^* + D = C_2^* < L$.

This contract is feasible as long as it obeys the nonnegativity constraint on C_1. Otherwise (and I omit the details of the derivation), we have $C_1^* = 0$. Q.E.D.