Insuring Uninsurable Risks from AI: A Market-Based Solution

Anonymous Authors[1]

1. Background

Many experts believe AI systems will, sooner or later, pose uninsurable risks, including existential risks (Grace et al., 2024; Bengio et al., 2024). If so, it will be impossible to hold accountable the parties liable for such harms (or their insurers).

Weil (2024) proposes to solve this extreme judgment proof-problem by assigning punitive damages to harms that are *correlated* with uninsurable risks (where the correlation would be estimated by courts and juries). While of interest, this solution has several problems. First, is it's novelty: this would be an unprecedented application of punitive damages that may violate the Due Process Clause (2024, 40-44, 50-53), requiring a major doctrinal shift that would cut across all of tort law. Second, correlates of uninsurable risks might be difficult to find. Third, given the high uncertainty involved, correlation estimations by courts will likely be *ad hoc*, high variance, and fail to leverage all available information. Fourth and finally, punitive damages for correlated risks will send a very oblique and noisy signal to liable parties: its effectiveness at actually inducing greater care taken is doubtful. Liable parties might find powerful legal teams to be a safer investment than investments in safety.

Historically, the solution to uninsurable (albeit, non-existential risks) has been for government to step into its role as insurer of last resort (Moss, 2004), as seen in government provided reinsurance for terrorism risk insurance (Federal Insurance Office, 2022) or indemnification schemes for nuclear power operators (Commission, 2021, sec. 3.2). Such programs can be in the government's interest for several reasons. First, by creating a more predictable legal environment and making insurance more affordable, they spur the economy in the short-term while protecting it against future shocks by increasing insurance uptake (2021, sec. 1.1)(Michel-Kerjan & Pedell, 2006, 6, 7)(Hubbard et al., 2005, 178). Second, by encouraging or mandating *ex ante* contributions from the private sector, governments can lower their financial exposure to risk (Carroll et al., 2004). Governments cannot credibly commit to *not* bail out a critical economic sector or *not* provide relief to victims in a major disaster: the government is always implicitly exposed to such risk.[1] Third, governments might mitigate the moral hazard it generates as implicit insurer (or lender) of last resort (see e.g. the "Too Big To Fail" effect (Strahan, 2013)).

In the context of solving the judgment-proof problem, this last reason is the most interesting. While such programs might reduce moral hazard over the baseline (no program), they can only have done so in a crude manner due to their crude pricing.[2]

Risk-based pricing is wanted. A government agency could make risk estimates, but this would be costly and the agency would struggle to collapse the information asymmetry between itself and the well-resourced private actors it insures. This paper proposes another solution, leveraging advances in mechanism design: a survey mechanism, the Bayesian Truth Serum (Prelec, 2004), could be used to reliably extract and aggregate honest risk estimates from experts, *including the parties insured*. This, it's argued, better leverages all available information than other solutions.

It's further argued that if accurate risk-based pricing can be had, a government indemnification program is preferable over Weil's punitive damage regime for producing less litigation and more robustly signaling to insureds of what risks they must mitigate. From the insured's perspective, it would also more consistently transform large and uncertain *ex post* costs into manageable and certain *ex ante* costs.

Finally, this paper proposes using a contribution matching mechanism, Quadratic Financing (Buterin et al., 2019), to redistribute collected fees *back* to industry, funding the research required to reduce its uninsurable risks.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]This financial reason, based on *ex post* costs to governments, will not apply to existential risks: governments are also judgment-proof in the face of such risks.

[2]For example, in the commercial nuclear power case, indemnity fees were charged per plant and set by simply multiplying the maximum power output of said plant by a fixed multipliers: this does not encourage operators to take greater care along any dimension other than their choice of maximum power output in the initial design stage. (Commission, 2021, sec. 3.2)

2. A Risk-Priced Indemnification Program

An indemnification program is preferred over reinsurance as this removes the intermediary of insurers, allowing the government to directly manipulate incentives of risk-generating parties. Elsewhere I've argued that it's developers who should be strictly and exclusively liable for said risks (largely based on their being least-cost avoiders(Author, 2024)).

Participation would be mandatory for developers of AI models trained over a certain effective compute[3] threshold. The core of the program is a risk-priced indemnity fee that developers must pay per training run. It's recommended the fee be a function of effective compute.[4]

A government agency would estimate the disutility of various disaster scenarios, but risk-estimation would rely on a survey of public and private experts, *including indemnified developers*. The Bayesian Truth Serum (BTS) is employed to incentivize *effortful* and *honest* risk-estimations from respondents (Prelec, 2004). BTS rewards responses with high information scores – high "surprisingness" relative to respondents' predictions of how other respondents will respond. Scaling the BTS payout incents greater effort in information gathering. Honest reporting is a Bayes-Nash equilibrium under BTS. I.e. absent other incentives, a respondent will maximize their expected payout by reporting honestly *if* they believe a large enough majority of other respondents will also report honestly.

A developer (who must pay the fee) obviously has a conflicting incentive to lie (underreport the risks), and can expect other developers to lie. This conflict of interest can be overcome by dramatically scaling BTS' payout or removed by silencing a developer's risk-estimation when their *individualized* fee is calculated.[5] An expectation of overwhelming honesty can be created by ensuring the vast majority of respondents are *not* developers but instead independent experts with no conflicts of interest. Government agencies and relevant insurers could also be respondents.

The survey should be run at regular intervals (e.g. yearly), with the fee scale fixed for that interval. Any developer who wants to train during the current window must have participated in the last survey.[6] The government chosen survey questions[7] set the safety agenda, clearly signaling to indemnifieds what risks they must mitigate to lower their fees.

Using BTS to solve the information asymmetry between the government and indemnifieds has several advantages over relying heavily on government risk estimates and inspecting indemnifieds. It should be cheaper, incenting parties to compete to provide the most informative risk-estimates, all while more reliably aggregating a wider range of private information. It should also be more secure – developers can divulge the *risk implications* of their private information *without* exposing security-sensitive information. Finally, it creates a more cooperative relationship between developers and the government, lending the regime greater legitimacy.

3. Quadratic Financing for Safety Research

It's recommended the revenue from the indemnification scheme be used to fund AI related programs in the public's interest. One such program should aim to directly help developers shoulder the cost of the Safety Research (SR) they need to reduce their fees.

Because SR is a public good, developers face a coordination problem.[8] To help them solve their coordination problem, ensure an optimal supply of SR, and contribute its own fair share of funds, it's recommended the government employ a fund-matching mechanism, Quadratic Finance (QF), designed to achieve all the above (Buterin et al., 2019).

Under QF, developers and the broader public would propose various SR projects. Developers then choose to fund to whatever extent whichever projects it likes, knowing the government will top-up a project's total funds according the QF formula. This top-up scales quadratically in the number of contributors to a project. (As with BTS, QF will require basic defenses against collusion and fraud.)

Projects would essentially be competing for private contributions, signaling where to send public funds. Because of the agenda setting achieved by the liability and indemnification scheme, projects would require minimal vetting. The market then determines which projects achieve that agenda most effectively.

4. Further Research

While confident in the theoretical soundness of its claims, the paper acknowledges the need for further empirical research into the effectiveness of BTS and QF. Available studies align with expectations (Weaver & Prelec, 2013; Pasquini, 2022), but more tests are needed.

[3] Effective compute = FLOPs * an algorithmic efficiency factor.

[4] Cf. the industry's currently voluntary "responsible scaling policies" e.g. (Anthropic, 2023).

[5] This second option puts developers in a prisoner's dilemma: they could lower their fees by coordinating, but it's individually rational to defect, increasing fees for one's competitors.

[6] A small discount on the fee should also be offered for having participated in the last several surveys.

[7] E.g. "For an AI trained on compute x, what's the likelihood of disaster D within time frame t?"

[8] Note that the coordination problem for supplying SR is greatly simplified by the liability and indemnification regime: instead of countless potential victims needing to coordinate, only *developers* need coordinate.

References

Anthropic. Anthropic's Responsible Scaling Policy, 2023. URL https://www.anthropic.com/news/anthropics-responsible-scaling-policy.

Author, N. N. Suppressed for Anonymity. In *Under Submission*, 2024.

Bengio, Y., Hinton, G., Yao, A., Song, D., Abbeel, P., Darrell, T., Harari, Y. N., Zhang, Y.-Q., Xue, L., Shalev-Shwartz, S., Hadfield, G., Clune, J., Maharaj, T., Hutter, F., Baydin, A. G., McIlraith, S., Gao, Q., Acharya, A., Krueger, D., Dragan, A., Torr, P., Russell, S., Kahneman, D., Brauner, J., and Mindermann, S. Managing extreme AI risks amid rapid progress. *Science*, 384 (6698):842–845, May 2024. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.adn0117. URL http://arxiv.org/abs/2310.17688. arXiv:2310.17688 [cs].

Buterin, V., Hitzig, Z., and Weyl, E. G. A Flexible Design for Funding Public Goods. *Management Science*, 65(11): 5171–5187, November 2019. ISSN 0025-1909, 1526-5501. doi: 10.1287/mnsc.2019.3337. URL http://arxiv.org/abs/1809.06421. arXiv:1809.06421 [econ, q-fin].

Carroll, S. J., LaTourrette, T., Chow, B. G., Jones, G. S., and Martin, C. Assessing the Effectiveness of the Terrorism Risk Insurance Act. Technical report, RAND Corporation, January 2004. URL https://www.rand.org/pubs/research_briefs/RB9153.html.

Commission, U. S. N. R. The Price-Anderson Act: 2021 Report to Congress, Public Liability Insurance and Indemnity Requirements for an Evolving Commercial Nuclear Industry. 2021. URL https://www.nrc.gov/docs/ML2133/ML21335A064.pdf.

Federal Insurance Office, U. D. o. t. T. Report on the Effectiveness of the Terrorism Risk Insurance Program. Technical report, June 2022. URL https://home.treasury.gov/system/files/311/2022%20Program%20Effectiveness%20Report%20%28FINAL%29.pdf.

Grace, K., Stewart, H., Sandkühler, J. F., Thomas, S., Weinstein-Raun, B., and Brauner, J. Thousands of AI Authors on the Future of AI, April 2024. URL http://arxiv.org/abs/2401.02843. arXiv:2401.02843 [cs].

Hubbard, R. G., Deal, B., and Hess, P. The Economic Effects of Federal Participation in Terrorism Risk. *Risk Management and Insurance Review*, 8(2):177–209, 2005. ISSN 1540-6296. doi: 10.1111/j.1540-6296.2005.00056.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6296.2005.00056.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-6296.2005.00056.x.

Michel-Kerjan, E. and Pedell, B. How Does the Corporate World Cope with Mega-Terrorism? Puzzling Evidence from Terrorism Insurance Markets. *Journal of Applied Corporate Finance*, 18(4):61–75, 2006. ISSN 1745-6622. doi: 10.1111/j.1745-6622.2006.00112.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1745-6622.2006.00112.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-6622.2006.00112.x.

Moss, D. A. *When All Else Fails: Government as the Ultimate Risk Manager*. Harvard University Press, October 2004. ISBN 978-0-674-01609-5. Google-Books-ID: PEdOEAAAQBAJ.

Pasquini, R. A. Quadratic Funding and Matching Funds Requirements, July 2022. URL http://arxiv.org/abs/2010.01193. arXiv:2010.01193 [econ, q-fin].

Prelec, D. A Bayesian Truth Serum for Subjective Data. *Science*, 306(5695):462–466, October 2004. doi: 10.1126/science.1102081. URL https://www.science.org/doi/abs/10.1126/science.1102081. Publisher: American Association for the Advancement of Science.

Strahan, P. E. Too Big to Fail: Causes, Consequences, and Policy Responses. *Annual Review of Financial Economics*, 5(Volume 5, 2013):43–61, November 2013. ISSN 1941-1367, 1941-1375. doi: 10.1146/annurev-financial-110112-121025. URL https://www.annualreviews.org/content/journals/10.1146/annurev-financial-110112-121025. Publisher: Annual Reviews.

Weaver, R. and Prelec, D. Creating Truth-Telling Incentives with the Bayesian Truth Serum. *Journal of Marketing Research*, 50(3):289–302, June 2013. ISSN 0022-2437. doi: 10.1509/jmr.09.0039. URL https://doi.org/10.1509/jmr.09.0039. Publisher: SAGE Publications Inc.

Weil, G. Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence, January 2024. URL https://papers.ssrn.com/abstract=4694006.