Legal Priorities
Project

Ordinary meaning of existential risk

Eric Martínez
Christoph Winter

LPP WORKING PAPER N° 7-2022

LPP Working Paper No. 7-2022 **Very preliminary draft**
Last update: December 2022
Comments welcome: ericmart@mit.edu, christoph_winter@fas.harvard.edu

Ordinary Meaning of Existential Risk

Eric Martínez[1] & Christoph Winter[2]

Abstract

Recent legislative efforts across multiple jurisdictions aim at regulating existential risk, a new category of extreme risks that pose a greater threat to humanity's future than any previous risk. This article investigates the ordinary meaning of legally relevant concepts in the existential risk literature. Four experiments (n=6,814) reveal that the ordinary meaning of "existential risk": (a) like the technical meaning, is narrower than other related terms, such as "global catastrophic risk" and "extreme risk"; (b) is mostly unaffected by exposure to definitions, except those containing probability thresholds; (c) mostly, though not entirely, resembles an expected harm calculation; and (d) differs widely between abstract and concrete scenarios but not across concrete risk type (such as climate change, pandemics, and nuclear war). These results provide crucial insights for those tasked with drafting and interpreting existential risk laws, and for the coherence of ordinary meaning analysis more generally. This study also lays the foundation for a new research program we refer to as "*ex ante* ordinary meaning analysis"—focused not only on how judges can and should interpret legal provisions once they have been drafted, but on how lawmakers can and should draft legal provisions so as to best achieve their policy aims.

Keywords

Experimental jurisprudence; existential risk; global catastrophic risk; ordinary meaning analysis

[1] Massachusetts Institute of Technology (MIT), Cambridge, MA, USA; Legal Priorities Project, Cambridge, MA, USA. Email: ericmart@mit.edu.

[2] Instituto Tecnológico Autónomo de México (ITAM), Mexico City, Mexico; Harvard University, Cambridge, MA, USA; Legal Priorities Project, Cambridge, MA, USA. Email: christoph_winter@fas.harvard.edu.

For comments, discussion, and critique, we are grateful to Matthijs Maas and José Jaime Villalobos. We are particularly grateful to Suzanne Van Arsdale for excellent research assistance and many useful suggestions.

"Member States and stakeholders clearly expressed that a Declaration for Future Generations should state a firm commitment to securing the interests of future generations in all decision making; by identifying, managing and monitoring global existential risks ..."

United Nations, Elements Paper for the
Declaration for Future Generations (2022)[3]

I. Introduction

Recent scholarship in the emerging field of existential risk studies has identified risks that pose greater threats to the future of humanity than any faced before.[4] Few legal protections address these risks—which include climate change[5], nuclear war[6], pandemics[7], and artificial intelligence[8]—thereby endangering both present and future

[3] Permanent Representatives of the Netherlands and Fiji to the United Nations, 'Elements Paper for the Declaration for Future Generations' (9 September 2022), United Nations General Assembly, https://www.un.org/pga/76/wp-content/uploads/sites/101/2022/09/Elements-Paper-Declaration-for-Future-Generations-09092022.pdf.

[4] Much of this research is synthesized in T. Ord, *The Precipice: Existential Risk and the Future of Humanity* (New York: Hachette Books, 2020).

[5] See eg P. U. Clark et al., 'Consequences of Twenty-First-Century Policy for Multi-Millennial Climate and Sea-Level Change' (2016) 6 *Nature Climate Change* 360; S. J. Beard et al., 'Assessing Climate Change's Contribution to Global Catastrophic Risk' (2021) 127 *Futures* 102673; C. E. Richards, R. C. Lupton, and J. M. Allwood, 'Re-Framing the Threat of Global Warming: An Empirical Causal Loop Diagram of Climate Change, Food Insecurity and Societal Collapse' (2021) 164 *Climatic Change* 49; P. Kareiva and V. Carranza, 'Existential Risk Due to Ecosystem Collapse: Nature Strikes Back' (2018) 102 *Futures* 39; J. Halstead (2022), Climate Change and Longtermism, Supplementary Material in W. MacAskill, *What We Owe The Future* (Basic Books, 2022), https://drive.google.com/file/d/14od25qdb4sdDoXVDMoiSrTwuzYAMSpxK/view. Christian Huggel et al., 'The Existential Risk Space of Climate Change' (2022) 174(1) *Climatic Change* 8; L. Kemp et al., 'Climate Endgame: Exploring Catastrophic Climate Change Scenarios' (2022) 119(34) PNAS e2108146119.

[6] See eg A. Witze, 'How a Small Nuclear War Would Transform the Entire Planet' (2020) 579 *Nature* 485; R. P. Turco et al., 'Nuclear Winter: Global Consequences of Multiple Nuclear Explosions' (1983) 222(4630) *Science* 1283; J. Jägermeyr et al., 'A Regional Nuclear Conflict Would Compromise Global Food Security' (2020) 117(13) PNAS 7071.

[7] See eg M. Schoch-Spana et al., 'Global Catastrophic Biological Risks: Toward a Working Definition' (2017) 15(4) *Health Security* 323; K. Esvelt, 'Inoculating Science Against Potential Pandemics and Information Hazards' (2018) 14(1) *PLOS Pathogens* e1007286.

[8] See eg K. Vold and D. R. Harris, 'How Does Artificial Intelligence Pose an Existential Risk?' in C. Véliz (ed), *The Oxford Handbook of Digital Ethics* (OUP, 2021); S. J. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (Penguin Books, 2019); N. Bostrom, *Superintelligence: Paths, Dangers, Strategies* (OUP, 2014); B. Christian, *The Alignment Problem: Machine Learning and Human Values* (W. W. Norton & Company, 2020); E. Yudkowsky, 'Artificial Intelligence as Positive or Negative Factor in Global Risk' in N. Bostrom and M. M. Ćirković (eds), *Global Catastrophic Risks* (New York: OUP, 2008); R. Ngo, 'AGI Safety From First Principles' (September 2020) at https://drive.google.com/file/d/1uK7NhdSKprQKZnRjU58X7NLA1auXlWHt/view (last accessed 7 December 2022); R. Ngo, 'The Alignment Problem From a Deep Learning Perspective' (30 August 2022) at https://arxiv.org/abs/2209.00626 (last accessed 7 December 2022); D. Hendrycks and M. Mazeika, 'X-Risk Analysis for AI Research' (20 September 2022) at https://arxiv.org/abs/2206.05862 (last accessed 7

generations. Calls to fill this gap have resulted in draft legal provisions that would cover scenarios involving some of these risks, including at the United Nations,[9] United States of America,[10] and the United Kingdom.[11] In addition to draft legal provisions, several prominent legal scholars have drawn attention to the importance of mitigating catastrophic risks through legal action.[12] Nevertheless, it remains an open question how a legal provision ought to be drafted to ensure effective protection against such risks. Or, in the words of the Secretary-General of the United Nations when discussing relevant lawmaking efforts in Our Common Agenda (2021): "An effort is warranted to better define ... the extreme, catastrophic and existential risks that we face."[13]

While the answer to this question may vary by jurisdiction, similar methods may prove useful across most, if not all, jurisdictions. Around the world's jurisdictions, judges (whether by tradition or by law) tend to interpret words in a legal provision according to their ordinary meaning.[14] Although there is some debate as to what ordinary meaning actually means, jurists generally agree that it refers to how a typical or reasonable person

December 2022); J. Carlsmith, 'Is Power-Seeking AI an Existential Risk?' (16 June 2022) at https://arxiv.org/abs/2206.13353 (last accessed 7 December 2022).

[9] United Nations Secretary-General, Our Common Agenda (10 September 2021), 64; Permanent Representatives of the Netherlands and Fiji to the United Nations, n 3 above.

[10] Global Catastrophic Risk Management Act of 2022, S. 4488, 117th Cong. (2022) at https://www.congress.gov/bill/117th-congress/senate-bill/4488/text; U.S. Senate Committee on Homeland Security & Governmental Affairs, 'Portman, Peters Introduce Bipartisan Bill to Ensure Federal Government is Prepared for Catastrophic Risks to National Security' (24 June 2022) Minority Media, at https://www.hsgac.senate.gov/media/minority-media/portman-peters-introduce-bipartisan-bill-to-ensure-federal-government-is-prepared-for-catastrophic-risks-to-national-security- (last accessed 7 December 2022).

[11] UK Cabinet Office, 'The National Resilience Strategy: A Call for Evidence' (2021), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1001404/Resilience_Strategy_-_Call_for_Evidence.pdf (last accessed 7 December 2022), 17, 18; see also House of Lords Select Committee on Risk Assessment and Risk Planning, 'Preparing for Extreme Risks: Building a Resilient Society' (3 December 2021) HL Paper 110, https://publications.parliament.uk/pa/ld5802/ldselect/ldrisk/110/110.pdf (last accessed 7 December 2022); for an overview of ongoing (legal) advocacy against existential risk, see J. Bliss, 'Existential Advocacy' (LPP Working Paper No. 4-2022) (10 September 2022) at https://dx.doi.org/10.2139/ssrn.4217687 (last accessed 7 December 2022).

[12] See eg R. A. Posner, 'Catastrophe: Risk and Response' (Oxford: OUP, 2004); C. R. Sunstein, 'Irreversible and Catastrophic,' (2006) 91 *Cornell Law Review* 841; C. R. Sunstein, 'The Catastrophic Harm Precautionary Principle' (2007) 6 *Issues in Legal Scholarship* 1, 3.

[13] United Nations Secretary-General, Our Common Agenda (10 September 2021), 64.

[14] Examples of jurisdictions that explicitly employ some version of ordinary meaning analysis include Australia (eg *Electricity Generation Corporation* v *Woodside Energy*, 2014 HCA 7), the United Kingdom (*River Wear Commissioners* v. *Adamson*, 2 App Cas 742, 1877), South Africa (*Venter* v. *R*, 1907 TS 910; Terrence R. Carney, 'Legal Fallacy? Testing the Ordinariness of "Ordinary Meaning"' (2016) 137 *South African Law Journal* 269), and the United States (see generally Brian G. Slocum, *Ordinary Meaning: A Theory of the Most Fundamental Principle of Legal Interpretation* (Chicago: University of Chicago Press, 2016)), and Singapore (Interpretation Act Sec. 9A, 1993), as well as international law (Vienna Convention on the Law of Treaties art. 31, 1969). Ordinary meaning has also been found to be relevant in civil-code jurisdictions, including Argentina, Finland, France, Germany, Italy, Poland, and Sweden (see generally D. N. MacCormick and R. S. Summers, *Interpreting Statutes: A Comparative Study* (London: Routledge, 2016).

understands and uses a given word or concept.[15] Thus, investigating how people understand and use “existential risk” and related terms would be directly informative of the ordinary meaning of existential risk, and by extension, would provide insight into how judges might interpret a legal provision containing these terms. This would, in turn, help lawmakers choose which term, definition, and examples to include in provisions to guard against specific risks so as to maximize the chance that these laws are interpreted as intended.

For example, suppose a lawmaker wants to design a statute that requires the government to spend 1% of annual GDP on reducing existential risk to humanity.[16] In particular, the lawmaker has in mind a narrow definition of existential risk, in order to avoid government spending on things that only tangentially relate to existential risk, such as general military defense. If it turns out that the ordinary meaning of “existential risk” is broader[17] than alternatives, such as “global catastrophic risk” or “extreme risk”, then according to the lawmaker’s own aims, *ceteris paribus*, they should prefer one of the alternatives to serve as the wording of the provision.

Here, across several studies, we investigated the ordinary meaning of existential risk and related terms. In Study 1a, we investigated how laypeople interpret the term “existential risk” relative to other terms referenced in the associated scientific and legal literature as well as in current legislative proposals, finding that the ordinary meaning of existential risk (like the technical meaning) is narrower than that of related terms. In Study 1b, we investigated how laypeople’s interpretation of existential risk is affected by their being provided definitions of the term and illustrative examples of potential threats, finding that interpretations of existential risk are mostly unaffected. In Study 1c, we investigated how laypeople’s interpretation of existential risk is affected by definitions that specify a probability threshold, finding that while such thresholds do make a difference in their judgments, laypeople are still reluctant to consider low-probability

[15] See generally B. G. Slocum, *Ordinary Meaning: A Theory of the Most Fundamental Principle of Legal Interpretation* (Chicago: University of Chicago Press, 2015); W. N. Eskridge, Jr., *Interpreting Law: A Primer on How to Read Statutes and the Constitution* 33-35 (Paul, MN: Foundation Press, 2016); A. Scalia & B. Garner, *Reading Law: The Interpretation of Legal Texts*, ch 6 (St Paul, MN: Thomson/West 2012); L. M. Solan, *The Language of Statutes: Laws and Their Interpretation*, ch 3 (Chicago: University of Chicago Press, 2010). Some empirical studies have sought to test ordinary meaning with this understanding. See eg K. P. Tobia, ‘Testing Ordinary Meaning’ (2020) 134 *Harvard Law Review* 726; T. R. Lee and S. C. Mouritsen, ‘Judging Ordinary Meaning’ (2018), 127 *Yale Law Journal* 788; S. Klapper, S. Schmidt, and T. Tarantola, ‘Ordinary Meaning from Ordinary People’ (forthcoming).

[16] In a recent survey, legal academics rated this policy as granting the most protection to future generations, if incorporated into a country's constitution, across several options that also included protection against discrimination, legal standing (locus standi), a commission or ombudsperson, and a state goal to protect future generations. E. Martínez and C. Winter, ‘Protecting Future Generations: A Global Survey of Legal Academics’ (LPP Working Paper No. 1-2021) (10 September 2022) at https://dx.doi.org/10.2139/ssrn.3931304 (last accessed 7 December 2022), 31-33.

[17] Note that by “broader” we refer to terms that would be interpreted as having a lower and less restrictive threshold of harm to qualify, and accordingly cover a wider set of scenarios, including those that, if they occurred, would inflict lower levels of harm.

scenarios as constituting an existential risk (at least in the abstract), even when a legal provision explicitly states that existential risk includes very low-probability scenarios.

In Study 2, we first investigated whether laypeople's interpretation of existential risk differs depending on the type of scenario presented, finding that people generally have stable interpretations of the probability of and number of lives at risk needed for something to constitute an existential risk across different scenario types, except that they are slightly more likely to consider a scenario to be an existential risk if it pertains to climate change (as compared to artificial intelligence or pandemics). We further investigated whether people's evaluation of existential risk deviate from an expected value calculation, finding that people's judgments of whether a particular scenario constitutes an existential risk are sensitive to both the expected amount of harm from the scenario as well as the total number of lives threatened, suggesting that people's judgments roughly (though not entirely) follow an expected value calculation.

Taken together, these findings provide not only critical information for lawmakers drafting existential risk legislation, but also new insight into the coherence and justification of the ordinary meaning principle more generally. Moreover, our study lays the foundation for a potential new research program we refer to as "ex ante ordinary meaning analysis"—focused not only on how judges can and should interpret legal provisions once they are drafted, but also on how lawmakers can draft a legal provision using words that will best guide judges (and the public) into recovering their intended meaning and legislative aims.

II. Study 1: Terms and Definitions for Existential Risk

Study 1 was split up into three sub-studies that investigated how participants understood different terms and provisions for existential risk, in terms of the number of lives endangered and the probability of occurrence. In Study 1a, we investigated how people interpreted the term "existential risk" relative to other terms referenced in the associated scientific and legal literature as well as in current legislative proposals. In Study 1b, we investigated how people interpreted the term "existential risk" with or without different definitions used in the literature and in legislative proposals and with or without a list of example threats. In Study 1c, we further investigated how people interpreted definitions that each specified a different probability of risk involved. Here we describe the methods and results of each of these studies in turn.

A. Study 1a: Terms for Risk

1. Materials

To investigate how people interpret different terms for "existential risk," we constructed several versions of a short questionnaire, which asked participants to (a) read a short legal provision referencing a term such as "existential risk," (b) estimate the minimum number of lives that have to be endangered for a given scenario to fall under the confines of the provision, and (c) estimate the minimum probability that those lives will be endangered for the provision to apply to that scenario.[18] The wording of the scenario was as follows:

> Imagine a legal provision that requires governments to protect against "[risk]". Suppose the human population currently stands at 8 billion (8,000,000,000) people.

There were 11 versions of the questionnaire, each with one of the following terms in place of [risk]:

1. Existential risks to humanity
2. Existential risks
3. Extreme risks
4. Global catastrophic risks
5. Global collapse
6. Global disasters
7. Global existential catastrophes
8. Global existential risks
9. High-impact, low-probability risks
10. Large-scale risks
11. Risks of irreversible damage[19]

[18] All data, materials and analysis code are available at the following repository link: https://www.google.com/url?q=https://osf.io/96dzq/?view_only%3D3cfa5b40aa3940c6a766bd9892c82d39&sa=D&source=docs&ust=1670987900556730&usg=AOvVaw10TlF3ysWsnDRkY9LIvupm.

[19] As discussed above, these terms were selected based on various suggestions in the academic literature as well as existing legislation, international law and current policy discourse.

For "existential risks to humanity," see United Nations Secretary-General, Our Common Agenda (10 September 2021), 27 ("existential risk to humanity"); M. Boyd and N. Wilson, 'Existential Risks to Humanity Should Concern International Policymakers and More Could Be Done in Considering Them at the International Governance Level' (2020) 40(11) *Risk Analysis* 2303; J. Ginns, 'Policy Proposals: Risk Management: The Opportunity to Transform the UK's Resilience to Extreme Risks' (September 2022), Centre for Long-Term Resilience, 3 ("Existential Risks to Humanity");

for "existential risks," see B. Tonn and D. Stiefel, 'Evaluating Methods for Estimating Existential Risks' (2013) 33(10) *Risk Analysis* 1772, 1785 ("Our focus on existential risk is motivated by our ultimate goals of estimating the risk of human extinction …");

The two questions were presented as follows:

> What is the minimum number of lives that have to be endangered by a particular risk for something to constitute a “[risk]” according to this provision?

for “extreme risks,” see T. Ord, A. Mercer, and S. Dannreuther, ‘Future Proof: The Opportunity to Transform the UK’s Resilience to Extreme Risks’ (June 2021), 9 (“These threats to humanity—which we in this report refer to as ‘extreme risks’—define our time.”); M. Rees, *Our Final Century* (London: William Heinemann, 2003); S. Baum and A. M. Barrett, ‘Global Catastrophes: The Most Extreme Risks’, in V. Bier (ed), *Risks in Extreme Environments: Preparing, Avoiding, Mitigating and Managing* (New York: Routledge, 2018), 174 (“... manage these most extreme risks and keep human civilization safe.”); J. Ginns, ‘Policy Proposals: Risk Management: The Opportunity to Transform the UK’s Resilience to Extreme Risks’ (September 2022), Centre for Long-Term Resilience, 3;

for ”global catastrophic risks,” see N. Bostrom and M. M. Ćirković (eds), *Global Catastrophic Risks* (OUP, 2008), 2 (“global catastrophic risks facing humanity”); S. Avin et al., ‘Classifying Global Catastrophic Risks’ (2018) 102 *Futures* 20; Tonn and Stiefel, n 19 above, 1785 (“global catastrophic risks ... could kill many millions of humans”);

for “global collapse,” see P. R. Ehrlich and A. H. Ehrlich, ‘Can a Collapse of Global Civilization Be Avoided’ (2013) 280(1754) *Proceedings of the Royal Society B: Biological Sciences* 20122845 (“determining how to prevent ... a global collapse is perhaps the foremost challenge confronting humanity.”); L. Kemp and C. Rhodes, ‘The Cartography of Global Catastrophic Governance’ (2020), Centre for the Study of Existential Risk, at https://globalchallenges.org/wp-content/uploads/The-Cartography-of-Global-Catastrophic-Governance-Final.pdf, 2, Table 1 (“A global collapse could be considered as a lower bound for [existential risk] ...”);

for “global disasters,” see A. van Aaken, ‘Is International Law Conducive To Preventing Looming Disasters?’ (2016) 7(S1) *Global Policy* 81, 82 (“global disasters”); P. Susman, P. O’Keefe, and B. Wisner, ‘Global Disasters, a Radical Interpretation’ in K. Hewitt (ed) *Interpretations of Calamity* (Routledge, 1983);

for “global existential catastrophes,” see L. Rifkin, ‘The Survival of Humanity’ (13 September 2013) Scientific American at https://blogs.scientificamerican.com/guest-blog/the-survival-of-humanity/ (last accessed 7 December 2022) (“For global existential catastrophes, the "extent of harm" part of this equation would be astronomical.”); cf. also S. Baum, ‘Quantifying the Probability of Existential Catastrophe: A Reply to Beard et al.’ (2020) 123 *Futures* 102608 (“global and existential catastrophes”);

for “global existential risks,” see Permanent Representatives of the Netherlands and Fiji to the United Nations, n 3 above (“Member States and stakeholders clearly expressed that a Declaration for Future Generations should state a firm commitment to securing the interests of future generations in all decision making; by identifying, managing and monitoring global existential risks ...”);

for “high-impact, low-probability risks,” see 2021 Wellbeing of Future Generations Bill, Bill 253 [HL] at https://publications.parliament.uk/pa/bills/cbill/58-02/0253/210253.pdf, section 16.1.c (“an assessment of risks, including high-impact, low-probability risks”);

for “large-scale risks,” see U.N. Doc. A/RES/69/283 (2015) [Sendai Framework for Disaster Risk Reduction 2015-2030], para 15 (“risk of large-scale”);

for “risks of irreversible damage,” cf. United Nations Framework Convention on Climate Change (New York, 9 May 1992) 1771 U.N.T.S. 107, 31 I.L.M. 849 (1992), *entered into force* 21 Mar. 1994 [UNFCCC], art. 3.3 (“threats of ... irreversible damage”).

Note also that some related terms were necessarily left out, due to convenience and their closeness to other terms. Examples of other such terms include “existential catastrophe,” N. Bostrom, ‘The Vulnerable World Hypothesis’ (2019) 10(4) *Global Policy* 455, 458 et seq.; “ultimate harm,” I. Persson and J. Savulescu, *Unfit for the Future: The Need for Moral Enhancement* (Oxford: OUP, 2012); “oblivion,” B.E. Tonn, ‘Transcending Oblivion’ (1999) 31 *Futures* 351; and “threat to civilization,” D. Steel, C. T. DesRoches, and K. Mintz-Woo, ‘Climate Change and the Threat to Civilization’ (2022) 119(42) PNAS e2210525119.

> What is the minimum probability (as a %) that these lives will be endangered for something to constitute a "[risk]" according to this provision?

The first question required that participants enter a number between 0 and 8 billion for the minimum number of lives. The second question required that participants enter a number between 0 and 100 for the minimum probability, with decimals allowed.

In addition to the main questionnaire, materials also included an attention check (a simple multiplication problem) and a demographics questionnaire, which asked about age, politics (ranging from "strongly liberal" to "strongly conservative," with "centrist" in the middle), and gender identity ("male," "female," "non-binary," "prefer not to specify," and "prefer to self-identify").

2. Participants and Procedure

Participants (n=2,563) were recruited via the online platform Prolific. Participants were required to reside in the United States and speak English fluently. Participants were retained in the final analysis as long as they successfully completed the study and answered the attention check correctly.

With regard to procedure, participants were first shown (a) the demographics questionnaire, followed by (b) the attention check, and then (c) the main questionnaire. Parts (a), (b), and (c) were shown on separate screens. With respect to (c), participants saw just one version of the questionnaire, that is, using only one term for risk, such as "existential risk to humanity" but not "catastrophic risk to humanity," "global collapse," etc.

3. Results

Results are visualized in Figure 1. Of all ten terms tested, "existential risks to humanity" had the highest mean rating for both minimum lives harmed and minimum probability of harm, and thus also had the highest mean rating for minimum expected lives harmed.

With regard to minimum lives harmed, the mean participant rating for each wording tested was between 500 million (M) and 2 billion (B). The mean participant rating for "global existential risks" was highest at 1.876 B (95% CI: 1.574 B to 2.214 B), followed by "existential risks to humanity" (1.779 B; 95% CI: 1.467 B to 2.117 B), "existential risks" (1.734 B; 95% CI: 1.442 to 2.052 B), "global collapse" (1.721 B; 95% CI: 1.437 B to 2.010 B); "global existential catastrophe" (1.360 B; 95% CI: 1.17 B to 1.627 B); "risks of irreversible damage" (1.268 B; 95% CI: 1.021 B to 1.558 B);

"high-impact, low-probability risks" (1.187 B; 95% CI: 939 M to 1.433 B); "global catastrophic risks" (1.174 B; 95% CI: 937 M to 1.426 B); "global disasters" (987 M; 95% CI: 777 M to 1.196 B); "extreme risks" (849 M; 95% CI: 652 M to 1.060 B); and "large-scale risks" (744 M: 95% CI: 539 M to 934 M).

With regard to the minimum probability of those lives being harmed, the mean participant rating for each wording tested was between 20 and 40%. The mean participant rating for "existential risks to humanity" was highest at 37.6% (95% CI: 34.0 to 41.4), followed by "global collapse" (37.5%; 95% CI: 34.0 to 41.1), "global existential catastrophes" (34.5%; 95% CI: 31.4 to 37.7), "global existential risks" (34.0%; 95% CI: 30.3 to 37.6), "global disasters" (32.5%; 95% CI: 29.3 to 36.2), "global catastrophic risks" (37.5%; 95% CI: 34.0 to 41.1), "extreme risks" (32.3%; 95% CI: 28.8 to 35.8), "risks of irreversible damage" (31.6%; 95% CI: 28.1 to 35.4), "existential risks" (31.2%; 95% CI: 27.6 to 34.7), "large-scale risks" (29.6%: 95% CI: 26.2 to 32.8), and "high-impact, low-probability risks" (20.3%: 95% CI: 17.3 to 23.6).

With regard to our regression analyses comparing "existential risk" to other terms, our model found that participant ratings of minimum lives harmed were significantly higher for the existential risk condition than for "extreme risk," "global catastrophic risk," "global disaster," "global existential catastrophe," "high-impact low-probability risks," "irreversible damage," and "large-scale risk." There was no significant difference between "existential risk" and "global collapse," "existential risk to humanity," and "global existential risk" with regard to participant ratings of minimum lives harmed.

Our probability model found that participant ratings of minimum probability of harm were significantly higher for the "existential risk" condition than for "high-impact low-probability risk" and significantly lower than for "global collapse" and "existential risk to humanity." There was no significant difference between the "existential risk" condition and any other condition.

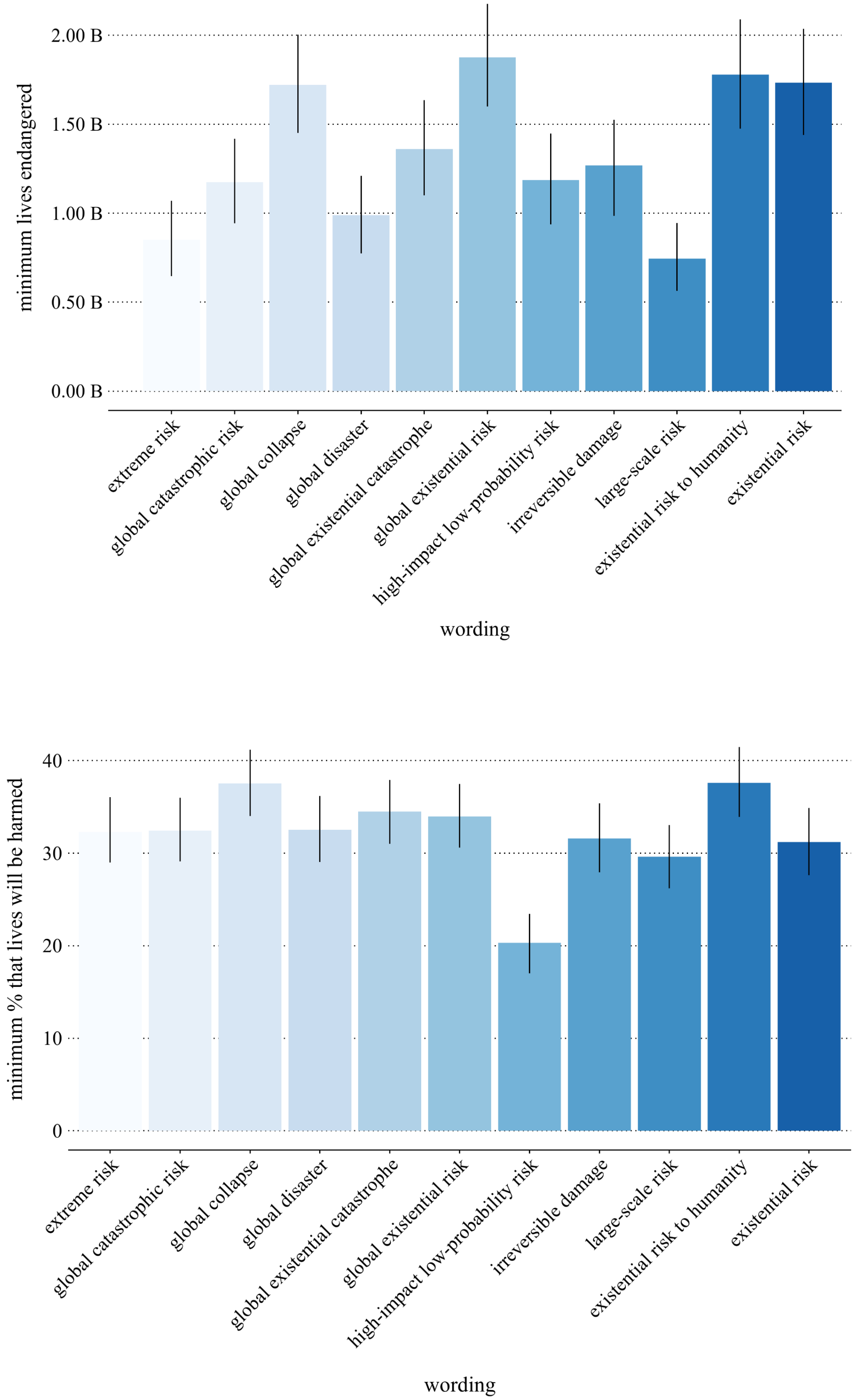


Figure 1: Mean response for (a) minimum number of lives endangered and (b) minimum probability those lives will be harmed for scenarios to constitute a particular risk.

B. Study 1b: Definitions and Examples

Having established the scope of people's understanding of "existential risk" and related terms with respect to minimum harm and likelihood of that harm, in Study 1b we sought to investigate how people's understanding of existential risk varied when accompanied by different types of definitions and illustrative examples of threats.

1. Materials

We constructed a set of materials similar to those in Study 1a, with a few deviations. First, every version of the questionnaire used only the term "existential risk to humanity" and none of the alternatives used in Study 1a (such as "catastrophic risk to humanity"). Second, the questionnaire provided a definition to accompany the term "existential risk," such that the structure of the scenario was as follows:

> Imagine a legal provision that requires governments to protect against "existential risks to humanity." The provision defines an existential risk as [definition].

There were five different basic definitions used in the materials: three drawn from definitions used in the existential risk literature, one drawn from proposed legislation in the United States, and one additional legal definition constructed for the purpose of this study. These definitions (and source, if applicable) were as follows:

1. FLI Definition: "any risk that has the potential to eliminate all of humanity or, at the very least, kill large swaths of the global population, leaving the survivors without sufficient means to rebuild society to current standards of living."[20]
2. Ord Definition: "[any risk] that threaten[s] the destruction of humanity's longterm potential."[21]
3. Bostrom Definition: "[any risk] that threatens the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development."[22]

[20] A. Conn, 'Existential Risk' Future of Life Institute, 16 November 2015 at https://futureoflife.org/existential-risk/existential-risk/ (last accessed 7 December 2022).

[21] Ord, n 4 above, 6.

[22] N. Bostrom, 'Existential Risk Prevention as Global Priority' (2013) 4(1) *Global Policy* 15, 15; see also 'Overview' Forethought Foundation for Global Priorities Research at https://www.forethought.org/research-overview (last accessed 7 December 2022).

4. United States Definition: "the potential for an outcome that would result in human extinction."[23]
5. Martínez-Winter Definition: "any risk of human extinction or the permanent destruction of human rights, interests or well-being"

For each of these basic definitions, we constructed an additional version of the questionnaire that contained the following language elaborating an exemplary list of types of threats that might fall within the scope or purview of the definition:

> The provision also provides examples of threats that might constitute existential risks, including nuclear war, biotechnology, artificial intelligence, and climate change.

Thus, in total, there were ten versions of the questionnaire: two versions for each of the five different definitions of existential risk, one with and one without the list of threats. The remainder of the questionnaire (ie the wording of the two questions) was the same across versions and identical to Study 1a. That is, the questionnaire asked participants to rate the minimum lives harmed and probability of those lives being harmed for a scenario to constitute an existential risk according to the provision. As in Study 1a, the materials included a demographics questionnaire that asked about age, politics, and gender, as well as an attention check that asked participants to solve a simple multiplication question.

2. Participants and Procedure

Participants (n=2,579) were recruited via the online platform Prolific. As in Study 1a, participants were required to reside in the United States and speak English fluently. Participants were retained in the final analysis if they successfully completed the study and answered the attention check correctly.

The procedure for Study 1b was the same as in Study 1a. Participants were first shown (a) the demographics questionnaire, followed by (b) the attention check, and then (c) the main questionnaire. All parts were shown on separate screens. With respect to (c), participants saw just one version of the questionnaire, that is, the two questions

[23] Senate Amendment 6438, proposed amendment to S.A. 5499 for National Defense Authorization Act for Fiscal Year 2023, H.R. 7900, Congressional Record Vol. 168, No. 158, 117th Cong. (29 September 2022) at https://www.congress.gov/congressional-record/volume-168/issue-158/senate-section/article/S6025-1; Senate Amendment 6464, proposed amendment to S.A. 5499 for National Defense Authorization Act for Fiscal Year 2023, H.R. 7900, Congressional Record Vol. 168, No. 162, 117th Cong. (11 October 2022) at https://www.congress.gov/congressional-record/volume-168/issue-158/senate-section/article/S6025-1; see also Global Catastrophic Risk Management Act of 2022, n 10 above ("The term 'existential risk' means the risk of human extinction.").

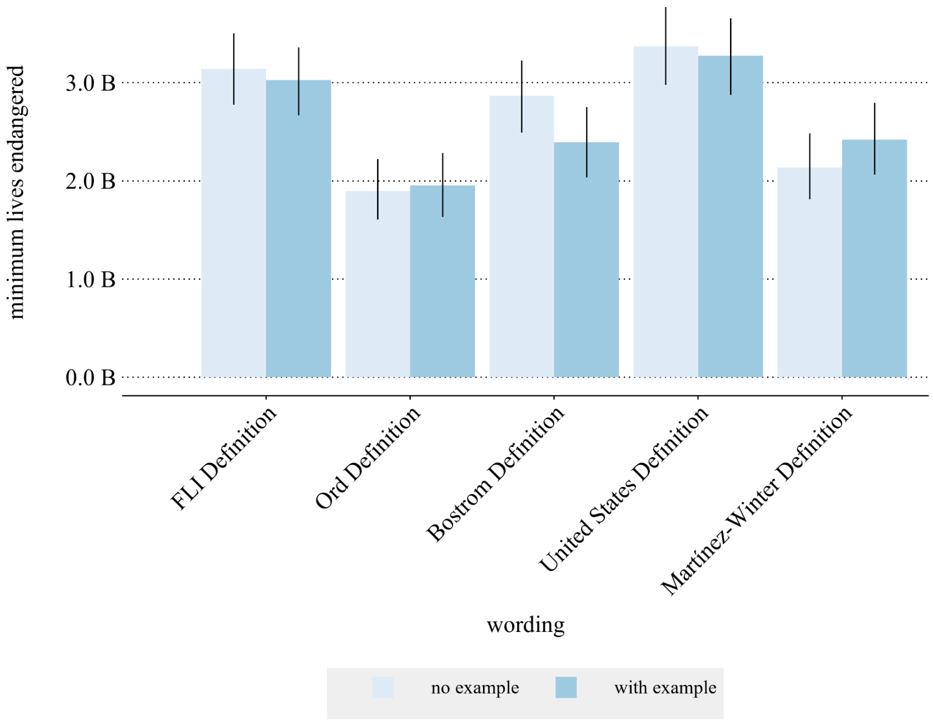
accompanying a provision with one definition of existential risk, either with or without examples of potential threats.

3. Analysis Plan

To investigate whether there was a significant effect of definition[24] and examples on ratings for minimum number of lives harmed, we conducted a mixed-effects linear regression with (a) definition (definition vs. no definition) and examples as fixed effects, (b) definition type and participant as random effects, and (c) number of lives as the outcome variable. To investigate whether the definition and examples of threats similarly affected ratings for minimum probability of harm, we conducted a mixed-effects linear regression with the same fixed and random effects but with minimum probability as the outcome variable.

4. Results

Results are visualized in Figures 2a and 2b.



[24] Note that for our models we included the "existential risk" data from Study 1a.

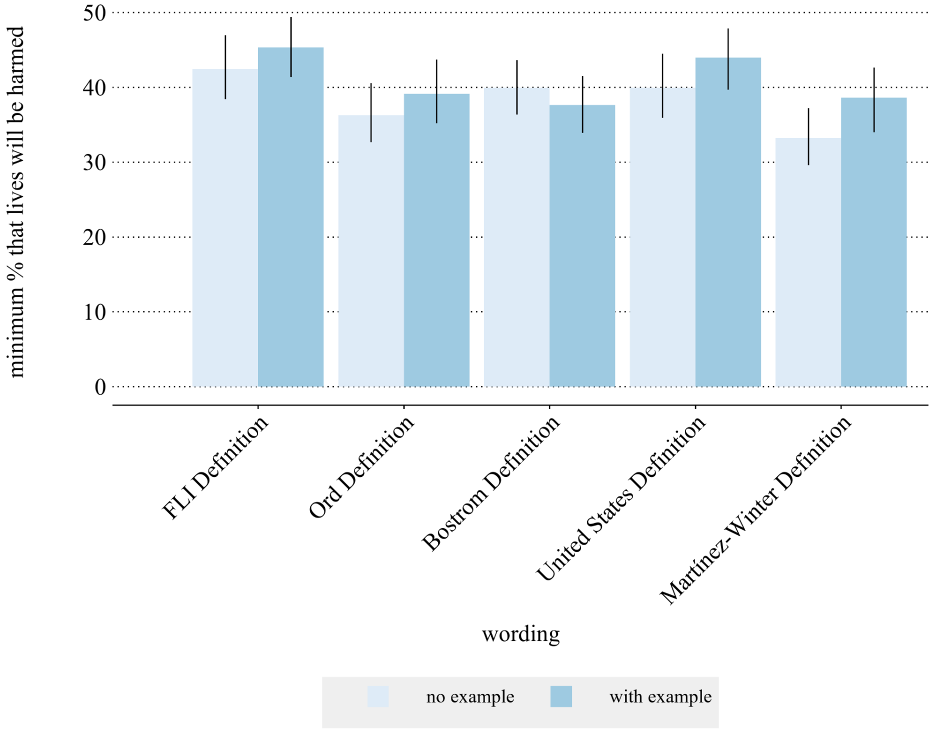


Figure 2: Mean response for (a) minimum number of lives endangered and (b) minimum probability those lives will be harmed for scenarios to constitute an existential risk, by definition type.

Across all conditions, the mean rating for minimum number of lives harmed was 2.654 B (95% CI: 2.535 B to 2.778 B), and the mean rating for minimum probability of harm was 38.7% (95% CI: 38.4 to 41.0).

Comparing different definitions, the definition with the highest mean rating for minimum number of lives harmed was the United States definition at 3.323 B (95% CI: 3.023 to 3.613 B); followed by the FLI definition at 3.085 B (95% CI: 2.817 B to 3.376 B), the Bostrom definition at 2.635 B (95% CI: 2.385 to 2.892 B), the Martínez-Winter definition at 2.278 B (95% CI: 2.034 B to 2.562 B), and the Ord definition at 1.924 B (95% CI: 1.706 B to 2.178 B).

With regard to the probability prompt, the definition with the highest mean rated minimum probability was the FLI definition at 43.9% (95% CI: 41.2 to 46.7), followed by the United States definition at 42.0% (95% CI: 38.9 to 44.8), the Bostrom definition at 38.8% (95% CI: 35.9 to 41.5), the Ord definition at 37.7% (95% CI: 34.8 to 40.5), and the Martínez-Winter definition at 36.0% (95% CI: 33.3 to 38.8).

Comparing example vs no-example conditions, we find that the mean rating for minimum number of lives harmed was 2.619 B (95% CI: 2.442 B to 2.790 B) across all example conditions and 2.689 B (95% CI: 2.531 B to 2.864 B) across all no-example conditions. With regard to probability, the mean rating for minimum probability of harm was 41.0% (95% CI: 39.3 to 42.9) across all example conditions and 38.4% (95% CI: 36.7 to 40.3) across all no-example conditions.

With regard to the regression, our models did not find a significant effect of definition vs. no definition on participant ratings of minimum lives harmed or minimum probability of lives harmed. Similarly, our models did not find a significant effect of examples vs. no examples.

C. Study 1c: Specified Probability Threshold

1. Materials

In Study 1c, we further sought to investigate whether people's interpretation of existential risk was sensitive to whether the definition specified the probability of the risk. To do so, we prepared a set of materials (n=4 conditions) similar to those in Study 1b, with some deviations. As in Study 1b, every version of the questionnaire used the term "existential risk to humanity." For the sake of convenience, and because the specific wording of the definition (beyond the specification of the probabilities) was not the main focus of Study 1c, every version of the questionnaire used the same basic definition (#5 in Study 1b):

> Imagine a legal provision that requires governments to protect against "existential risks to humanity." The provision defines an existential risk as any risk, including [probability] risks, of human extinction or the permanent destruction of humanity's potential.

Each of the versions had a different specification of the probability, as follows:

1. low-probability
2. very low-probability
3. extremely low-probability

For the very low-probability condition, we constructed an additional version of the questionnaire that contained the following language with a numerical value for probability:

> The provision further defines very low-probability risks as including "risks with an estimated likelihood of occurrence of as low as 1%, according to the best evidence available."

In total, there were four versions of the questionnaire. The remainder of the questionnaire (ie the wording of the two questions) was the same across versions and identical to Studies 1a and 1b in that it asked participants to rate the minimum lives harmed and probability of those lives being harmed for a scenario to constitute an existential risk according to the provision. The materials included a demographics questionnaire that asked about age, politics, and gender, as well as an attention check that asked participants to solve a simple multiplication question.

2. Participants and Procedure

Participants (n=922) were recruited via the online platform Prolific. As in Studies 1a and 1b, participants were required to reside in the United States and speak English fluently. Participants were retained in the final analysis if they successfully completed the study and answered the attention check correctly.

For procedure, as in Studies 1a and 1b, participants were first shown (a) the demographics questionnaire, followed by (b) the attention check and (c) the main questionnaire. All parts were shown on separate screens. With respect to (c), as in Studies 1a and 1b, participants saw just one version of the questionnaire (ie one provision with a definition including one specification of probability).

3. Analysis Plan

For each condition, we calculated a confidence interval of the mean response to both the probability and minimum lives harmed questions using the bias-corrected and accelerated (BCa) bootstrap method based on 5,000 replicates of the sample data. To better understand the distribution of participant responses for each condition, we also computed the middle 50% range of responses to the probability question—that is, the difference between the 25th and 75th percentile of participant responses.

4. Results

Results of Study 1c are visualized in Figures 3a and 3b.

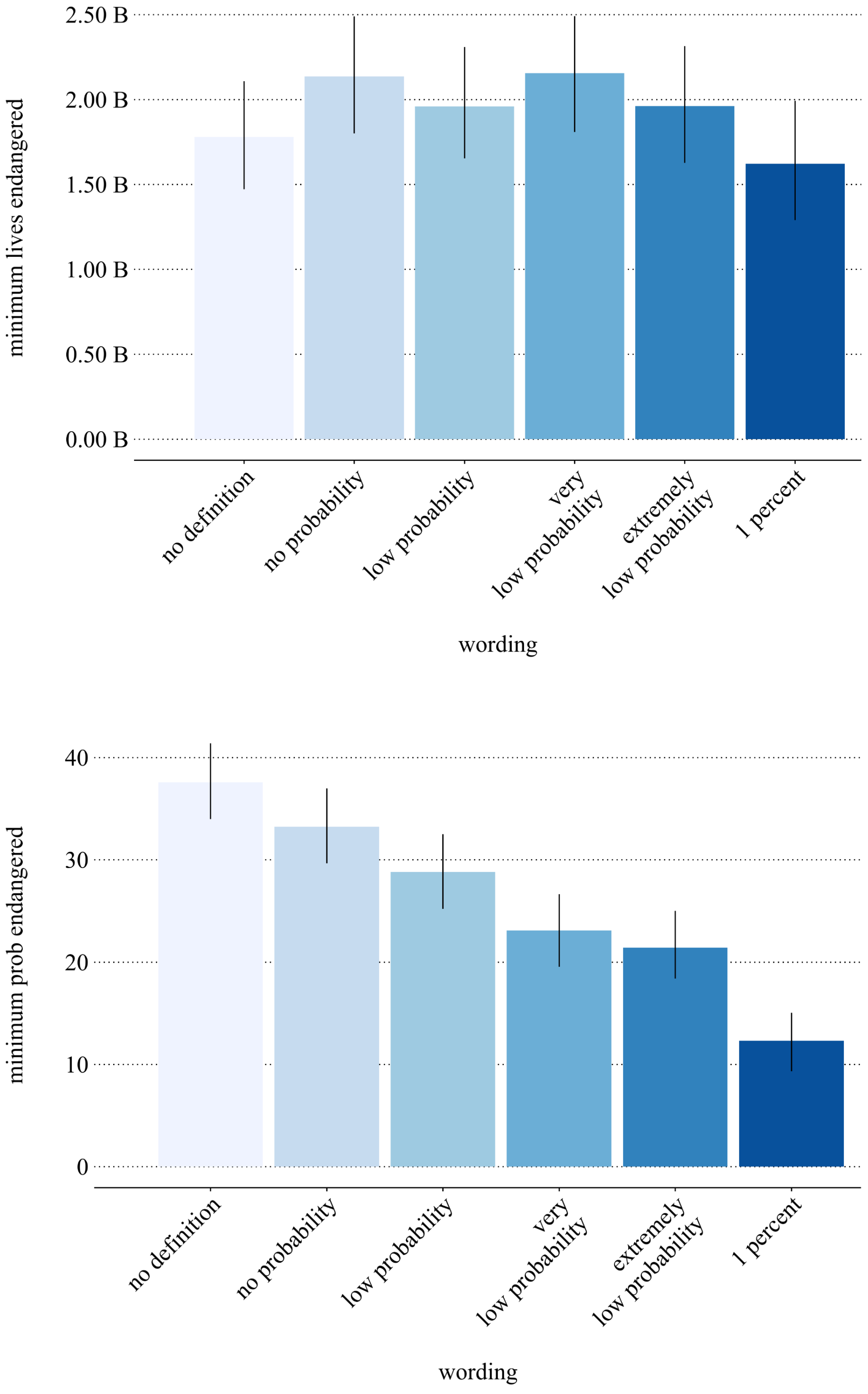


Figure 3: Mean response for (a) minimum number of lives endangered and (b) minimum probability those lives will be harmed for scenarios to constitute a an existential risk, by specified probability threshold.

With regard to the minimum number of lives harmed, the specification with the highest mean response was the definition with very low probability at 2.155 B (95% CI: 1.824 B to 2.488 B), followed by no probability specification at 2.136 B (95% CI: 1.816 B to 2.520 B), extremely low probability at 1.961 B (95% CI: 1.633 B to 2.305 B), low probability at 1.959 B (95% CI: 1.641 B to 2.330 B), and as low as 1% at 1.621 B (95% CI: 1.288 B to 1.986 B).

With regard to the minimum probability of harm, the specification with the highest mean response was the definition with no probability specification at 33.2% (95% CI: 29.4 to 36.9); followed by the specification of low probability at 28.8% (95% CI: 25.4 to 32.5), very low probability at 23.1% (95% CI: 19.8 to 26.3), extremely low probability at 21.4% (95% CI: 18.0 to 24.6), and less than 1% at 12.3% (95% CI: 9.5 to 15.3).

When looking at the middle 50% of participant responses (25th to 75th percentile) for minimum probability of harm, the definition with the widest range specified low probability with a range of 45 percentage points (5 to 50), followed by no probability at 40 (10 to 50), very low probability at 39 (1 to 40), extremely low probability at 31.5 (1 to 32.5), and as low as 1% at 9 (1 to 10).

III. Study 2: Existential Risk Scenarios

A. Materials

For Study 2, we constructed a series of questions asking participants to read a legal provision related to the protection of existential risk and evaluate whether a particular scenario constitutes an existential risk according to the provision. The basic wording of each of these questions was as follows:

> Imagine a legal provision that requires governments to protect against "existential risks to humanity." The provision defines an existential risk as [definition]. Suppose the human population currently stands at 8 billion (8,000,000,000) people.
>
> Now imagine [scenario] which, if it occurred, would kill [lives] billion people. According to the best available evidence, there is a [probability] chance that [scenario will occur].
>
> In your estimation, does the risk of this scenario occurring constitute an "existential risk to humanity" according to the legal provision? (yes/no)

The values for [definition], [scenario], [probability], and [lives] varied in the following manner. First, our materials contained three different options for [definition] as follows (differences among the formulations in **bold**):

1. (no probability reference) any risk of human extinction or the permanent destruction of humanity's potential.
2. (low probability reference) any risk, **including low-probability risks**, of human extinction or the permanent destruction of humanity's potential.
3. (low probability w/ definition) any risk, **including low-probability risks**, of human extinction or the permanent destruction of humanity's potential. **The provision further defines low-probability risks as including "risks with an estimated likelihood of occurrence of as low as 1%, according to the best evidence available."**

Our materials contained four different options for [scenario], including a generic abstract scenario, an AI scenario, a pandemic scenario, and a climate change scenario. The wording of each of these formulations were as follows:

1. (generic) Now imagine there is a potential scenario which, if it occurred, would kill 8 billion people. According to the best available evidence, there is a 1% chance that this scenario will occur.
2. (artificial intelligence) Now imagine a company has created an AI system that, were it to be corrupted, would kill 8 billion people. According to the best available evidence, there is a 1% chance that this AI system will be corrupted.
3. (pandemic) Now imagine two countries are preparing for a pandemic, which, if it occurred, would kill 8 billion people. According to the best available evidence, there is a 1% chance that this pandemic will occur.
4. (climate change) Now imagine there is a potential extreme climate crisis which, if it occurred, would kill 8 billion people. According to the best available evidence, there is a 1% chance that this climate crisis will occur.

To assess whether participant ratings were sensitive to expected harm, we varied the expected number of lives harmed such that the product of [probability] and [lives] equaled one of three values: 80 million, 160 million, and 400 million. Moreover, to assess whether participants' judgments deviated from an expected value calculation,[25] within each expected value condition we constructed 10–12 combinations of [probability] and [lives]. These combinations are visualized in Table 1.

Probability	Lives (80 million EV condition)	Lives (160 million EV condition)	Lives (400 million EV condition)
1%	8 billion	n/a	n/a
2%	4 billion	8 billion	n/a
5%	1.6 billion	3.2 billion	8 billion
10%	800 million	1.6 billion	4 billion
20%	400 million	800 million	2 billion
50%	160 million	320 million	800 million
80%	100 million	200 million	500 million
90%	90 million	180 million	450 million
95%	85 million	170 million	425 million
98%	82 million	164 million	410 million
99%	81 million	161 million	402 million
100%	80 million	160 million	400 million

Table 1: Probability and lives shown in expected value conditions

In addition to these main materials, we also constructed several attention-check versions of the scenario where the expected value either equaled 0 people or 8 billion people, as well as a demographic questionnaire identical to those used in Study 1.

B. Participants and Procedure

Participants (n=750) were recruited via Prolific using the same eligibility criteria as in Study 1. With regard to procedure, participants randomly saw 12 versions of the prompt. Randomization was set up such that participants saw (a) exactly one prompt of

[25] That is, whether participant ratings were sensitive to the probability of harm or number of lives, independent of expected harm.

each probability value and (b) exactly one prompt of each scenario and existential risk definition combination (ie three prompts with each scenario type and four prompts with each existential risk definition).

Interspersed among those 12 trials were two attention-check trials, in which participants were presented with scenarios in which the expected lives harmed equaled either 0 people or 8 billion people. Participants were retained in the final analysis if they completed the study and answered the attention-check trials “correctly” by responding “yes” for scenarios with 8 billion expected lives harmed and “no” for 0 expected lives harmed.

C. Analysis Plan

We conducted a mixed-effects logistic regression with (a) expected harm, scenario type, probability specification, and lives/probability as fixed effects, (b) participant as random effect, and (c) response as the outcome variable, with “yes” coded as a 1 and “no” coded as a 0.

D. Results

Results for Study 2 can be visualized in Figure 4. With regard to different expected value conditions, the condition with the highest rate of “yes” responses to whether a particular scenario constituted an existential risk was the 400 million lives condition at 76.8% (95% CI: 74.9 to 78.8), followed by the 160 million lives condition at 68.8% (95% CI: 66.9 to 71.0) and the 80 million lives condition at 65.3% (95% CI: 63.0 to 67.4).

With regard to different probability specifications, the highest rate of “yes” responses participants responded that scenarios constituted an existential risk most often in the 1% condition, with 77.4% responding "yes" 77.4% (95% CI: 75.5 to 79.2), followed by low probability condition at 68.5% (95% CI: 66.5 to 70.5) and no probability specified condition at 65.7% (95% CI: 63.7 to 67.8).

With regard to different types of scenarios, participants responded that scenarios constituted an existential risk most often in the climate change scenario, with 71.5% responding "yes" (95% CI: 69.3 to 73.6), followed by the AI scenario at 71.2% (95% CI: 68.9 to 73.5), the generic scenario at 69.2% (95% CI: 66.5 to 71.7), and the pandemic scenario at 68.9% (95% CI: 66.5 to 77.1).

Our model revealed a significant effect of expected value on the responses, in that participants responded that scenarios constituted an existential risk significantly more often in the 400 million lives condition than in the reference condition of 160 million lives (β=.4048, SE=.0743, z=5.446, p<.001), and significantly less often in the 80 million lives condition (β=-.2195, SE=.0690, z=-3.180, p=.0015). With regard to scenario, just as participants were more likely to respond that a scenario constituted an existential risk if it involved climate change, the climate change scenario also had a significantly higher % yes response rate than the reference condition (generic scenario) (β=.1617, SE=.0803, z=2.013, p=.0442). No significant differences were found among the other conditions.

Our model also revealed a significant effect of lives and probability, in that conditions with a lower probability of harm and higher number of lives harmed had a significantly higher % yes response rate than conditions with a higher probability of harm and lower number of lives harmed (β=-.0648, SE=.0091, z=-7.097, p=.0442), even for the same expected value. Our model also found a significant effect of probability definition in that the as low as 1% condition had a significantly higher % yes response rate than the low probability condition (β=-.4640, SE=.0754, z=-6.150, p<.001) and the no probability condition (β=-.6088, SE=.0747, z=-8.154, p<.001).

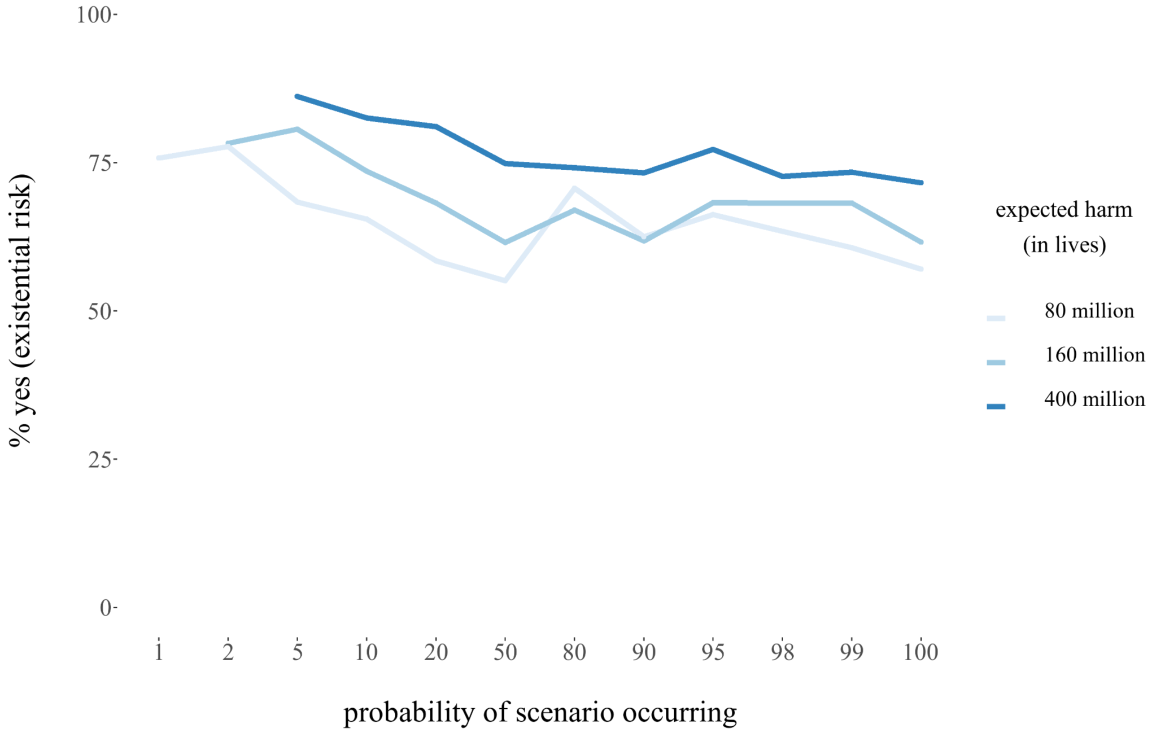


Figure 4: Percent of participants who responded “yes” to whether a particular scenario constituted an existential risk, as a function of probability of the scenario occurring (x-axis), and the expected number of lives harmed by the scenario (lines).

IV. Discussion

This Part turns to the implications of these studies for theory and practice of legal interpretation and lawmaking. Section A summarizes the descriptive findings of the studies and describes to what extent they advance our understanding of how people interpret existential risk and other terms in the context of legal provisions. Section B considers the implications of these findings for judicial interpretation, both in the context of existential risk and in general and identifies an abstract/concrete paradox, in which the interpretation of a term varies depending on the abstract or concrete nature of the scenario described. Section C discusses the significance of the experimental results for existential risk legislation, as well as future research aimed at informing policy efforts more generally.

A. Cognitive Implications

The first question we set out to answer was how laypeople interpret the term "existential risk" relative to other terms referenced in the existential risk and international law literature. In Study 1a, respondents interpreted the term "existential risk" as requiring, on average, a higher minimum number of lives to be endangered compared to other terms. Respondents also interpreted "existential risk" as requiring a higher minimum probability of those lives being harmed and, in turn, a higher minimum expected harm (calculated by multiplying the minimum number of lives times the minimum probability of those lives being harmed). These findings indicate that laypeople, like experts, interpret the term "existential risk" as referring to a risk that is more serious in magnitude of lives endangered and expected harm compared to terms such as a "catastrophic risk" or "large-scale risk."[26] However, laypeople diverge from experts by interpreting versions of "existential risk" as requiring, on average, a higher minimum probability of lives being endangered compared to other terms. In contrast, experts interpret the term "existential risk" to include certain scenarios of low likelihood.[27] Laypeople also seem to interpret existential risk as requiring a lower minimum number of lives endangered compared to experts, who interpret the term as only referring to risks that would endanger virtually all of humanity.[28] At the same time, other terms we looked at also had a very high probability threshold (the lowest was "high-impact, low-probability risk" at 20.4%, compared to "existential risk" at 38%) and a lower minimum lives endangered threshold, indicating that laypeople's abstract interpretation of these terms also fails to cover the scenarios considered by experts to fall under the category of existential risks.

[26] Ord, n 4 above; Conn, n 20 above.

[27] Ord, n 4 above; Bostrom, n 22 above.

[28] Ord, n 4 above; Bostrom, n 22 above; Conn, n 20 above.

The second question we set out to answer was how laypeople's interpretation of existential risk is affected by their being provided definitions and examples of the term. When comparing participant responses to provisions with a definition (Study 1b) to those without a definition (Study 1a), our regression models did not find a significant difference between the two conditions with respect to participant ratings of minimum lives harmed, nor to participant ratings of minimum probability of lives harmed, suggesting that laypeople's understanding of existential risk is not meaningfully affected by the different definitions provided. Our model revealed a similar lack of a main effect of examples vs no examples of potential threats, suggesting that laypeople's understanding of existential risk is likewise not meaningfully affected by being presented with examples of types of threats ("nuclear war, biotechnology, artificial intelligence, and climate change") that fall within the definition of the term.

The third question we set out to answer was how laypeople's interpretation of existential risk is affected by definitions that specify a probability threshold. The fact that our regression models in Study 1c revealed a main effect of probability specification on participant ratings of the minimum probability required for a scenario to count as an existential risk suggests, perhaps unsurprisingly, that people's interpretation of existential risk is sensitive to the specification of probability. At the same time, the average minimum probability threshold given by participants even for the as low as 1% probability condition was 11.6%, suggesting that participants are reluctant to consider low-probability scenarios as constituting an existential risk (at least in the abstract), even when the legal provision so states.[29]

The fourth question we set out to answer was whether people's interpretation of the probability and lives threshold of existential risk differs depending on the type of scenario presented. In Study 2, our model revealed that participants were slightly more likely to endorse the climate change scenario as an existential risk compared to the reference condition (generic scenario). No significant differences were found among the generic, artificial intelligence, and pandemic conditions, indicating that people otherwise have stable perceptions of existential risk across scenario types.

[29] Note that, while the mean was 11.6%, many participants (43.4%) did, in fact, choose 1% as the minimum probability threshold. Furthermore, while the mean may be useful in determining the ordinary meaning of a term, it does not necessarily mean that it should be the *only* measure used to determine the ordinary meaning of the term, nor does it necessarily reflect how judges will or should interpret a term. In particular, some argue that "ordinary meaning" refers not to how an ordinary person understands a given term but rather how a reasonable reader might interpret a legal term. T. L. Grove, 'Testing Textualism's "Ordinary Meaning"' (20 August 2022) at https://dx.doi.org/10.2139/ssrn.4190031 (last accessed 7 December 2022) ('many prominent textualists have long treated "ordinary meaning" as a legal concept—one that must be elucidated through the understanding of a hypothetical reasonable reader'). According to this view, one might argue that a reasonable reader would still interpret the provision as referring to probabilities of as low as 1%, even if, on average, the provision was interpreted to refer only to probabilities higher than that, and that, by extension, the provision should be interpreted as referring to probabilities of as low as 1%.

The fifth question we set out to answer was whether people's evaluation of existential risk deviates from an expected value calculation. Our regression model in Study 2 revealed a significant effect of expected value, indicating that people's interpretation of whether a scenario counts as an existential risk is sensitive to the amount of harm expected from that scenario. At the same time, the fact that the conditions with a lower probability of harm and higher number of lives harmed had a significantly higher % yes response rate than conditions with a higher probability of harm and lower number of lives harmed, even for the same expected value, indicates that people's judgments of whether a particular scenario constitutes an existential risk is not sensitive only to the expected amount of harm overall, but also to the total number of lives threatened.

B. Doctrinal Implications

As stated in the introduction, judges around the world tend to interpret words in a legal provision according to their *ordinary meaning*–that is, how a typical or reasonable person generally understands and uses a given word or concept.[30] Since our study investigated how people generally understand the "existential risk" and related terms, the results of our study are directly informative of their ordinary meaning, and, by extension, are directly informative of how judges ought to apply the ordinary meaning doctrine to legal provisions related to existential risk.

Our results indicate several areas of convergence between the ordinary meaning of existential risk and its generally understood technical meaning. For example, laypeople in our study interpreted "existential risk" as referring to a risk that is more serious in magnitude of lives endangered and expected harm compared to risks encompassed by other terms, such as a "catastrophic risk" or "large-scale risk." This understanding of "existential risk" is narrower than these other terms with respect to the minimum lives threshold and the minimum expected value threshold—consistent with their respective technical definitions. Moreover, the fact that laypeople's perceptions of existential risk in our study were mostly stable across scenario types indicates that judges who apply the ordinary meaning doctrine should likewise apply the doctrine stably across scenarios (perhaps with slightly higher weight to climate change scenarios than other scenarios).

Our results also suggest two areas of divergence between the ordinary and technical meaning of existential risk. First, the ordinary meaning of existential risk may be narrower on some dimensions than its technical meaning, as laypeople's abstract

[30] See Tobia, n 15 above; Lee and Mouritsen, n 15 above; Klapper, Schmidt, and Tarantola, n 15 above. For example, in *Addison* v. *Holly Hill Fruit Products, Inc.*, 322 U.S. 607, 618 (1944), the court stated that "legislation when not expressed in technical terms is addressed to the common run of men and is therefore to be understood according to the sense of the thing, as the ordinary man has a right to rely on ordinary words addressed to him."

threshold for what counts as an existential risk in terms of the minimum probability and level of expected harm is higher than that often described by experts.

Second, ordinary understandings of existential risk may be dynamic, changing depending on context and the amount of detail provided, compared to static technical definitions. Participants were more willing to consider something an existential risk when presented with a more concrete scenario—such as the generic, climate change, artificial intelligence, or pandemic scenarios of Study 2—compared to the more abstract prompts of a term, definition, or even probability threshold (with no concrete scenario) in Study 1. Participants in Study 2 not only displayed a greater willingness to rate low-probability scenarios as existential risks, but were in fact more likely to rate low-probability scenarios as existential risks than high-probability scenarios of equivalent expected value. These apparently dichotomous results suggest a tension between the "abstract" ordinary meaning of existential risk and the "concrete" ordinary meaning of existential risk, similar to other abstract/concrete paradoxes observed in legal contexts[31] and non-legal contexts[32] in previous studies.[33]

In addition to deviating from the technical definition, this abstract/concrete paradox poses a potential problem for judges attempting to apply the ordinary meaning doctrine in existential risk cases. On the one hand, given that judges interpret the words in legal provisions as applied to specific cases as opposed to merely in the abstract, one might argue that people's concrete judgments in Study 2 are a more reliable indicator of the ordinary meaning of existential risk than people's abstract judgments in Study 1a, and by extension, are a more accurate reflection of how judges should interpret existential risk in real-world cases. On the other hand, some might argue that people's abstract interpretation of a given word or concept is a more reliable indicator of that word's meaning, just as some have argued that more abstract intuitions are more reliable than concrete intuitions[34].

[31] See eg N. Struchiner, G. Almeida, and I. Hannikainen, 'Legal Decision-Making and the Abstract/Concrete Paradox' (2020) 205 *Cognition* 1; P. Bystranowski et al., 'Do Formalist Judges Abide By Their Abstract Principles? A Two-Country Study in Adjudication' (2021) 35 *International Journal for the Semiotics of Law* 1903; D. Lewinsohn-Zamir, I. Ritov, and T. Kogut, 'Law and Identifiability' (2017) 92 *Indiana Law Journal* 505; K. M. Carlsmith et al., 'Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment' (2002) 83 *Journal of Personality and Social Psychology* 284.

[32] See eg L. Caviola, S. Schubert, and A. Mogensen, 'Should You Save the More Useful? The Effect of Generality on Moral Judgments About Rescue and Indirect Effects' (2021) 206 *Cognition* 104501; S. Nichols and J. Knobe, 'Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions' (2007) 41 *Noûs* 663; C. Freiman and S. Nichols, 'Is Desert in the Details?' (2011) 82 *Philosophy & Phenomenological Research* 121; D. H. Bostyn, S. Sevenhant, and A. Roets, 'Of Mice, Men, and Trolleys: Hypothetical Judgment Versus Real-Life Behavior in Trolley-Style Moral Dilemmas' (2018) 29 *Psychological Science* 1084.

[33] The abstract/concrete paradox refers to the tendency to activate inconsistent intuitions (and generate inconsistent judgments) depending on whether a problem to be analyzed is framed in abstract terms or is described as a concrete case. Bystranowski et al., n 31 above.

[34] See eg H. Sidgwick, *The Methods of Ethics* (London: Macmillan and Co., 7th ed, 1907).

This abstract/concrete paradox poses problems, not just for the interpretation of existential risk, but for ordinary meaning analysis more generally when people's reported interpretation of a term in the abstract differs from their interpretation of that term in a concrete case. How can one get at the "true" ordinary meaning of a provision if people's reported understanding differs based on how the question is asked? One method, as alluded to above, is to try to determine which of several framings is a more accurate reflection of participants' true understanding of the term. A second method is to try to reconcile participants' apparently contradictory responses in such a way as to arrive at a coherent ordinary meaning of the term.[35] To the extent that these strategies are unsatisfactory, the abstract/concrete paradox may call into question the presumption of using ordinary meaning analysis, either overall or in cases where framing effects and other cognitive biases are most likely to manifest (eg cases that involve specific numbers)[36].[37]

[35] For example, previous work has found that people given both conditions at the same time (abstract/concrete) give the same response for both. Struchiner, Almeida, and Hannikainen, n 31 above, 7 ("When simultaneously evaluating concrete and abstract cases, we observed a tendency toward consistency across levels of abstraction—as predicted by the bias hypothesis"). This suggests that the different answers when presented separately may be due to some bias, and that there is some unbiased version of both answers that are non-conflicting.

[36] cf. D. Kahneman and A. Tversky, 'Prospect Theory: An Analysis of Decision under Risk' (1979) 47 *Econometrica* 263; C. Winter, 'The Value of Behavioral Economics for EU Judicial Decision-Making' (2020) 21 *German Law Journal* 240; E. Yudkowsky, 'Cognitive Biases Potentially Affecting Judgment of Global Risks' in N. Bostrom and M. M. Ćirković (eds), *Global Catastrophic Risks* (New York: OUP, 2008); S. Schubert, L. Caviola, and N. S. Faber, 'The Psychology of Existential Risk: Moral Judgments about Human Extinction' (2019) 9 *Scientific Reports* 15100.

[37] Note that similar issues have plagued judges attempting to apply ordinary meaning analysis by appealing to dictionary definitions as opposed to empirical methods. Dictionaries often have multiple definitions that may conflict with one another and/or may conflict with the definitions provided by other dictionaries (eg in the United States: *Muscarello* v. *United States*, 524 U.S. 125, 1998; *Tamiguchi* v. *Kan Pacific Saipan, Ltd*, 566 U.S. 560, 2012; *Babbit* v. *Sweet Home Chapter*, 515 U.S. 687, 1995), which poses problems for judges attempting to determine which definition is an accurate reflection of a word's ordinary meaning. Indeed, this very problem has been observed to be the reason why certain judges turn to other interpretive tools besides ordinary meaning analysis, as well as the stated reason for why scholars have recently turned to empirical methods (such as corpus linguistics and surveys) as a means of uncovering ordinary meaning as opposed to dictionary definitions. See Lee and Mouritsen, n 15 above; S. C. Mouritsen, 'Hard Cases and Hard Data: Assessing Corpus Linguistics as an Empirical Path to Plain Meaning' (2011) 13 *Columbia Science & Technology Law Review* 156; Tobia, n 15 above; Klapper, Schmidt, and Tarantola, n 15 above.

Note also that the heavy reliance on dictionaries is a recent phenomenon in the United States. J. J. Brudney, & L. Baum, 'Oasis or Mirage: The Supreme Court's Thirst for Dictionaries in the Rehnquist and Roberts Eras' (2011) 5 *William & Mary Law Review* 483 (explaining that, while the United States Supreme Court's use of dictionaries was virtually non-existent before 1987, by 2010 as many as one-third of statutory decisions cited dictionary definitions). Outside the United States, dictionaries are used as a less dispositive, more supportive tool, as in international treaties. See Richard K. Gardiner, *Treaty Interpretation*, (Oxford: OUP, 2015); I. Van Damme, 'On "Good Faith Use of Dictionary in the Search of Ordinary Meaning under the WTO Dispute Settlement Understanding"—A Reply to Professor Chang-Fa Lo' (2010) 2 *Journal of International Dispute Settlement* 231 (stating that "undisputedly, dictionaries are used to determine the ordinary meaning. Some, or rather, most courts and tribunals rely on them without any express statement to

Recent work has called into question the presumption of ordinary meaning on other grounds. For example, Tobia et al.[38] found that ordinary people regularly take terms in law to communicate technical meanings as opposed to ordinary ones, suggesting that the underlying aims of the ordinary meaning doctrine—democratic interpretation, fair notice, and rule-of-law values–may be better met by looking to *technical* as opposed to ordinary meanings. Our results present new grounds for appealing to the technical meaning of a term over its ordinary meaning—namely, that in some cases (particularly those that involve specific numbers), the ordinary meaning of a term may simply be incoherent or undiscoverable according to this method of evaluation. Given previous work showing that judges, like laypeople, display abstract/concrete paradoxes in their decision-making,[39] as well as difficulties in reasoning about small probabilities,[40] one might favor not only applying the technical definition of a term, but appealing directly to expert interpretation of the technical definition in a concrete case.

C. Policy Implications

The prevalence of the *ordinary meaning* principle across the world's jurisdictions has implications, not only for those applying the law, but for those creating the law as well (ie lawmakers). As alluded to above, the fact that judges tend to interpret words in a law according to how they are ordinarily interpreted by laypeople[41] implies that lawmakers should ensure that laypeople (and, thus, ultimately judges) would interpret the words in a proposed law in their intended manner so as to achieve the lawmakers' intended policy aims.

In terms of choosing among potential terms to use in existential risk legislation, our results suggest that lawmakers who prefer to cover a smaller set of risks that endanger more lives and are more likely to occur (and therefore have a greater expected harm) should strongly consider using the term "existential risk" over alternatives discussed here. Laypeople interpret it as having the highest requirements for minimum probability of harm, minimum number of lives, and minimum expected harm, therefore a judge

that effect. Other dispute settlement bodies, like those of the WTO, prefer explicit references to dictionaries ...").

[38] K. P. Tobia, B. G. Slocum, and V. Nourse, 'Ordinary Meaning and Ordinary People' (forthcoming) 171 *University of Pennsylvania Law Review* at https://dx.doi.org/10.2139/ssrn.4034992 (last accessed 7 December 2022).

[39] Struchiner, Almeida, and Hannikainen, n 31 above.

[40] S. I. Gatowski et al., 'Asking the Gatekeepers: A National Survey of Judges on Judging Expert Evidence in a Post-Daubert World' (2001) 25 *Law and Human Behavior* 433.

[41] Note that there are several other interpretive tools used and endorsed by judges and academics, particularly when the ordinary meaning is unclear. cf. Lee and Mouritsen, n 15 above. For example, other tools endorsed by the majority of United States legal academics include looking to legislative intent or purpose, or evaluating the consequences of applying one interpretation over another. E. Martínez and K. P. Tobia, 'What Do Law Professors Believe about Law and the Legal Academy? An Empirical Inquiry' (5 August 2022) at https://dx.doi.org/10.2139/ssrn.4182521 (last accessed 7 December 2022).

applying ordinary meaning analysis would likely interpret it more narrowly than other candidate terms. Conversely, a lawmaker intending to cover risks with wider bands of probability and lives endangered might prefer other terms that better match their preferences, given that laypeople (and existential risk experts) interpret it as having a broader scope of application along these dimensions.

In terms of choosing whether to include a definition of existential risk in a particular provision, the fact that certain definitions (particularly those that explicitly defined a low probability threshold) had a much lower minimum probability threshold than existential risk without a definition indicates that insofar as lawmakers intend a provision to cover lower probability scenarios, they should include in the provision an explicit definition of their minimum probability threshold to increase the likelihood that those scenarios are intercepted by the eventual judge as falling within the scope of the provision. More broadly, the fact that these and previous studies have found that verbal probabilities (such as "very unlikely") are interpreted in a much more disparate fashion from person to person than specific numbers[42] further suggests that lawmakers intending a provision to cover a specific probability range should specify that probability range using specific numbers as opposed to vague verbal descriptions of that range, as judges (whether or not they engage in ordinary meaning analysis) may otherwise interpret the range differently from that which the legislator had in mind.

Our results did not reveal a significant difference in how people interpreted the minimum lives harmed or probability of harm when provided with a definition compared to no definition. While choice of definition did influence how people interpreted "existential risk," the mere inclusion of a definition did not shift interpretations in a particular direction. There was also no significant difference when including or omitting examples of types of existential risks: nuclear war, biotechnology, artificial intelligence, and climate change, suggesting that the inclusion of such examples may not be as helpful as might otherwise be assumed for lawmakers intending a provision to cover a specific harm or probability range (though perhaps still useful to ensure that specific *types* of risk are covered). However, the more elaborate, concrete scenarios presented in Study 2 did change how participants interpreted "existential risk," making them more likely to consider low-probability scenarios (as well as those of lower expected harm) to be existential risks. Given the potential for this context to influence ordinary meaning of the term chosen, as well as the fact that judges applying ordinary meaning analysis may likewise be influenced by this context (or lack thereof), laws intended to cover lower probability risks might benefit from descriptions of specific scenarios beyond a mere listing of types of threats.

[42] B. C. Wintle et al., 'Verbal Probabilities: *Very Likely* To Be *Somewhat* More Confusing Than Numbers' (2019) 14(4) PLoS One e0213522.

In terms of specific scenarios, the fact that participants' judgments of whether a particular scenario constitutes an existential risk were mostly insensitive to the type of scenario indicates that lawmakers generally may not need to include a list of example scenarios in order to for a judge to interpret a particular scenario as falling within the scope of that provision.[43] In terms of expected value, the fact that participants displayed a general tendency to judge scenarios as being an existential risk based on their expected harm (with some extra weight to low probability scenarios of equivalent expected value)[44]

[43] The one exception to this—that people are slightly more likely to judge a particular scenario as an existential risk if it pertains to climate change—only weakly implies that lawmakers that are particularly concerned about a specific type of scenario being interpreted as an existential risk may want to include that as an example in the provision. Moreover, although participants were likely to judge all types of scenarios as falling within the scope of the provision, it is conceivable that specifying a list of examples may make it even more likely that a particular judge will rate those scenarios as falling within the scope of the provision. At the same time, providing a list of examples (even if the list is not meant to be exhaustive), may also make it likely that a judge will deem scenarios that are not explicitly mentioned in that list as not falling within the scope of the provision. See e.g. Scalia and Garner, n 15 above at 107; (explaining that United States judges frequently invoke the canon of construction known as *expressio unius est exclusio alterius*, whereby when a statute expresses something explicitly, as in a list, anything not expressed explicitly does not fall within the statute). See also Tobia, Slocum, and Nourse, n 38 above at 269 (finding, in a sample of 122 law students and 1478 laypeople, that 67% of law students and 59% of laypeople implicitly invoked the *expressio unius est exclusio alterius* canon in legal contexts).

[44] At first glance, this result may seem counterintuitive. Human beings frequently treat low-probability risks as if they were zero, so one might expect them to give lower rather than higher weight to low probability scenarios. See M. Bazerman and M. D. Watkins, Predictable Surprises: The Disasters You Should Have Seen Coming, and How to Prevent Them (Boston: Harvard Business Review Press, 2004), 84-87; see also Sunstein, 'The Catastrophic Harm Precautionary Principle' n 12 above, 4; C. R. Sunstein, 'The Catastrophic Harm Precautionary Principle' (2007) 6 *Issues in Legal Scholarship* 1, 4 suggesting that "catastrophic risks may be entirely ignored for just this reason." However, other well-established findings in the behavioral sciences may explain our results. One explanation is that the low probabilities in our examples may not be sufficiently low to be perceived as zero. cf. Winter, n 36 above, 259. Another explanation could be that participant's choices are influenced by the imagination of potential real-world scenarios, in which, according to Sunstein, "it is plausible to think that the loss of 200 million people is more than 1000 times worse than the loss of 2000 people. ... Here too we are speaking of expected value, but emphasizing that the expected value of a catastrophic harm is much higher than what emerges by a simple exercise in multiplication, as in the idea that 200 million deaths is worse than 200 times a million deaths." Sunstein, 'The Catastrophic Harm Precautionary Principle' n 12 above, 5-6. While Sunstein further suggests that the (un)availability heuristic—, ie the tendency to heavily weigh judgments based on information that is available and can be readily recalled— may explain why catastrophic risks are neglected (see also Yudkowsky, n 36 above; Ord, n 4 above; E. Martínez and C. Winter, 'Experimental Longtermist Jurisprudence' in S. Magen and K. Prochownik (eds), *Advances in Experimental Philosophy of Law* (Bloomsbury Academic, forthcoming)), our study design may have made these risks more cognitively "available," thus mitigating the effect of this bias. Moreover, participants in our study may have suffered from the so-called "zero-risk bias", ie the tendency to prefer the elimination of low probability risks, even if alternative options produce a greater expected value. In other words, individuals overweigh small risks and are willing to pay more than the expected value to eliminate them altogether. For short overviews of the zero-risk bias, see D. Kahneman, *Thinking Fast and Slow* (Farrar, Straus and Giroux, 2011), 315; Winter, n 36 above, 258, 259.

Regardless of the precise reason for why participants chose as they did, these results cast some doubt on Sunstein's assumption that a version of his Catastrophic Harm Precautionary Principle, which is based on expected value considerations, "might well provide more protection than accords with ordinary intuitions." Sunstein, 'The Catastrophic Harm Precautionary Principle' n 12 above, 4. Instead, we find here implies that, insofar as lawmakers do not want the determination of whether a scenario is an existential risk to be based on a scenario's expected level of harm, they should specify that accordingly in their provision.

Moreover, although the focus of our study was on lay adults in the United States, this research program can and ought to be extended beyond this jurisdiction. The ordinary meaning doctrine is used in jurisdictions across the world, and further research could determine the extent to which these findings hold across the English-speaking world and beyond. For example, previous work in experimental jurisprudence has investigated and verified the degree to which ordinary people's beliefs regarding certain procedural principles (eg laws applied retrospectively or unintelligible laws[45]) and interpretive principles of law (eg textualism vs purposivism[46]) are stable across cultures. Similarly, one might investigate how laypeople in countries outside the United States interpret different terms for existential risk, so lawmakers in other countries can make a better-informed decision about what term to use when drafting an analogous provision.

In addition to existential risk, this research program has broader implications for ordinary meaning analysis more generally. Currently, the standard approach in ordinary meaning research—which we refer to as *ex post* ordinary meaning analysis—is backwards facing; that is, it focuses on evaluating tools judges might use to determine (a) what the [idealized] legislature intended a particular term to mean or (b) the textual meaning of a term, independent of what the legislature intended[47]. However, ordinary meaning analysis as conducted in this study—which we refer to as *ex ante* ordinary meaning analysis—is also forward facing; that is, it can help lawmakers decide which words to use by anticipating how an ordinary person might understand key terms, and in doing so guide judges (and the public) toward their intended meaning, such that legislative rules are appropriately understood, enforced, and adjudicated.

that ordinary intuitions at least in this particular study follow expected value, and if anything, are even more risk averse (but cf. also Study 1a), and therefore arguably provide even more protection, than some versions of the Catastrophic Harm Precautionary Principle.

Further note that placing some extra weight on low probability scenarios might additionally be justified by existing legal mechanisms, such as the precautionary principle, if there is a plausible risk that is deemed to be "low probability" due to lack of conclusive evidence. Cf., among others, Treaty on the Functioning of the European Union 2016 (OJ C 202), Art. 191. For international law, see generally A. Trouwborst, *Evolution and Status of the Precautionary Principle in International Law* (Netherlands: Springer, 2002). For criticism of the precautionary principle, see C. Sunstein, 'Beyond the Precautionary Principle' (2003) 151 *University of Pennsylvania Law Review* 1003.

[45] I. R. Hannikainen et al., 'Are There Cross-Cultural Legal Principles? Modal Reasoning Uncovers Procedural Constraints on Law' (2021) 45 *Cognitive Science* e13024.

[46] I. R. Hannikainen, K. P. Tobia, et al., 'Coordination and Expertise Foster Legal Textualism' (2022) 119 PNAS e2206531119.

[47] See eg Tobia, n 15 above; Lee and Mouritsen, n 15 above; Klapper, Schmidt, and Tarantola, n 15 above; see also Tobia, Slocum, and Nourse, n 38 above.

V. Conclusion

Recent scholarly and legislative efforts have sought to protect present and future generations from existential and catastrophic threats associated with climate change, nuclear war, artificial intelligence, and pandemics. This article informs these and other efforts through four experimental studies investigating how ordinary people interpret legal provisions referencing existential risk and related terms. The results of these studies advance our understanding of the *ordinary meaning* of existential risk, providing important insights both for lawmakers drafting existential risk legislation and for judges tasked with interpreting and applying this legislation.

Beyond existential risk, our study also offers insight into the coherence and justification of the ordinary meaning principle more generally, and lays the foundation for a new research program within legal interpretation research that we refer to as "*ex ante* ordinary meaning analysis"—focused not only on how judges can and should interpret legal provisions once they have been drafted, but on how lawmakers can and should draft legal provisions so as to best achieve their policy aims.