On Modeling Human Perceptions of Allocation Policies with Uncertain Outcomes

HODA HEIDARI
Carnegie Mellon University
and
SOLON BAROCAS
Microsoft Research & Cornell University
and
JON KLEINBERG
Cornell University
and
KAREN LEVY
Cornell University

Many policies allocate harms or benefits that are *uncertain* in nature: they produce distributions over the population in which individuals have different probabilities of incurring harm or benefit. Comparing different policies thus involves a comparison of their corresponding probability distributions, and we observe that in many instances the policies selected in practice are hard to explain by preferences based only on the expected value of the total harm or benefit they produce. In cases where the expected value analysis is not a sufficient explanatory framework, what would be a reasonable model for societal preferences over these distributions? Here we investigate explanations based on the framework of *probability weighting* from the behavioral sciences, which over several decades has identified systematic biases in how people perceive probabilities. We show that probability weighting can be used to make predictions about preferences over probabilistic distributions of harm and benefit that function quite differently from expected-value analysis, and in a number of cases provide potential explanations for policy preferences that appear hard to motivate by other means. In particular, we identify optimal policies for minimizing perceived total harm and maximizing perceived total benefit that take the distorting effects of probability weighting into account, and we discuss a number of real-world policies that resemble such allocational strategies. Our analysis does not provide specific recommendations for policy choices, but is instead interpretive in nature, seeking to describe observed phenomena in policy choices.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Economics

General Terms: Economics, Human factors, Theory

Additional Key Words and Phrases: probability weighting, uncertain allocations, human perceptions, harm minimization, benefit maximization

1. INTRODUCTION

Societies frequently wrestle with tough decisions regarding the allocation of benefits or burdens among their populations (see, e.g., [Calabresi and Bobbitt 1978; Viscusi 2018]). These decisions—particularly those that involve harm—are immensely dif-

Exemplary Applied Modeling Track Paper at EC'21.
Authors' addresses: hheidari@cmu.edu, {sbarocas, kleinberg, karen.levy}@cornell.edu

ficult yet often unavoidable. As Sunstein points out, governments regularly pursue policies that lead to harms, including death, among the public: *"If government allows new highways to be built, it will know that people will die on those highways; if government allows new power plants to be built, it will know that some people will die from the resulting pollution. [...] Of course it would make sense, in most or all of these domains, to take extra steps to reduce risks. But that proposition does not support the implausible claim that we should disapprove, from the moral point of view, of any action taken when deaths are foreseeable."* [Sunstein 2003] These considerations remain true even when the prospective harms are reduced as much as possible; to the extent that harms remain, we must reason about the impact of policies that produce foreseeable harms.

To make matters more complicated, many of these allocations deal in *probabilities* of some outcome occurring: when we raise the speed limit by a certain amount, for example, we can estimate to some approximate level the number of additional traffic fatalities that will result [Farmer 2019], but we can say much less about who in particular will die. Thus, for matters involving harm, the policy process necessarily involves a set of choices (even if these choices arise only implicitly) between different *distributions* of harm over the population. For example, policy P might produce a probability p_i that individual i is harmed, while policy Q might produce a probability q_i that individual i is harmed, for each individual in the population. (To keep the discussion simple, we will think about a single kind of "harm" that can befall people as a result of the policy, rather than adding the complexity of different types or degrees of harm.)

How should we compare the two distributions of harm that arise from policies P and Q? Much of the work that underpins mathematical models in these domains, including many of the loss functions that go into algorithmic decisions, tend to be based on expected cost—the idea that we should favor the policy that produces the lower expected harm. In our case, policy P produces a sequence of probabilities $(p_1, p_2, ..., p_n)$ over the n members of the population, and its expected harm is the sum $p_1 + p_2 + \cdots + p_n$; we can write a similar expression for the probabilities of harm $(q_1, q_2, ..., q_n)$ produced by policy Q.

Of course, real-life policymaking is complex, and it is not clear that minimization of expected harm is typically the chief criterion in selecting among policy options. But there is a more basic problem with using expected harm as the criterion: many policy questions about competing distributions of harm begin after we've already reduced the total amount of harm to a roughly fixed, low target level, and so the debate is among distributions that all have the same expected level of harm. How, then, should we think about preferences among these competing policy proposals?

1.1 A Real-life Example

We can see the outlines of such debates in a number of settings where a risk of harm is being allocated across a population. In the policies for drafting people into the military in the United States, for example, the government has considered a number of different implementations for randomizing the selection of inductees. (Here, required service in the military is the cost, or harm, that is being allocated according to a probability distribution.) Under a given policy P, individual i would learn that they had a probability p_i of being drafted. Crucially, difficult questions

about the implementations of draft systems persist regardless of the desired size of the military; that is, for a given size of the military, the sum of the draft probabilities p_i over the population is pinned to this number, but some distributions of these probabilities have been nonetheless viewed as preferable to others.

What accounts for these preferences? We note that discussions of revisions to the draft framed uncertainty itself as a cost being borne by members of the population. As the U.S. Selective Service System notes, prior to the introduction of a structured process for randomization, men knew only that they were eligible to be drafted from the time they turned 18 until they reached age 26; *"[this] lack of a system resulted in uncertainty for the potential draftees during the entire time they were within the draft-eligible age group. All throughout a young man's early 20's he did not know if he would be drafted"* [Selective Service System 2020]. The systems that were subsequently introduced specified priority groups according to age, which had the effect of deliberately producing non-uniform probabilities of being drafted in any given year; under these systems, some people were selected with higher-than-average probability and others with lower-than-average probability.[1] Viewed in terms of distributions, these policy changes had the effect of *concentrating* the probabilities more heavily on a subset of the eligible population each year, rather than *diffusing* the probabilities more evenly across everyone.

The quote from the Selective Service System points out that a process that diffuses probabilities too widely seems to create unnecessary (and harmful) levels of uncertainty; but there are, of course, corresponding objections that could be raised to processes that concentrate probabilities too heavily on too small a group.

An abstraction of these questions would therefore consider multiple probability distributions of harm—for example, policy P producing $(p_1, p_2, ..., p_n)$, policy Q producing $(q_1, q_2, ..., q_n)$, and perhaps others—and ask which of these should be preferred as a choice for society. In posing such questions, we are guided by the belief that studying reactions to distributions of harm should draw closely on those parts of the behavioral sciences that have considered how people subjectively evaluate probabilities. We therefore develop a framework based on the concept of *probability weighting* from behavioral economics.

Our model will allow us to evaluate the Selective Service System's argument, and similar arguments in other domains, at a broad level—the contention that completely uniform randomization over the draft-eligible population is a sub-optimal policy because the cumulative level of uncertainty felt by the population is unnecessarily high. At first glance, this argument is counter-intuitive: since the size of the military is the same under all the draft policies being considered, isn't the cumulative level of uncertainty felt by the population also the same under all policies? On closer inspection, though, we find that this decision—to shift the probabilities in a non-uniform direction, and to interpret this as reducing cumulative uncertainty—is

[1]Specifically, men were drafted according to "priority year," with the youngest men being drafted first. During the year a man was 20 years old, he was in the top priority group, with reduced likelihood of being called up each subsequent year. Within each group, call-up order was randomized by lottery according to birthday [Selective Service System 2020]. This prioritization based on a known random ordering of birthdays served as an additional way of concentrating the probabilities on a subset of the population.

very much consistent with the predictions of probability weighting.

2. MOTIVATING THE MODEL

We can adapt our discussions about harm allocations—and complex scenarios such as the military draft—into a stylized example in which a fixed amount of harm must be allocated across a given population. We will argue that different allocations of harm have very different subjective resonances, and it is these differences that behavioral theories of probability weighting aim to illuminate.

Thus, as a thought experiment, consider the following hypothetical example. Suppose we need to allocate 1 unit of harm among 100 individuals. For simplicity, let's assume all 100 individuals are equally deserving and willing to bear the harm. We might allocate the harm to one specific person (say, Bob), while giving the other 99 people certainty that they are not at risk—hence the probability distribution $(1, 0, \cdots, 0)$. Feeling sorry for Bob, we might instead divide the risk between him and another member of the population, Chloe—and ultimately flip a coin to decide which of them is to bear the harm, while the other 98 people are free and clear; i.e. the distribution $(1/2, 1/2, 0, \cdots, 0)$. Or we could have a third person, David, join Bob and Chloe in the risk pool, lowering the risk for each of them to one-third $(1/3, 1/3, 1/3, 0, \cdots, 0)$. Finally, we might allocate the risk evenly among all 100 individuals, and select the recipient of the harm by random lottery: $(0.01, \cdots, 0.01)$.

How might a policymaker select among these policies? Each of them, ultimately, results in the same amount of harm (1 unit) befalling the population, yet they strike us as intuitively quite different. We may consider it blatantly unfair to single Bob out as a certain victim by concentrating the risk completely on him; and indeed, a long line of work in psychology on the so-called *identifiable victim effect* suggests that we tend to find such outcomes particularly troubling [Jenni and Loewenstein 1997].[2] On the other hand, a random lottery distributes the risk equally among all 100 individuals—but in the interim, it forces *everybody* to worry about their chances of being harmed. (This is the form of uncertainty, and corresponding psychological cost, that the Selective Service System was concerned with in our example of the draft lottery.) The second and third options provide intermediate alternatives. In the second alternative, no one person is harmed with *certainty*, while, at the same time, the smallest possible number of individuals need bear the risk.

The fact that we may prefer some of the above alternatives to others immediately suggests that a cost-benefit analysis based on expected harm is not sufficient to capture our intuitions—since all the options involve the same expected amount of harm. Likewise, our intuitive reactions to these different proposals do not neatly map onto common concerns with distributive justice, where we tend to worry about the relative impact of allocations on different social groups or subgroups within

[2]Philosophy has also grappled with the observation that we tend to recoil at the idea of, for example, harvesting one person's organs to save the lives of five other people. Such cases reveal an intuitive distaste for distributions that aim to reduce the overall amount of harm experienced by a population by focusing those harms on a small subset of people [Thomson 1976]. Note that our framework does not apply to these cases because concentrating costs in these instances actually reduces the total cost (e.g., reducing the total number of deaths from five to one); in our settings, the way a policy allocates harms does not affect the amount of harm imposed on the overall population.

the population, given existing social inequalities. In this case, our reactions have nothing to do with any details about who Bob, Chloe, and Dave happen to be or the social groups to which they belong. What we perceive to be the more desirable allocation instead seems to rest on how we perceive the benefits or harms of being subject to uncertain outcomes.[3]

An interpretive analysis: Our intention in exploring people's subjective perceptions of risk probabilities is, emphatically, *not* to prescribe a "best" mode of allocating probabilities of risk, nor to endorse the underlying policy decisions that give rise to a need to allocate such risk in the first place, nor to treat superficially the variety of other procedural and moral concerns that attend the allocation of harms and benefits to people. Ours is a purely *interpretive* undertaking; we find that preferences for certain allocation policies involving probabilities are difficult to explain unless we take probability weighting into account.

Policy experts disagree about the extent to which cognitive errors ought to be explicitly incorporated into account in public decision-making. While some consider it inappropriate to base policies on what are essentially misunderstandings, others suggest that we might reasonably consider the "psychic benefits" to the public of protecting against "imaginary" risks [Schneier 2008; Viscusi 2018; Portney 1992; Pollak 1998]. We stake no claim in this debate; our goal is to explore descriptively how people's subjective perceptions of probabilities *might* impact preferences regarding such allocations—and how these impacts potentially explain peculiar real-life allocation policies. In this way, our work follows a style of research that seeks to shed light on observed policy outcomes by linking them to our behavioral understanding of latent human preferences for certain types of outcomes over others (see, e.g., [Srivastava et al. 2019; Zhu et al. 2018; Lee et al. 2019] for earlier work in this genre).

All of this still leaves us with a basic question. We have seen examples so far (with others to come) of policy-making favoring some level of randomization, while also steering away from completely uniform randomization that would spread risk of harm diffusely across a population. Is there a model that predicts this type of "intermediate" position that avoids both a concentration of risk on identifiable victims as well as too diffuse a distribution over the whole population? And can such a model be derived from known psychological models of human behavior? In this work we will argue that a preference for these types of intermediate distributions of risk can be derived naturally from the concept of *probability weighting*, one of the most empirically well-grounded human biases studied in behavioral economics [Kahneman and Tversky 2013], to which we now turn.

[3]To put it differently, the purpose of our work is *not* to argue that probability weighting tends to result in distributions that disproportionately harm members of specific social groups. Rather, we study human perceptions toward distributions that allocate the same type of harm unevenly across *otherwise-equal* individuals (without specifying their group memberships). Given the centrality of probability weighting in the empirical study of behavioral biases around uncertainty, we believe that showing how a range of distributional considerations arises purely from probability weighting is of interest independent of the possibility of additional biases.

2.1 A Model Based on Probability Weighting

Motivated by the premise that understanding people's perceptions of harm/benefit allocations are crucial in designing acceptable policies, we posit that models that solely rely on expected value comparisons may miss crucial aspects of human perceptions toward uncertain allocations—which are in part shaped by probability weighting. As a result, expected value-optimizing algorithms may produce allocations that are behaviorally repugnant to people. Our model can partially explain these reactions using one of the fundamental principles in behavioral sciences. To our knowledge, our work is the first to explore the attractiveness of different uncertain allocation policies by exploring *optimal allocations* under probability weighing. We make several connections between the optimal allocation patterns suggested by our theory and real-world policy choices that would be otherwise difficult to explain.

Probability weighting begins from the qualitative observation that people tend to overweight small probabilities—behaving as though they are larger than they actually are—and tend to underweight large probabilities—behaving as though they are smaller than they actually are. More generally, probability weighting is the premise that when faced with an uncertain event of probability p, people will tend to behave with respect to this event—for example, when determining risks or evaluating gambles involving the event—as though its probability were not p but a value $w(p)$, the *weighted version* of the probability. This weighting function $w(p)$ has the two properties noted above: that $w(p)$ is larger than p when p is small, and $w(p)$ is smaller than p when p is large. If we think in terms of the graph of $w(p)$ as a function of p, people refer to these properties as the "inverse S-shaped" nature of the probability weighting curve. There are a number of different models that derive inverse S-shaped probability weighting curves from simple observations; one influential functional form was provided by Prelec [Prelec 1998], who derived it from a set of underlying axioms about preferences for different types of gambles. The concept of probability weighting has been invoked to explain a number of peculiar behavioral patterns; one of the canonical examples is people's participation in gambling and lotto games [Quiggin 1991; Kahneman 2011].[4]

We use probability weighting here to ask the following basic question. Suppose there are r units of harm to be allocated across a population of n people, and we are evaluating policies that assign individual i a probability p_i of receiving harm, subject to the constraint that the sum of p_i over all individuals i is r. In the motivating settings discussed so far, it is natural to think of the cost borne by individual i as the perceived probability $w(p_i)$—either because individual i perceives it this way (via the psychological cost of their own uncertainty) or because the rest of society views it this way (via our discomfort at the idea that i is an identifiable victim with a perceived probability $w(p_i)$ of being harmed). We can therefore ask: which probability distribution minimizes this total cost, the sum of $w(p_i)$ over all

[4]To elaborate further on this connection, note that the cost of buying a lotto ticket is always set to be higher than the expected benefit (i.e., the likelihood of winning times the prize); otherwise, lottery operators would lose money. Nonetheless, people participate in these games in large numbers. Work in behavioral economics has advanced probability weighting as one explanation for this irrational behavior, via the tendency to over-weigh small probabilities—here, the chance of winning the lottery [Quiggin 1991; Kahneman 2011].

individuals i? Notice that this question allows for distinctions among probability distributions that all produce the same total expected harm for the population: in particular, all the distributions under consideration have a total expected harm of r, but they can nevertheless differ substantially in the sum of $w(p_i)$ over all individuals i.

3. OVERVIEW OF FINDINGS

We find that the distributions minimizing the weighted sum of harm probabilities $w(p_i)$ in fact correspond to intermediate distributions of the type we have been discussing qualitatively: distributions that concentrate the risk on a subset of the population, such that each member of the at-risk subset has a probability of harm that is strictly less than 1, while most of the population has a probability of harm equal to 0. The analysis leading to this conclusion involves some subtlety: sums of S-shaped functions do not exhibit the nice properties that simpler function classes do, and so minimizing them requires additional complexity in the analysis.

With this model in place, we can also explore the natural complement to this dynamic. Our discussion thus far has focused on probabilities of *harm*, but there is an analogous class of questions about distributing probabilities of *benefit* across a population—for example, in the availability of opportunities like higher education or financial assistance programs. Suppose there are r units of benefit available to the population as a whole, and we are considering policies that assign a probability p_i that individual i receives the benefit. Which distributions maximize the sum of $w(p_i)$ over all individuals i—that is, maximizing the total perceived benefit? As with risks of harm, we do not argue that such a policy is necessarily desirable, only that it may have added or diminished attractiveness in its perceived impact; to the extent that such policies are favored in practice, the theory of probability weighting might therefore offer a suggestive description.

We find that the distributions maximizing this sum of perceived probabilities of benefit are quite different from the distributions minimizing the sum of perceived probabilities of harm. In particular, when the total available benefit r is small relative to the size of the population under consideration, the maximizing distribution is a uniform lottery which assigns all n people a probability of r/n; but as r increases, the maximizing distribution changes abruptly to one in which a subset of the population receives a portion of the benefit with certainty, and the rest of the population is given a uniform lottery for the remainder.

4. IMPLICATIONS

Given that a society developing policy seems to favor some probability distributions of harm or benefit over others, even when they have the same expected value, it is natural to ask whether a model based on probability weighting can shed light on the nature of these preferences. Our modeling activity thus works out what the favored policies would look like if society were seeking to maximize or minimize the total weighted probability. As we discuss in our work, properties of these minimizing and maximizing distributions *can* be observed in a variety of real-world settings. We consider a number of allocation policies that have been adopted in practice that involve distributions of uncertain harms and benefits that closely resemble

what our model suggests are optimal under probability weighting. Because the attractiveness of these policies is difficult to explain otherwise, we present them as inductive evidence that probability weighting may be playing a meaningful role in guiding societal preferences for certain allocations and in determining the actual distributions of harms and benefits in society.

REFERENCES

Calabresi, G. and Bobbitt, P. 1978. *Tragic choices.* Norton.

Farmer, C. M. 2019. The effects of higher speed limits on traffic fatalities in the United States, 1993–2017.

Jenni, K. and Loewenstein, G. 1997. Explaining the identifiable victim effect. *Journal of Risk and Uncertainty 14,* 3, 235–257.

Kahneman, D. 2011. *Thinking, fast and slow.* Macmillan.

Kahneman, D. and Tversky, A. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I.* World Scientific, 99–127.

Lee, M. K. et al. 2019. Webuildai: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction 3,* CSCW, 1–35.

Pollak, R. A. 1998. Imagined risks and cost-benefit analysis. *The American Economic Review 88,* 2, 376–380.

Portney, P. R. 1992. Trouble in happyville. *Journal of Policy Analysis and Management 11,* 1, 131–132.

Prelec, D. 1998. The probability weighting function. *Econometrica*, 497–527.

Quiggin, J. 1991. On the optimal design of lotteries. *Economica*, 1–16.

Schneier, B. 2008. The psychology of security. In *International conference on cryptology in Africa*. Springer, 50–79.

Selective Service System. 2020. Changes from vietnam to now. `https://www.sss.gov/history-and-records/changes-from-vietnam-to-now/`. Accessed: 2020-10-06.

Srivastava, M., Heidari, H., and Krause, A. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2459–2468.

Sunstein, C. R. 2003. Hazardous heuristics. *The University of Chicago Law Review 70,* 2, 751.

Thomson, J. J. 1976. Killing, letting die, and the trolley problem. *The Monist 59,* 2, 204–217.

Viscusi, W. K. 2018. *Pricing lives: Guideposts for a safer society.* Princeton University Press.

Zhu, H., Yu, B., Halfaker, A., and Terveen, L. 2018. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction 2,* CSCW, 194.