A Theory of Good Intentions*

Paul Niehaus
UC San Diego

November 15, 2013

Abstract

Why is other-regarding behavior often misguided? I study a new explanation grounded in the idea that altruists want to *think* they are helping. Frictions arise because perception and reality can diverge ex post when feedback is limited (as for example when donating to international development projects). Among other things the model helps explain why donors have a limited interest in learning about effectiveness; why intermediaries may market based on need, effectiveness, or neither; and why beneficiaries may not be able to do better than accept this situation. For policy-makers, the model implies a generic tradeoff between the quantity and quality of generosity.

*I thank Nageeb Ali, Jim Andreoni, Navin Kartik, Joel Sobel, Adam Szeidl, and seminar participants at Microsoft Research New England, Columbia, UCLA, and NEUDC for helpful comments. Microsoft Research New England provided generous hospitality.

1 Introduction

Other-regarding behavior poses a challenge for social scientists. On the one hand, some people are remarkably generous. Americans give about 2% of GDP to charity each year, for example.[1] This suggests that they care deeply about helping others. Yet in many cases generous people are also quite poorly informed about how to help effectively. For example, only 3% of charitable givers even *claim* to give based on research comparing the effectiveness of alternatives.[2] This pattern is so common that it is embodied in colloquial language, where "well-intentioned" is a euphemism for "poorly informed." Yet if people really are well-intentioned, why don't they *become* well-informed?

Economists have predominantly taken the view that funders want to be effective but find it difficult to learn how. Krasteva and Yildirim (2013) emphasize that the costs of learning may exceed the benefits for small donors. Development economists highlight the role of market failures: information about effectiveness is a public good (Duflo and Kremer, 2003; Levine, 2006; Ravallion, 2009), and communication from practitioners to funders is often distorted by strategic considerations (Pritchett, 2002; Duflo and Kremer, 2003; Levine, 2006). Institutions that produce and disseminate effectiveness research (e.g. CGD, J-PAL, IPA, CEGA) were created in part to address these concerns.

This paper examines an alternative (and complementary) interpretation: funders do not want to be more effective. Instead, they want to *think* that they are effective. Yet perception and reality can diverge. To illustrate the core premise, consider donating to help feed malnourished African children. This induces agreeable thoughts of children eating. Now suppose you learn that the charity in question is ineffective – perhaps an exposé reveals serious fraud. Presumably this reduces your satisfaction. What is more interesting is the counterfactual: if you had not learned of the fraud, you would have continued to experience "warm glow" (Andreoni, 1989) thinking about your impact *even though in reality no such impact existed*. Your altruistic preferences cannot literally be over children's outcomes as these occur on another continent, outside of your experience. Instead, perceptions count. This raises the question: how and how well will learning work in a market where perceptions *are* the product?

I study this question in a model of a single benefactor and beneficiary; the model thus abstracts from public goods issues. The benefactor does not know ex ante how his decisions will affect the beneficiary ex post. The unusual feature of the model is that this uncertainty persists ex post with positive probability. As a result the benefactor may face residual *ambiguity* which he must interpret. For example, a donor may receive no news about whether the charity he gave to was honest and have to decide what this implies. He cannot learn the correct interpretation through repeated experience, precisely because the true state remains unobserved. He therefore adopts the interpretation that maximizes his expected utility. This approach builds on evidence from psychology and economics that people tend to interpret information in a self-serving manner (Mobius et al., 2012).

[1] Author's calculation using data from The Giving Institute (2013) and the Bureau of Economic Analysis (`http://www.bea.gov/national/index.htm#gdp`, accessed 7 August 2013).

[2] See Hope Consulting (2012). The Hope sample over-represents wealthier donors and thus if anything likely overstates the amount of research done by the average donor.

The beliefs this yields have a seemingly innocuous structure: they are (endogenously) Bayesian, and they are consistent with the distribution of all observable data. As a result they are not readily falsifiable. For example, a well-intentioned donor correctly forecasts the probability that he will learn about a scandal involving his chosen charity. On learning of no scandals, however, the same donor assumes that "no news is good news" and views the charity as definitely honest. Because this effect appears only in the presence of ambiguity, the model predicts relatively standard decision-making when outcomes are observable (such as helping a neighbor) but relatively distorted behavior when outcomes are unobserved (such as helping internationally).

Given this interpretation strategy, the benefactor has mixed feelings about learning. On the one hand, he always prefers to avoid ex-post feedback as this constrains his beliefs. A donor who learns that his donation was stolen, for example, is directly worse off as this makes it difficult to believe that it was effective. On the other hand, the benefactor does want to obtain a limited amount of information ex ante, precisely in order to avoid such disappointments. Before donating, for example, a donor would like to know whether an unpleasant scandal will later break. The general result is that the benefactor prefers to do just enough research ex ante to accurately forecast the feedback he will receive ex post, but no more.

These motives in turn shape the marketing strategies that maximize revenue for intermediary organizations such as charities. Critics argue that these organizations provide too little information about effectiveness, with one writing that "useful information about what different charities do and whether it works isn't publicly available anywhere."[3] In the model, however, there is a sense in which this is simply good marketing. Intermediaries see their revenue fall in expectation if they commit to conducting an impact evaluation (formally, generating information about parameters that complement the benefactor's action). The reason is that the benefactor's interests are already aligned with those of the intermediary: he actually *wants* to believe the best about impact, and so further information is more likely to hurt than to help. Conversely, the intermediary benefits from marketing based on need. Formally, revenue increases in expectation with information about parameters that substitute for the benefactor's action. Need is a compelling strategy because of a conflict of interest between the parties: the benefactor wants to believe that things are not that bad, while the intermediary wants him to confront a harsher reality. This may help explain why nonprofit organizations often market using graphic depictions of need (e.g. "poverty pornography" images) and "awareness-raising" campaigns rather than cost-effectiveness claims or research on impact.

The result on effectiveness illustrates a broader theme: a tradeoff between the *quality* and the *quantity* of giving. From the point of view a policy-maker, good intentions are problematic as they may direct resources to relatively ineffective causes. For example, a new approach to poverty reduction with little concrete evidence may capture funders' imagination and attract large sums. The policy-maker could address this by sponsoring rigorous impact evaluation research. If (as expected) the results do not live up to the hype, funders will turn to alternatives. Definitionally, however, funders will be less excited about these alternatives than they originally were about the new approach. As a result, total giving will tend to fall. The

[3]GiveWell, `http://www.givewell.org/about/story`, accessed 10 September 2013.

policy-maker must therefore choose between a larger volume of poorly-informed funding and a smaller flow of better-informed giving. For the same reason even the beneficiary may prefer not to reveal the true state to the benefactor, as it may be better to receive large amounts of help in inefficient ways than to disillusion the giver.

Because it is explicitly built on utility from thoughts and perceptions, the model conveniently organizes a set of facts related to saliance. The link is simply that whatever brings those thoughts to mind tends to raise the return on giving. This helps explain, for example, why donors are more likely to support work on problems that have affected their loved ones (Small and Simonsohn, 2008), and why charities spend money to thank donors repeatedly for past gifts. It may also explain why charities encourage donors to think of their gifts as buying discrete, memorable item (e.g. cows) even when in reality (and in the fine print) they have no influence over fund allocations.

The results above for a "pure" altruist might plausibly set an upper bound on the effectiveness of people with more nuanced motives. Several have been proposed in the literature. Duncan (2004) argues, for example, that some donors care not about beneficiary's welfare per se but about the *impact* of their actions. More recently, Andreoni et al. (2012) present evidence suggesting that *guilt* plays a role. Section 4 applies the good intentions framework to a class of preferences that nest these motives as special cases, depending on the reference point against which the benefactor evaluates outcomes. For the benefactor this leaves matters qualitative unchanged: he continues to do a limited amount of research ex ante and avoid feedback ex post. For other players in the market, however, incentives may reverse. Impact philanthropists are a nonprofit's ideal customers, as they are completely aligned in their desire to believe that donations have a large marginal impact. Guilty givers, on the other hand, pose a challenge; they seek to assuage their guilt by convincing themselves that needs are exaggerated and that nothing they could do would ever really make a difference. They thus provide the sole case in which intermediaries can benefit from marketing based on effectiveness, as well as need.

How broadly applicable are these ideas? The model could be interpreted as describing any sort of other-regarding preference. Empirically the link is tightest to individual charitable giving where, as mentioned above, donors give but do not conduct much research. Donor's qualitative comments further highlight their use of interpretation and assumption. Donors told Hope Consulting (2012), for example, that "with known nonprofits, unless there is a scandal, you assume they are doing well with your money" (p. 38) and that "I don't research, but I am sure that the nonprofits to which I donate are doing a great job." (p. 42), leading the authors to conclude that "this creates a big challenge to getting people to do more research – they see no need to do so." (p. 44) Evidence from laboratory experiments corroborates this. Fong and Oberholzer-Gee (2011) find, for example, that only 1/3 of subjects are willing to pay \$1 to learn whether they are playing a \$10 dictator game with a disabled person or a drug user. Similarly Dana et al. (2007) find that only 56% of dictators choose to observe *free* information on the relationship between their actions and the recipient's payoffs.[4]

[4]The arguments may also apply to more localized gift-giving. Unwanted Christmas gifts, for example, are so common that there are websites devoted to displaying bad examples: knick-knacks, ugly sweaters, and so on (see for example `www.badgiftemporium.com` or `whydidyoubuymethat.com`). Waldfogel (2009) argues that holiday gift-giving is so wasteful that people should stop it entirely.

For institutional funders data are scarcer, but industry veterans have similar concerns. As recently as 2006 David Levine wrote to argue for "Building Learning into the Global Aid Industry;" by his count, "rigorous evaluations of the impacts of development programs remain rare. In its first 55 years, the World Bank published exactly zero. The U.S. Agency for International Development (USAID) had a better record: that organization funded one randomized study in the 1970s and another one in the 1990s" (Levine, 2006). Pritchett (2002) describes his years in the aid industry as "ignorant armies clashing by night," with "very rarely any firm evidence presented and considered about the likely impact of... proposed actions." Interestingly, Easterly (2006) emphasizes the role of faith and desire: "I feel like kind of a Scrooge... I speak to many audiences of good-hearted believers in the power of Big Western Plans to help the poor, *and I would so much like to believe them myself*" (emphasis added). No doubt this desire is only part of the story, alongside political and organization forces. But it is consistent with the idea that there is something fundamentally different about spending money on others' behalf.[5]

Conceptually the paper draws on and extends two strands of theoretical research. First, it takes quite literally Andreoni's (1989) influential idea that altruists benefit from the "warm glow" that their acts induce. Andreoni has emphasized that "the warm-glow hypothesis simply provides a direction for research rather than an answer to the puzzle of why people give – the concept of warm-glow is a placeholder for more specific models of individual and social motivations" (Andreoni et al., 2012). The present paper offers one such model linking warm glow to perceived outcomes.

Second, it draws inspiration from Brunnermeier and Parker's (2005) theory of optimal expectations. The key technical difference is that, unlike in their model, the decision-maker gets no utility from anticipation or remembrance and faces no tradeoff between anticipatory and flow utility; instead his *sole* objective is to hold pleasant thoughts. As a result he exhibits no cognitive dissonance – that is, no desire to hold beliefs other than those he holds in "equilibrium." More broadly, the paper builds on a tradition that emphasizes the effect of beliefs on well-being (e.g. Akerlof and Dickens (1982)). While this literature has focused on self-regard, its tenets must be at least as important for understanding other-regard.

The rest of the paper is organized as follows. Section 2 presents the framework and characterizes optimal interpretations. Section 3 characterizes learning, beginning with a simple example and concluding with general results. Section 4 extends this analysis to alternative motives for giving, and Section 5 discusses open questions for further research.

2 The Good Intentions Framework

2.1 Timing

There are two players, a benefactor and a beneficiary. Nature initially determines the value of a finite-valued parameter $\theta \in \Theta$ after which the timing of play is as follows:

[5]See also Brigham et al. (2013) who find that micro-finance institutions were unlikely to respond to emails mentioning research that microfinance was ineffective, but significantly more likely to respond to emails that mentioned positive results.

1. A signal $s_1 \in S_1$ is revealed and the benefactor forms subjective ex ante beliefs $\hat{\pi}(\theta, s_2|s_1)$
2. The benefactor chooses a decision $d \in D$
3. A signal $s_2 \in S_2$ is revealed and the benefactor forms subjective ex post beliefs $\hat{\pi}(\theta|d, s_2, s_1)$
4. Payoffs are realized

Let $\pi(\theta, s_2, s_1)$ describe the joint distribution of the observable data (s_1, s_2) and the unobservable parameter θ. No assumption is made that the benefactor knows this distribution, and its relationship to his beliefs is discussed below. The distribution π is fixed for now but will later be endogenized to characterize incentives for learning and communication.

2.2 Payoffs

The beneficiary's payoff depends on the decision d and state θ according to

$$v(d, \theta) \tag{1}$$

In the standard approach to modeling "pure" altruism, the benefactor's payoff would be

$$u(d) + v(d, \theta) \tag{2a}$$

The first term represents the benefactor's private concerns. For example, if $d \in [0, y]$ is a donation to a charitable cause then $u(d) = U(y - d)$ might be the benefactor's consumption utility. The second term represents the utility the benefactor obtains from the beneficiary's outcome. Note that this specification implies that the benefactor is *aware* of the ex-post realization of v. To allow for ex-post ambiguity, the benefactor's payoff must depend on his *perception* of v:

$$u(d) + \mathbb{E}_{\hat{\pi}(\theta|d,s_2,s_1)}[v(d, \theta)] \tag{2b}$$

This perception is captured by $\hat{\pi} \in \Delta(\Theta)$, the benefactor's ex-post subjective belief about the state of the world. The fact that $\hat{\pi}$ may be non-degenerate embodies the idea that uncertainty about θ may not completely resolve by the end of the game.

The altruism described by (2b) is still *pure* in the sense that, conditional on the level of u, the benefactor uses the same function v to assess the beneficiary's well-being as the beneficiary himself. The model thus abstracts from some of the wedges that earlier work has explored. A benefactor might have paternalistic preferences, for example, and care more about keeping the beneficiary from starving than about her other needs (e.g. Garfinkel (1973)). A benefactor might also help in part to signal his type (e.g. Glazer and Konrad (1996), Ali and Benabou (2013)). For simplicity I study pure altruism through Section 3 and then show in Section 4 how the framework can be extended to alternative motives.

2.3 Optimization

Given beliefs, the benefactor's decision-making process is standard: he chooses a decision d to maximize his subjective expected utility. Adopting the shorthand $\hat{\pi}$ for the complete

contingent belief profile $(\hat{\pi}(\theta, s_2|s_1), \hat{\pi}(\theta|d, s_2, s_1))$, we have

$$d^*(\hat{\pi}, s_1) = \arg\max \mathbb{E}_{\hat{\pi}(\theta, s_2|s_1)}[u(d) + v(d, \theta)] \tag{3}$$

The focus of the analysis will be on the evolution of beliefs and their effects on behavior through (3). I restrict the beliefs the benefactor may hold as follows:

Assumption 1 (Admissible beliefs)**.** *Subjective beliefs $\hat{\pi}(\theta, s_2|s_1)$ satisfy*

(a) $\hat{\pi}(\theta, s_2|s_1)$ is a probability measure on $\Theta \times S_2$ for any s_1

(b) $\hat{\pi}(\theta, s_2|s_1) = 0$ if $\pi(\theta, s_2|s_1) = 0$ for any (θ, s_2, s_1)

Subjective beliefs $\hat{\pi}(\theta|d, s_2, s_1)$ satisfy analogous conditions.

Part (a) of this assumption simply says that beliefs are well-defined. Part (b) is substantive and imposes a degree of logical consistency: the benefactor understands that some compound events are impossible and does not hold beliefs that are clearly incompatible with the facts. Beyond this, however, the relationship between probabilistic events may be ambiguous. For example, if the set $\{\theta : \pi(s_2, s_1|\theta) > 0\}$ has more than one element for some given (s_2, s_1) then it is unclear how the benefactor should weight their relative likelihood. Moreover, this problem does not go away with repetition of the game: because the benefactor does not observe θ ex post, he cannot learn about $\pi(\theta|s_2, s_1)$ regardless of how many i.i.d. draws of (s_2, s_1) he observes. I resolve this indeterminacy by studying beliefs that are optimal in the sense that they maximize expected utility.

$$\max_{\hat{\pi}} \mathbb{E}_{\pi}\Big[u(d^*(\hat{\pi}, s_1)) + \mathbb{E}_{\hat{\pi}(\theta|d^*, s_2, s_1)}[v(d^*(\hat{\pi}, s_1), \theta)]\Big] \text{ such that } \hat{\pi} \text{ is admissible} \tag{4}$$

Note the distinct roles played here by ex ante and ex post beliefs: while the former determine the mapping from signals s_1 into actions, the latter determine how the benefactor interprets the consequences of those actions.

2.4 Interpretation & Discussion

The "good intentions" framework departs from standard modeling techniques in two ways. First, the benefactor holds preferences over beliefs as well as over outcomes. This idea builds on a literature dating at least as far back as Akerlof and Dickens (1982), who model a employee who prefers to believe that his risk of workplace injury is low. More recently Caplin and Leahy (2001) study the effects on decision-making of anxiety about future payoffs, while Brunnermeier and Parker (2005) study the general problem of optimal beliefs when expectations about the future affect current happiness. As these examples illustrate the literature has focused on self-regarding beliefs; the argument here is that thoughts or beliefs are at least as important for understanding other-regard. When giving to Africa, for example, it is hard to see how anything *other* than beliefs could matter.

Second, the model explicitly endogenizes beliefs through optimization, in the spirit of Akerlof and Dickens (1982) and Brunnermeier and Parker (2005). A natural question is whether this leads to beliefs that are coherent either internally or with what the benefactor observes.

To examine this, note first that the benefactor's ex post belief $\hat{\pi}(\theta|d, s_2, s_1)$ affects his payoffs only through $\mathbb{E}_{\hat{\pi}(\theta|d,s_2,s_1)}[v(d,\theta)]$. He will therefore choose to be as optimistic as possible ex post about the beneficiary's situation. Formally, optimal beliefs put full weight on the state

$$\overline{\theta}(d, s_2, s_1) = \arg \max_{\theta \in \Theta : \pi(\theta|s_2,s_1)>0} [v(d,\theta)] \tag{5}$$

which is the best state of the world consistent with the information history. Given this, the benefactor's ex ante problem reduces to

$$\max_{\hat{\pi}} \mathbb{E}_{\pi}\Big[u(d^*) + v(d^*, \overline{\theta})]\Big] \tag{6}$$

where I have suppressed arguments for brevity. This says that the benefactor holds ex ante beliefs that induce optimal behavior, given that he will ultimately take the optimistic interpretation $\overline{\theta}$. One can then show that optimal beliefs are, without loss of generality, Bayesian.

Lemma 1 (Baysian Updating)**.** *There exist optimal subjective beliefs satisfying Bayes' rule, i.e.*

$$\hat{\pi}(\theta, s_2|s_1)\hat{\pi}(s_1) = \hat{\pi}(\theta, s_2, s_1)$$
$$\hat{\pi}(\theta|d, s_2, s_1)\hat{\pi}(s_2, s_1) = \hat{\pi}(\theta, s_2, s_1)$$

for all (θ, s_2, s_1)*.*

The proof (see Appendix A) is constructive and shows that beliefs derived as conditional probabilities from the prior

$$\hat{\pi}(\theta, s_2, s_1) = 1(\theta = \overline{\theta}(d^*(s_1), s_2, s_1))\pi(s_2, s_1) \tag{7}$$

are optimal. The interpretation of this specification is that the benefactor holds an unbiased view $\pi(s_2, s_1)$ of the *likelihood* of the various kinds of feedback he might receive, but chooses to *interpret* this feedback as proving that an appealing state of the world $\overline{\theta}$ has been realized. This has four noteworthy implications.

First, optimal beliefs have the usual mathematical properties of beliefs: for example, they behave as martingales. This implies that an empirical researcher cannot identify beliefs as "well intentioned" without ancillary data such as the empirical distribution π.

Second, optimal beliefs are consistent with observable data. Formally, the marginal distribution over (s_2, s_1) implied by (7) is the empirical distribution $\pi(s_2, s_1)$. This implies that the beliefs of a benefactor with unbounded time to learn about the model environment through repeated experience could converge to optimal beliefs. It is a corollary that optimal beliefs differ from the objective distribution only in describing data that are unobservable, i.e. the conditional distribution of θ given (s_2, s_1). Optimization is in this sense a mild assumption here relative to the literature, which has argued that people maintain optimistic interpretations even when these directly conflict with observable data. Brunnermeier and Parker (2005) argue, for example, that "psychological theories provide many channels through which the human mind is able to hold beliefs inconsistent with the rational processing of objective data"

(p. 1093). Mobius et al. (2012) show empirically that subjects interpret data about their ability with self-serving biases even when the data generating process is specified unambiguously and beliefs are elicited incentive-compatibly. In contrast, our focus here is on ambiguous questions such as the likelihood that a nonprofit executive is corrupt conditional on the absence of scandal, which provide even greater scope for the imagination.

Third, optimal beliefs are self-consistent: a benefactor holding them would not wish to alter them. To see this note that if the agent believes the true distribution is some $\hat{\pi}$ satisfying (7), and then uses (7) to re-calculate optimal beliefs, he arrives again at $\hat{\pi}$. (Note also that this need not hold for the empirical distribution π.) This property is one point of distinction between the model and others such as Brunnermeier and Parker (2005) in which agents hold self-inconsistent beliefs, reflecting the tension between between utility from actions and utility from beliefs. Here there is no such tension.

Fourth, the model nests the benchmark case of preferences over outcomes. To see this, consider evidence (s_2, s_1) that is consistent with only a single state $\theta : \pi(\theta|s_2, s_1) > 0$. For such evidence the only admissable interpretation is $\hat{\pi}(\theta|s_2, s_1) = \pi(\theta|s_2, s_1)$. Next, call feedback *fully revealing* if it always uniquely identifies the state, i.e. $\{\theta \in \Theta : \pi(\theta|s_2, s_1) > 0\}$ is single-valued for any (s_2, s_1) such that $\pi(s_2, s_1) > 0$. Then the following holds:

Lemma 2 (Role of Feedback)**.** *Beliefs derived via Bayesian updating from the prior $\pi(\theta, s_2, s_1)$ are optimal if feedback is fully revealing.*

In other words, the good intentions framework and the standard one coincide precisely when the benefactor expects no ex-post ambiguity about θ.[6] Intuitively, such cases are like decisions the benefactor makes which affect only himself. In these cases he directly experiences the consequences, which we can think of as a way in which he "learns" the realization of θ.

3 Effective Giving

How much will the benefactor learn in equilibrium when choosing how to help? I first illustrate the main ideas in an example and then provide generalizations in Section 3.5. For concreteness the narrative describes charitable giving.

3.1 An Example

Don, a marketing executive in Manhattan, considers giving to an NGO working to help Ben, a farmer in Africa. Don can donate any amount d up to total income y. Ben's welfare depends both on this donation and on other exogenous factors such as the level of rainfall or the effectiveness of the NGO. For simplicity, the situation is either Good ($\theta = \theta^g$) or Bad ($\theta = \theta^b$), where Ben's utility v satisfies $v(\theta^g, d) > v(\theta^b, d)$ for all d. Don's prior is that $\pi(\theta = \theta^g) \equiv \gamma \in (0, 1)$. Don genuinely wants to see Ben better-off, but since Ben is thousands of miles away this desire is reflected in preferences over *thoughts* about what is happening in

[6]The antecedent can be made both necessary and sufficient by adding appropriate sensitivity conditions.

Africa. Formally, Don maximizes

$$y - d + \hat{\gamma}_2 v(\theta^g, d) + (1 - \hat{\gamma}_2) v(\theta^b, d) \tag{8}$$

where $\hat{\gamma}_2$ is his subjective ex-post assessment of the likelihood that the situation is good. In each period he either observes θ or learns nothing. For example, interpreting θ as a measure of NGO effectiveness, he might or might not learn about an impact evaluation of its work. Interpreting θ as growing conditions, he might or might not read news about the state of African agriculture. Let p be the probability that he learns the truth before donating, and q the conditional probability that he learns it after donating if he had not learned it before.

If Don learns θ before donating then this pins down beliefs and he chooses

$$d^*(\theta) \equiv \arg\max_d y - d + v(\theta, d) \tag{9}$$

In the more interesting case where he does not learn before donating, he anticipates the views he will hold in the future. With probability q he will learn the true state, while with probability $1 - q$ he will obtain ambiguous information which he will interpret as meaning that all is well ($\theta = \theta^g$). His future perception is thus $\hat{\gamma}_2 = 0$ with probability $q(1 - \gamma)$ and $\hat{\gamma}_2 = 1$ with probability $1 - q(1 - \gamma)$. Given this, he optimally interprets the absence of news at time $t = 1$ to mean that matters in Africa are good with probability $\hat{\gamma}_1 = 1 - q(1 - \gamma)$[7] and gives

$$d^*(\emptyset) \equiv \arg\max_d y - d + \hat{\gamma}_1 v(\theta^g, d) + (1 - \hat{\gamma}_1) v(\theta^b, d) \tag{11}$$

3.2 Learning to Help

Don's tendency to take an optimistic view of things shapes his motives for learning. Consider first the effect on his payoffs of learning the truth ex post. If he already knew it then of course it has no effect. If it is news to him, however, then it cannot be welcome news. The reason is that when uninformed Don optimally reasons that "no news is good news" and believes all is well ($\theta = \theta^g$), while becoming informed may force him to confront the reality that things are in fact not well ($\theta = \theta^b$).

Observation 1. *Don's expected payoff strictly decreases in the probability that he becomes informed after donating.*

This observation highlights the idea that information is a *constraint*, ruling out hypotheses that formerly were plausible. Yet somewhat paradoxically, the constraining nature of ex post information can also make ex ante information endogenously valuable. To see this, suppose that Don knew he would definitely learn the truth ex post, and consider his demand for

[7] To see this note that this belief uniquely ensures

$$\arg\max_d y - d + \mathbb{E}_{\hat{\gamma}_1}[v(\theta, d)] = \arg\max_d y - d + \mathbb{E}_{(1-q(1-\gamma))}[v(\theta, d)] \tag{10}$$

Note that $\hat{\gamma}_1 = \mathbb{E}_\pi[\hat{\gamma}_2]$ so that the evolution of Don's beliefs satisfies the law of iterated expectations and with it Bayes' rule.

information ex ante. In this case his expected payoff is

$$(\gamma)\left[\max_d y - d + v(\theta^g, d)\right] + (1-\gamma)\left[\max_d y - d + v(\theta^b, d)\right] \tag{12}$$

when informed and

$$\max_d y - d + (\gamma)v(\theta^g, d) + (1-\gamma)v(\theta^b, d) \tag{13}$$

when uninformed. It follows directly from optimization and continuity that the former is strictly greater than the latter, so that Don values information. But now suppose that Don expects not to learn the truth ex post. In this case his payoff when informed ex ante is again given by (12), but his payoff when uninformed ex ante is

$$\max_d y - d + v(\theta^g, d) \tag{14}$$

He thus obtains a benefit from being *uninformed* proportional to

$$\max_d \left[y - d + v(\theta^g, d)\right] - \max_d \left[y - d + v(\theta^b, d)\right] \geq \max_d \left(v(\theta^g, d) - v(\theta^b, d)\right) > 0 \tag{15}$$

The intuition here, just as for ex post learning, is that information constrains the imagination. Absent any threat of real consequences, Don prefers maximum scope to "think positive."

Observation 2. *Don's payoff increases (decreases) in the probability he learns the truth before donating when he will (will not) learn the truth after donating.*

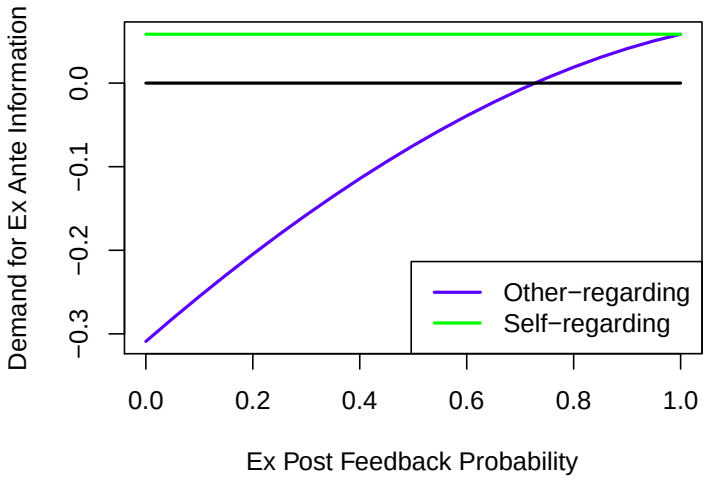
This observation summarizes a novel way of thinking about learning. The primal role of information is as a constraint on the imagination: it limits what thoughts one can reasonably entertain about the world. This makes it undesirable. On the other hand, given that such constraints are to be encountered, there is some value in knowing now what tomorrow's thoughts may be and acting so as to avoid disappointment. This generates positive demand.

Figure 1 illustrates the tension between the costs and benefits of learning with a parameterized example. When the probability that Don will learn the truth ex post is low he is strongly averse to learning the truth ex ante, as in all likelihood this will simply constraint his beliefs. As the probability of ex post learning rises his demand for ex ante research rises correspondingly until, past some threshold, it becomes positive. At all interior points his demand is strictly lower, however, than would be the case if he were making the decision for himself rather than for Ben.

Note that this result implies ex post feedback can stimulate demand for ex ante research. This is consistent with economicsts' arguments that *measuring* outcomes is necessary in order to force those spending money to pay attention to them. For example, Muralidharan (2012) writes of education policy in India that

> The Indian state has done a commendable job in improving the education indicators that were measured (including school access, infrastructure, enrollment, and inclusiveness in enrollment) but has fallen considerably short on the outcome indicators that have not been measured (such as learning outcomes). While independently measuring and administratively focusing on learning outcomes will not by itself

Figure 1: Demand for Ex Ante Information on Effectiveness



Notes: plots Don's willingness to pay for information for the case where $v(d, \theta) = \theta \log(d)$, $\theta^g = 2$, $\theta^g = 1$, and $\gamma = 0.2$, as a function of the probability he will learn the truth ex post.

lead to improvement, it will serve to focus the energies of the education system on the outcome that actually matters..."

3.3 Intermediaries

Don's ambivalent attitude towards learning in turn shapes the incentives of other players in the market. In this section I focus on a revenue-maximizing intermediary seeking to obtain donations from Don – for example, a charity. What marketing strategies maximize these donations? I focus here on the expected returns to generating various kinds of information which will then be disclosed to the public. This might correspond, for example, to commissioning an academic study by J-PAL.

Consider first the impact of better ex-post outcome measurement:

Observation 3. *Ex post feedback increases (decreases) expected generosity if v is submodular (supermodular).*

The probability of ex post feedback affects Don's decision only in the case where he is uninformed ex ante, so that his donation is given by (11). The comparative static is

$$\frac{\partial d}{\partial q} = \frac{(1-\gamma)[v_d(\theta^g, d) - v_d(\theta^b, d)]}{(1 - q(1-\gamma))v_{dd}(\theta^g, d) + q(1-\gamma)v_{dd}(\theta^b, d)} \tag{16}$$

which shares the sign of $v_d(\theta^b, d) - v_d(\theta^g, d)$. Next consider ex-ante research:

Observation 4. *Suppose that ex ante information does not affect expected generosity when ex post feedback is perfect. Then ex ante information strictly increases (decreases) expected generosity if v is submodular (supermodular) and feedback is limited.*

To see this, first consider the case of perfect feedback. Define $d^*(\gamma)$ as

$$d^*(\gamma) \equiv \arg\max_d y - d + \gamma v(\theta^g, d) + (1-\gamma)v(\theta^b, d) \tag{17}$$

If feedback is perfect ($q = 1$) then Don gives $d^*(\gamma)$ when uninformed, $d^*(1)$ if he obtains good news ex ante, and $d^*(0)$ if he learns bad news ex ante. Ex ante information thus has no average effect if $d^*(\gamma) = \gamma d^*(1) + (1-\gamma)d^*(0)$. Suppose this holds. Now consider the case with imperfect ex post feedback. If informed ex ante Don's expected donation is again $\gamma d^*(1) + (1-\gamma)d^*(0)$. If uninformed his donation solves (11). The solution to this equation is decreasing (increasing) in q if v is supermodular (submodular), and hence Don gives less (more) than $d^*(\gamma)$ when uninformed.

The mechanism underlying both these results is that Don prefers to believe that things are going well, so that information generally forces him to revise his beliefs negatively. How this affects his donation d then depends on whether giving is more or less impactful when the situation θ is bad. If θ complements donations – for example, if it measures effectiveness – then forcing Don to confront reality will lower his perception of marginal returns and depress giving. The intermediary has no incentive to do this. If, on the other hand, θ substitutes for donations – for example, if it measures Ben's baseline income – then forcing Don to confront reality will raise his perception of marginal returns and increase giving. Put another way, Don wishes to believe Ben is doing well, but the charity needs him to realize that Ben is desperately needy.

These results may help explain nonprofit marketing practice. Critics often lament how little rigorous information nonprofits provide about what they do and how impactful it is. Yet the model predicts that ambiguity on these dimensions is actually helpful, since it leaves space for donors to imagine the best. On the other hand, nonprofits often present information about need or use "awareness-raising" campaigns; these will be especially effective when altruists have a generic bias towards believing that others are doing better than they really are.

More broadly, the negative result for effectiveness research highlights a generic tradeoff in the model between the *quantity* and *quality* of altruistic activity. This is easiest to see from the perspective of a social planner seeking to maximize beneficiary well-being and choosing whether or not to sponsor research on effectiveness. While the research has the potential to increase the effectiveness of a *given* dollar of funding, it will also tend (according the result above) to reduce the total number of dollars given. It is thus unclear whether the beneficiary benefits. This has obvious implications for policy-makers allocating funds to development research. It also explains why the beneficiary may choose not to disillusion a well-intentioned donor even when given the chance (see Appendix B for a formal result).

3.4 Salience and Charitable Giving

By shifting emphasis from outcomes to thoughts, the good intentions model also provides a helpful framework for organizing some features of charitable marketing and giving related to salience that are hard to accomodate in standard models. To illustrate this, consider extending the model trivially by introducing a parameter $\rho \in (0,1)$ which measures the probability that Don thinks about Ben ex post. Then his expected payoff is

$$y - d + \rho \left[\hat{\gamma}_2 v(\theta^g, d) + (1 - \hat{\gamma}_2) v(\theta^b, d)\right] \tag{18}$$

This has several direct implications.

1. Donors give more to causes that are more memorable for them (higher ρ). This may help explain why people are more likely to give to issues that have affected friends and loved ones (Small and Simonsohn, 2008). For example, a donor who has lost a loved one to cancer is more likely to remember a gift supporting anti-cancer research through the associate property of memory (e.g. Tulving and Schacter (1990)).
2. As a corollary, charities can increase donations by making them more memorable. The most direct such strategy is of course to frequently remind the donor of his gift, and indeed "thank-you" notes are generally considered a good marketing practice.[8] Less obviously, charities can enhance recall of a gift by associating it with something specific and memorable. Linking a donation to an "identifiable victim" is one such strategy and has been show to increase giving (Jenni and Loewenstein, 1997). The use of "gift catalogues" may play a similar role; these allow donors to visualize their donation as leading to the provision of some specific, tangible thing (e.g. a goat) which they themselves "chose."[9]

3.5 General Functional Forms

This section generalizes the observations made using specific functional forms above. Doing so requires language to compare the information content of signals: a sense in which two signals are the same, and the standard Blackwell sense in which one is more informative than the other.

Definition 1 (Information equivalence)**.** *Random variables X and Y are informationally equivalent if there exists a bijection f such that $Y = f(X)$.*

Definition 2 (Blackwell garbling)**.** *Let $h(x, y, z)$ give the joint distribution of the random variables (X, Y, Z). X is a Blackwell garbling of Y with respect to Z if $h(x|y, z)$ is independent of z.*

[8] See for example `https://www.blackbaud.com/files/resources/downloads/WhitePaper_RecurringGiving.pdf`. Note that in the model Don's taste for reminders is ambiguous because v has no absolute unit: intuitively, thinking about Ben may make Don either happy or sad. Modifying Don's preferences along the lines suggested by Duncan (2004), so that Don cares about the *difference* his contribution made, resolves this ambiguity in favor of reminders.

[9] Gift catalogues are harder to rationalize as mechanisms for control, for two reasons. First, altruistic donors should not want control as they are unlikely to have good information about which interventions are most needed. Second and more importantly, donors' "choices" are typically not legally binding, as the accompanying fine print makes clear that the nonprofit will do whatever it wants with the donation. See for example `http://philanthropy.com/article/Holiday-Gift-Catalogs-Are/64374/`.

The shorthand $X \succsim Y$ indicates that the benefactor's expected payoff is weakly greater when he observers the random variable X than when he observers Y. We can now generalize Observation 2 and show that the benefactor prefers as little ex post feedback as possible.

Proposition 1. *Let random variable S_2' be a garbling of S_2 with respect to (S_1, θ). Then $S_2' \succsim S_2$.*

As above, the intuition is that feedback constrains the benefactor without helping him make decisions.

Proposition 2.

- *Let S_1 be informationally equivalent to S_2. Then $S_1 \succsim S_1'$ for any S_1'.*
- *Let S_1 be a garbling of S_2 with respect to θ and let S_1' be a garbling of S_1 with respect to S_2. Then $S_1 \succsim S_1'$.*

This generalizes Observation 2. The first part states that the benefactor's weakly prefers to observe ex ante what he will eventually observe ex post. In particular, he has no demand for information prior to making his decision that he will not subsequently learn after that decision. The second part states that, among signals that are strictly less informative than what he will observe ex post, the benefactor weakly prefers more informative ones. It is a corollary that he places a (weakly) positive value on such signals, since a white-noise signal is trivially a member of this set.

Generalizing Observation 4 requires a bit more work, as we need a generalization of the idea that ex ante information does not affect expected generosity under standard preferences (or equivalently, when ex post feedback is perfect).

Definition 3. *Suppose d is real-valued. The benefactor's preferences respect expectation if*

$$\arg\max_d \mathbb{E}_\mu[u(d) + v(d, \theta)] = \mathbb{E}_\mu[\arg\max_d u(d) + v(d, \theta)] \tag{19}$$

holds for any $\mu \in \Delta(\theta)$.

This condition says that, while particular realizations of θ may influence generosity one way or another, disclosure of θ neither increases nor increases generosity *in expectation*. We can now state and prove a general result on complementary and substitutability:

Proposition 3. *Suppose that Θ is ordered, D is real-valued, and $v(\theta, d)$ is monotone increasing in both arguments.*

- *Let S_2' be a garbling of S_2 with respect to θ. Then $\mathbb{E}_\pi[d]$ is higher (lower) under S_2' than under S_2 if v is supermodular (submodular).*
- *Let S_1' be a garbling of S_1 with respect to (S_2, θ) and suppose that the benefactors preferences respect expectation. Then $\mathbb{E}_\pi[d]$ is higher (lower) under S_1' than under S_1 if v is supermodular (submodular).*

Like Observation 4, this result implies that generosity tends to increase when information about needs is disclosed, but tends to decrease when information about effectiveness is disclosed.

4 Alternative Motives for Giving

The results above describe an altruist who is "pure" in the sense that she cares about (her perceptions of) the beneficiary's welfare. One might reasonably guess, then, that they are an upper bound on the effectiveness of an altruist with less aligned motivations. This section examines that guess. I consider a family of *reference-dependent* preferences of the form

$$u(d) + \mathbb{E}_{\hat{\pi}(\theta|d,s_2,s_1)}[v(d,\theta) - v(\overline{d},\theta)] \tag{2b'}$$

This specification extends that in (2b) by allowing the benefactor to care about her perceptions, not of the beneficiary's payoff $v(d,\theta)$ per se, but of the difference between that payoff and some reference payoff $v(\overline{d},\theta)$. The reference payoff is itself determined by a reference decision $\overline{d} \in D$ the benefactor could have made. In other words, the beneficiary thinks about the difference between what she *did* and something that she *could have done*.

This family nests at least two cases of interest from the literature. The first is the model of "impact philanthropy" proposed by Duncan (2004). In Duncan's model, a charitable giver cares about the difference between the outcome obtained when he gives and the counterfactual outcome that would have obtained had he given nothing. These preferences can be represented using as a reference point the decision $\overline{d} = \arg\max_d u(d) \in D$, or "doing nothing." In this case the benefactor effectively asks himself, how much better off is the beneficiary than if I had simply pursued my own private interests? His payoff is proportional to the (perceived) size of this gap.

The second case is "guilty giving." Andreoni et al. (2012), among others, have argued that other-regarding behavior is often motivated by a desire to close the gap between what one is doing and what one feels one could or should do. One simple way of capturing this idea is to let $\overline{d}$ measure what could or should be done. Loosely speaking, the benefactor then experiences pride when she does "more" than $\overline{d}$ but guilt when she does "less." We can isolate the latter motive by letting $\overline{d}$ represent a maximally generous action. If D is real-valued, for example, let $\overline{d} = \max D$. Note that this case describes an opposite extreme to the impact philanthropy model; together the two cases thus bookend the set of possible reference points.

Despite their variety, it turns out that the models described by (2b') share most properties of the base "pure altruism" case.

Proposition 4. *Lemmas 1 and 2 and Propositions 1 and 2 continue to hold replacing (2b) with (2b').*

The proof is by simple redefinition: let

$$\tilde{v}(d,\theta) \equiv v(d,\theta) - v(\overline{d},\theta) \tag{20}$$

and the proofs go through as before replacing v with $\tilde{v}$, since nothing in them relies on anything special about the structure of v. Conceptually the point here is that the disinterest in learning captured by these propositions, which may help explain ineffective altruism, is a product of perceptions management per se. It does not matter (qualitatively) whether the perceptions at stake are perceptions of outcomes, of impact, or of something else; the beneficiary will seek to

learn things she must eventually learn anyway, but otherwise will prefer to avoid information.

What about the market? Here it turns out – as professional fundraisers will attest – that knowing your donor's motives is essential for effective persuasion. Information has very different average effects on impact philanthropists and guilty givers, both of whom are different in turn from the "pure" altruist described above. The most ironic result is that the optimal strategy for a fundraising intermediary marketing itself to an impact philanthropist is to provide *no information at all*. The intuition is quite simple. An impact philanthropist wants to believe that the marginal impact of his dollar is as high as possible. He therefore interprets any ambiguous information as implying that the need is great and the available means of helping effective. Any information the intermediary provides will tend to lower his expectation of marginal impact, which in turn leads him to lower his donation.

For guilty givers, on the other hand, the reverse is true. To minimize guilt, these donors want to believe ex-post that there is nothing they could possibly have done that would have made any difference. Such a donor might convince himself, for example, that the need is not very great, or that all foreign aid is corrupt so that none of his donation would actually reach people in need. Such beliefs would enable him to give very little without experiencing guilt over missed opportunities. A fundraiser pitching such a donor could thus benefit by providing incontrovertible evidence of both need and efficacy. The donor, of course, would do his best to avoid this pitch.

The following Proposition formalizes these points. Note that this result is consistent with Proposition 3; the latter cannot be applied directly here since $v(d,\theta) - v(\overline{d},\theta)$ need not be increasing in θ even if v itself is.

Proposition 5. *Suppose that the benefactor's preferences are as in (2b'), Θ is ordered, D is real-valued, $v(d,\theta)$ is increasing in both arguments, and $v_d(d,\theta)$ is monotonic in θ.*

- *Let S_2' be a garbling of S_2 with respect to θ. Then $\mathbb{E}_\pi[d]$ is higher (lower) under S_2' than under S_2 if $\overline{d} = \min D$ ($\overline{d} = \max D$).*
- *Let S_1' be a garbling of S_1 with respect to (S_2,θ) and suppose that the benefactors preferences respect expectation. Then $\mathbb{E}_\pi[d]$ is higher (lower) under S_1' than under S_1 if $\overline{d} = \min D$ ($\overline{d} = \max D$).*

The impact philanthropy model case may also help explain the success of "matching grant" vehicles in fundraising. In a typical matching setup, an organization obtains a promise from a large funder to match subsequent smaller donations. The puzzle for economists is why such arrangements are credible: if the small donations do not materialize, will the large funder – who was clearly excited about funding the organization – really refrain from giving? This is exactly the sort of question an economist would ask – but exactly the sort of question a well-intentioned donor would *not* ask. An impact donor wants very much to believe that the large funder's commitment is credible, since this increases his marginal impact. He can do so, moreover, as long as there is ambiguity about counterfactual states. After donating himself, the donor simply needs to believe that the large funder would not have contributed if he had not. Fortunately for him, there is unlikely to be unambiguous evidence to the contrary.

5 Conclusion

Standard models of other-regarding behavior model benefactors with preferences over a beneficiary's outcomes. This approach is unrealistic as it posits that the decision-maker has preferences over events he never experiences. I study an alternative framework in which the benefactor has preferences over his beliefs about the beneficiary's outcomes. This framework nests the standard model in the special case where the benefactor obtains complete ex post information about the beneficiary's outcomes; absent perfect feedback the models' predictions diverge. Consistent with the motivation for the framework, the benefactor in the model endogenously prefers to avoid ex post feedback and also avoids ex ante information about the beneficiary except to avoid subsequent disappointment. The results may help explain a range of puzzles about effective giving ranging from poorly chosen holiday gifts to misspent charitable donations and foreign aid.

While static, the framework developed here is dynamically consistent in the sense that the benefactor holds beliefs that match the true distribution of observable variables. Formally modelling a dynamic extension could potentially shed further light on the evolution of altruism. Two specific conjectures seem worth examining. First, benefactor behavior will be self-perpetuating. A benefactor who takes an arbitrary action at time t will be motivated to believe this action was effective at time $t+1$, which will in turn motivate him to repeat the action. This may explain why nonprofits place such priority on the initial acquisition of donors. Second, benefactors may tend to become "jaded" over time as the accumulation of evidence increasingly constrains the extent to which they can "think positive."

References

Akerlof, George A and William T Dickens, "The Economic Consequences of Cognitive Dissonance," *American Economic Review*, June 1982, *72* (3), 307–19.

Ali, Nageeb and Roland Benabou, "Image versus Information," Technical Report, UC San Diego 2013.

Andreoni, James, "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence," *Journal of Political Economy*, December 1989, *97* (6), 1447–58.

_ , Justin Rao, and Hannah Trachtman, "Avoiding The Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving," Technical Report, UC San Diego June 2012.

Brigham, Matthew, Michael Findley, William Matthias, Chase Petrey, and Daniel Nelson, "Aversion to Learning in Development? A Global Field Experiment on Microfinance Institutions," Technical Report, Brigham Young University March 2013.

Brunnermeier, Markus K. and Jonathan A. Parker, "Optimal Expectations," *American Economic Review*, September 2005, *95* (4), 1092–1118.

Caplin, Andrew and John Leahy, "Psychological Expected Utility Theory And Anticipatory Feelings," *The Quarterly Journal of Economics*, February 2001, *116* (1), 55–79.

Che, Yeon-Koo, Wouter Dessein, and Navin Kartik, "Pandering to Persuade," *American Economic Review*, February 2013, *103* (1), 47–79.

Dana, Jason, Roberto Weber, and Jason Kuang, "Exploiting moral wiggle room: experiments demonstrating an illusory preference for fairness," *Economic Theory*, October 2007, *33* (1), 67–80.

Duflo, Esther and Michael Kremer, "Use of randomization in the evaluation of development effectiveness," Technical Report, World Bank 2003.

Duncan, Brian, "A theory of impact philanthropy," *Journal of Public Economics*, August 2004, *88* (9-10), 2159–2180.

Easterly, Bill, *The White Man's Burden: Why the West's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*, Oxford University Press, 2006.

Fong, Christina and Felix Oberholzer-Gee, "Truth in giving: Experimental evidence on the welfare effects of informed giving to the poor," *Journal of Public Economics*, 2011, *95* (5), 436–444.

Garfinkel, Irwin, "Is In-Kind Redistribution Efficient?," *The Quarterly Journal of Economics*, May 1973, *87* (2), 320–30.

Glazer, Amihai and Kai A Konrad, "A Signaling Explanation for Charity," *American Economic Review*, September 1996, *86* (4), 1019–28.

Hope Consulting, "Money for Good: The US Market for Impact Investments and Charitable Gifts from Individual Donors and Investors," Technical Report, Hope Consulting May 2012.

Jenni, Karen and George Loewenstein, "Explaining the Identifiable Victim Effect," *Journal of Risk and Uncertainty*, 1997, *14* (3), 235–257.

Kalai, Ehud and Ehud Lehrer, "Rational Learning Leads to Nash Equilibrium," *Econometrica*, September 1993, *61* (5), 1019–45.

Krasteva, Silvana and Huseyin Yildirim, "(Un)Informed Charitable Giving," *Journal of Public Economics*, 2013, *106*, 14–26.

Levine, David, "Learning What Works – and What Doesn't: Building Learning into the Global Aid Industry," Technical Report, UC Berkeley 2006.

Milgrom, Paul and Chris Shannon, "Monotone Comparative Statics," *Econometrica*, January 1994, *62* (1), 157–80.

Mobius, Markus, Muriel Niederle, Paul Niehaus, and Tanya Rosenblat, "Managing Self-Confidence: Theory and Experimental Evidence," Technical Report, UC San Diego August 2012.

Muralidharan, Karthik, "Using Evidence for Better Policy The Case of Primary Education in India," Technical Report, UC San Diego 2012.

Pritchett, Lant, "It pays to be ignorant: A simple political economy of rigorous program evaluation," *Journal of Policy Reform*, 2002, *5* (4), 251–269.

Ravallion, Martin, "Evaluation in the Practice of Development," *World Bank Research Observer*, March 2009, *24* (1), 29–53.

Small, Deborah A. and Uri Simonsohn, "Friends of Victims: Personal Experience and Prosocial Behavior," *Journal of Consumer Research*, October 2008, *35* (3), 532–542.

The Giving Institute, *Giving USA 2013*, Giving USA Foundation, 2013.

Tulving, E. and D. L. Schacter, "Priming and human memory systems," *Science*, January 1990, *247* (4940), 301–306.

Waldfogel, Joel, *Scroogenomics: Why You Shouldn't Buy Presents for the Holidays*, Princeton University Press, 2009.

A Proofs

Proof of Lemma 1

Consider the following family of history-contingent subjective beliefs:

$$\hat{\pi}(\theta, s_2, s_1) = 1(\theta = \overline{\theta}(d^*(s_1), s_2, s_1))\pi(s_2, s_1) \tag{21}$$

$$\hat{\pi}(\theta, s_2|s_1) = 1(\theta = \overline{\theta}(d^*(s_1), s_2, s_1))\pi(s_2|s_1) \tag{22}$$

$$\hat{\pi}(\theta|d, s_2, s_1) = 1(\theta = \overline{\theta}(d, s_2, s_1)) \tag{23}$$

where

$$d^*(s_1) = \arg\max_d \mathbb{E}_{\pi(s_2|s_1)}[u(d) + \mathbb{E}_{\hat{\pi}(\theta|d,s_2,s_1)}[v(d,\theta)]] \tag{24}$$

is the action the benefactor takes given these beliefs. It is straightforward to verify that the beliefs thus defined satisfy Bayes rule following any signal realizations. Intuitively, the benefactor retains objective beliefs about the distribution of signals (s_2, s_1) but distorts their *interpretation*, i.e. what these signals reveal about θ. To show that these beliefs also maximize the benefactor's payoff we need to show that they satisfy two conditions. First, if $\Theta(s_2, s_1)$ denotes the set of admissible beliefs upon observation of (s_2, s_1) then $\hat{\pi}(\theta|d, s_2, s_1)$ must solve

$$\max_{\tilde{\pi}\in\Theta(s_2,s_1)} \mathbb{E}_{\tilde{\pi}}[v(d,\theta)] \tag{25}$$

which it evidently does by definition. Second, $\hat{\pi}(\theta, s_2|s_1)$ is optimal if (though not necessarily only if) it induces the action that is optimal, i.e.

$$\arg\max_d \left[u(d) + \mathbb{E}_{\hat{\pi}(s_2|s_1)}[v(d,\theta)]\right] = \arg\max_d \left[u(d) + \mathbb{E}_{\pi(\theta,s_2|s_1)}\mathbb{E}_{\hat{\pi}(\theta|d,s_2,s_1)}[v(d,\theta)]\right] \tag{26}$$

This condition holds if

$$\hat{\pi}(\theta|s_1) = \mathbb{E}_{\pi(s_2|s_1)}[\hat{\pi}(\theta|d, s_2, s_1)] \tag{27}$$

$$= \mathbb{E}_{\pi(s_2|s_1)}[1(\theta = \overline{\theta}(d, s_2, s_1))] \tag{28}$$

$$= \sum_{s_2} 1(\theta = \overline{\theta}(d, s_2, s_1))\pi(s_2|s_1) \tag{29}$$

which follows from the definition of $\hat{\pi}(\theta, s_2|s_1)$ above.

Proof of Lemma 2

Proof. Suppose (s_2, s_1) is fully revealing; then we can write $\theta = f(s_2, s_1)$ for some function f. This implies that $\overline{\theta}(d, s_2, s_1) = f(s_2, s_1)$ and also that $\pi(\theta, s_2, s_1) = 1(\theta = f(s_2, s_1))\pi(s_2, s_1)$. We can now apply the construction used to prove Lemma 1 to show that beliefs derived via Bayesian updating from $\hat{\pi}(\theta, s_2, s_1) = 1(\theta = f(s_2, s_1))\pi(s_2, s_1) = \pi(\theta, s_2, s_1)$ must be optimal. □

Proof of Proposition 1

Fix a realization s_1. The benefactor's expected payoff if he observes S_2 is

$$u(d^*) + \sum_{s_2} \left[\max_{\theta \in \Theta(s_2, s_1)} \{v(d^*, \theta)\} \right] \pi(s_2|s_1) \tag{30}$$

where d^* is a decision that maximizes this expression. Now suppose instead he observes the realization of S_2'. Since d^* remains a feasible decision his payoff cannot be less than

$$u(d^*) + \sum_{s_2} \sum_{s_2'} \left[\max_{\theta \in \Theta(s_2', s_1)} v(d^*, \theta) \right] \pi(s_2'|s_2, s_1)\pi(s_2|s_1) \tag{31}$$

Now consider some realization (s_2', s_2, s_1, θ) observed with positive probability such that $\pi(s_2, s_1, \theta) > 0$ so that $\theta \in \Theta(s_2, s_1)$. We can write

$$\begin{aligned} \pi(s_2', s_2, s_1, \theta) &= \pi(s_2'|s_2, s_1, \theta)\pi(s_2, s_1, \theta) \\ &= \pi(s_2'|s_2)\pi(s_2, s_1, \theta) \\ &> 0 \end{aligned}$$

where the second step follows from the fact that S_2' garbles S_2 with respect to (S_1, θ) and the third from the fact that s_2' is observed. Thus for any realization we have $\Theta(s_2, s_1) \subseteq \Theta(s_2', s_1)$. This implies that the maximum in (31) is at least as great as that in (30) for any particular (s_2', s_2) and hence (31) is also greater in expectation. Since (31) is a lower bound on the benefactor's payoff when observing S_2, his actual payoff must also be weakly greater.

Proof of Proposition 2

Proof. **Part 1**. Fix the distribution of S_2. First note that because the benefactor chooses d after observing s_1 but then chooses $\overline{\theta}$ after observing both s_2 and s_1, his payoff is bounded above by

$$U(s_2, s_1) \equiv \max_{d, \theta \in \Theta(s_2, s_1)} u(d) + v(d, \theta) \tag{32}$$

which is the payoff he would obtain if he could choose d after observing both signals. Next, observe that when S_1 is equivalent to S_2 then the benefactor achieves this upper bound. Finally, note that when S_1 is not equivalent to S_2 then

$$\begin{aligned} \Theta(s_2, s_1) &= \{\theta \in \Theta : \pi(\theta|s_2, s_1) > 0\} && (33) \\ &\subseteq \{\theta \in \Theta : \pi(\theta|s_2) > 0\} && (34) \\ &= \Theta(s_2) && (35) \end{aligned}$$

and hence the constraint in (32) is weakly tighter than when S_1 is equivalent to S_2, so that $U(s_2, s_1)$ is weakly lower. Since this is an upper bound on the benefactor's payoff it implies that his realized payoff must also be weakly lower than when S_1 is equivalent to S_2.

Part 2. The proof follows the standard argument showing that information weakly im-

proves decision-making, with the caveat that we must also establish that observing a garbling of S_2 does not impose any additional constraints on beliefs.

Fix a realization s_1 of S_1. The benefactor's payoff when he observes this is

$$u(d^*) + \sum_{s_2} v(d^*, \overline{\theta}(d^*, s_2, s_1)\pi(s_2|s_1) \tag{36}$$

where d^* is the decision that maximizes this expression. If instead the benefactor were to observe s_1' then his payoff, again conditional on the (unobserved) value of s_1, is

$$u(d(s_1')) + \sum_{s_2} v(d(s_1'), \overline{\theta}(d(s_1'), s_2, s_1')\pi(s_2|s_1', s_1) \tag{37}$$

where $d(s_1')$ is the optimal decision given s_1'. To simplify this expression note that

$$\begin{aligned}
\pi(s_2|s_1', s_1) &= \frac{\pi(s_1'|s_2, s_1)\pi(s_2|s_1)\pi(s_1)}{\pi(s_1', s_1)} \\
&= \frac{\pi(s_1'|s_1)\pi(s_2|s_1)\pi(s_1)}{\pi(s_1', s_1)} \\
&= \pi(s_2|s_1)
\end{aligned}$$

where the key second step follows since s_1' is a garbling of s_1 with respect to s_2. Note also that

$$\begin{aligned}
\Theta(s_2, s_1) &= \{\theta : \pi(\theta, s_2, s_1) > 0\} \\
&= \{\theta : \pi(s_1|s_2, \theta)\pi(s_2, \theta) > 0\} \\
&= \{\theta : \pi(s_1|s_2)\pi(s_2, \theta) > 0\} \\
&= \{\theta : \pi(s_2, \theta) > 0\}
\end{aligned}$$

where the third step follows since s_1 is a garbling of s_2 with respect to θ and the last since $\pi(s_1|s_2) > 0$ for any observed realization. This implies that $\overline{\theta}(d, s_2, s_1)$ does not depend on s_1. An analogous argument shows that $\overline{\theta}(d, s_2, s_1')$ does not depend on s_1'. Exploiting these two facts we can rewrite (37) as

$$u(d(s_1')) + \sum_{s_2} v(d(s_1'), \overline{\theta}(d(s_1'), s_2, s_1)\pi(s_2|s_1) \tag{38}$$

which must by definition be weakly less than (36) since d^* is defined as the decision that maximizes that expression. □

Proof of Proposition 3

Proof. **Part 1**. Conditional on s_1, we can write the benefactors objective function as

$$f(d, \{x(s_2', s_2, s_1)\}) \equiv u(d) + \sum_{s_2} \sum_{s_2'} v(d, x(s_2', s_2, s_1))\pi(s_2'|s_2)\pi(s_2|s_1) \tag{39}$$

where

$$x(s_2', s_2, s_1) = \max\{\theta : \pi(\theta, s_2, s_1) > 0\} \tag{40}$$

in the case where he observes S_2 and

$$x(s_2', s_2, s_1) = \max\{\theta : \pi(\theta, s_2', s_1) > 0\} \tag{41}$$

in the case where he observes S_2'. (Note that we can write the distribution of S_2' in this separable form because it garbles S_2 and that x does not depend on d since v is monotone in θ.) Examining f, its latter argument is an element of a lattice with dimension support$(S_2) \times$ support(S_2'); moreover since S_2' garbles S_2 we have $\max\{\theta : \pi(\theta, s_2', s_1) > 0\} \geq \max\{\theta : \pi(\theta, s_2, s_1) > 0\}$ for any realization (s_2', s_2), so that S_2' induces a weakly larger element of this lattice than S_2. It then follows from the monotone comparative statics theorem (Milgrom and Shannon, 1994) that the solution is weakly greater (smaller) under S_2' if v is supermodular (submodular).

Part 2. Conditioning on any realization s_1' of S_1', the expected effect of observing S_1 instead can be written as

$$\sum_{s_1} \left[\arg\max_d u(d) + \sum_{s_2} v(d, \overline{\theta}(s_2, s_1))\pi(s_2|s_1) \right] \pi(s_1|s_1') \\ - \arg\max_d u(d) + \sum_{s_2} v(d, \overline{\theta}(s_2, s_1'))\pi(s_2|s_1') \tag{42}$$

Note that this statement exploits the fact that S_1 is finer than S_1' to write $\pi(s_2|s_1, s_1') = \pi(s_2|s_1)$ and $\overline{\theta}(s_2, s_1, s_1') = \overline{\theta}(s_2, s_1)$. By adding and substracting we can decompose this difference further as follows:

$$\sum_{s_1} \left[\arg\max_d u(d) + \sum_{s_2} v(d, \overline{\theta}(s_2, s_1))\pi(s_2|s_1) \right] \pi(s_1|s_1') - \sum_{s_1} \left[\arg\max_d u(d) + \sum_{s_2} v(d, \overline{\theta}(s_2, s_1'))\pi(s_2|s_1) \right] \pi(s_1|s_1') \\ + \sum_{s_1} \left[\arg\max_d u(d) + \sum_{s_2} v(d, \overline{\theta}(s_2, s_1'))\pi(s_2|s_1) \right] \pi(s_1|s_1') - \arg\max_d u(d) + \sum_{s_2} v(d, \overline{\theta}(s_2, s_1'))\pi(s_2|s_1') \tag{43}$$

This decomposition highlights two distinct effects of information. The first is the constraint effect: observing S_1 rather than S_1' places additional restrictions on what the benefactor can reasonably believe ex post. The second is a prediction effect: observing S_1 gives the benefactor a more precise prediction of S_2. The proof proceeds by showing that (a) the constraint effect has the sign predicted by the theorem, and (b) the prediction effect is zero when the benefactor's preferences respect expectation.

(a) It is enough to show the result for any particular realization (s_1, s_1'). Consider therefore

$$\arg\max_d u(d) + \sum_{s_2} v(d, \overline{\theta}(s_2, s_1))\pi(s_2|s_1) - \arg\max_d u(d) + \sum_{s_2} v(d, \overline{\theta}(s_2, s_1'))\pi(s_2|s_1) \tag{44}$$

By the same argument used above to prove part 1 of the proposition this difference is negative (positive) if v is supermodular (submodular). Intuitively, information tends to force the donor to hold a less optimistic view of θ, which increases generosity if and only if d and θ are substitutes.

(b) The prediction effect can be written as

$$\mathbb{E}\left[\arg\max_d u(d) + \mathbb{E}[v(d,\overline{\theta})|S_1]\right] - \arg\max_d u(d) + \mathbb{E}\left[v(d,\overline{\theta})\right] \tag{45}$$

for appropriate priors (which I suppress for brevity). Since preferences respect expectation we know that

$$\mathbb{E}\left[\arg\max_d u(d) + v(d,\overline{\theta})\right] = \arg\max_d u(d) + \mathbb{E}\left[v(d,\overline{\theta})\right] \tag{46}$$

Moreover since this property holds for any prior we can apply it a second time after conditioning on a realization s_1 to show that

$$\mathbb{E}\left[\arg\max_d u(d) + v(d,\overline{\theta})|s_1\right] = \arg\max_d u(d) + \mathbb{E}[v(d,\overline{\theta})|s_1] \tag{47}$$

Taking expectations of both sides over S_1 yields

$$\mathbb{E}\left[\arg\max_d u(d) + v(d,\overline{\theta})\right] = \mathbb{E}\left[\arg\max_d u(d) + \mathbb{E}[v(d,\overline{\theta})|S_1]\right] \tag{48}$$

which together with (46) implies that (45) is zero.

□

Proof of Proposition 5

Proof. **Part 1**. Given d and the realization (s_2, s_1) the benefactor's ex-post problem is

$$\max_{\theta\in\Theta(s_2,s_1)} v(d,\theta) - v(\overline{d},\theta) \tag{49}$$

Since $v_d(d,\theta)$ is monotone in θ, the solution to this problem must also solve $\max_{\theta\in\Theta(s_2,s_1)} v_d(d,\theta)$ for *any* d if $d \geq \overline{d} = \min D$, and $\min_{\theta\in\Theta(s_2,s_1)} v_d(d,\theta)$ for *any* d if $d \leq \overline{d} = \max D$. It follows that further constraining the benefactor's ex-post beliefs by revealing additional information will decrease (increase) the expected value of $v_d(d,\theta)$ for any d, and thus weakly decrease (increase) his expected donation, when $\overline{d} = \min D$ ($\overline{d} = \max D$).

Part 2. The argument proceeds exactly as in the proof of Part 2 of Proposition 3. The effect of coarser information has two effects, a constraint effect and a prediction effect; the prediction effect is zero when preferences respect expectation, while the sign of the constraint effect depends on $\overline{d}$ as in Part 1 above.

□

B Communication

At the heart of the preceding analysis is the idea that other-regarding behavior is qualitatively different from self-regarding behavior because of the lack of directly experienced consequences. Benefactors do not experience the effects they produce for beneficiaries but instead learn about them indirectly. One channel for this indirect learning is of course communication between benefactor and beneficiary. For example, givers and receivers of holiday gifts may talk beforehand about the kinds of things the receiver likes, and often talk afterwards about the suitability or desirability of the gift chosen – the giver hoping to hear the receiver say that it was "just what I wanted."

To better understand good intentions in settings where such direct communication is possible it is necessary to model strategic communication between benefactors and beneficiaries. This section does so in an extended and adapted version of the parable of Don and Ben. Specifically, I enrich Don's choice set so that he decides between alternative methods of helping, and also allow Ben to communicate ex ante with Don.

B.1 An Example, Continued

Don, the Manhattan marketing executive, is again contemplating a donation to help Ben, the African farmer. Don has become aware of two different NGOs both of which work in Ben's village but which provide different services, and must decide how much to donate to each. Let $d = (d^a, d^b)$ represent his giving, where $d^a, d^b \geq 0$ and Don's budget constraint is $d^a + d^b \leq y$. Ben's preferences are represented by

$$v(\theta, d) = \theta^a d^a + \theta^b d^b \tag{50}$$

The interpretation is that θ^i measures the marginal impact of intervention i on Ben's welfare. Don is uncertain about these impacts, knowing only that they are drawn from distribution π with support on $[\underline{\theta}^a, \overline{\theta}^a] \times [\underline{\theta}^b, \overline{\theta}^b]$ where $\underline{\theta}^a > 0$, $\underline{\theta}^b > 0$. Don does want to help in the way he perceives to be most effective; he seeks to maximize

$$u(y - d^a - d^b) + \mathbb{E}_{\hat{\pi}}[\theta^a d^a + \theta^b d^b] \tag{51}$$

Don does not anticipate any feedback on the impact his donations have. Before he gives, however, Ben has an opportunity to send him a costless message m from some arbitrary set M.

Because he does not anticipate any feedback, Don finds it optimal to hold the same beliefs about the effectiveness of each intervention both before and after donating. In particular if he chooses to fund intervention i then he will optimally interpret Ben's message m to mean that

$$\hat{\pi}(\theta^i = x|m) = 1(x = \max\{\theta^i : \mathbb{P}(m|\theta^i) > 0\}) \tag{52}$$

In other words, Don holds the most optimistic view of the intervention he is funding that is

also consistent with Ben's message. Denoting by

$$\overline{\theta}^i(m) = \max\{\theta^i : \mathbb{P}(m|\theta^i) > 0\} \tag{53}$$

the most optimistic view of intervention i given message m, Don thus donates to intervention

$$i^*(m) = \arg\max_{i\in\{a,b\}} \{\overline{\theta}^i(m)\} \tag{54}$$

and gives a total donation $d^*(m)$ characterized by

$$u'(y - d^*(m)) = \overline{\theta}^{i^*(m)}(m) \tag{55}$$

Given this, Ben's problem is to choose a message m solving

$$\max_{m\in M} d^*(m)\theta^{i^*(m)} \tag{56}$$

This expression highlights the fact that Ben's communication decisions must trade off two goals: he wants to steer Don towards the more effective intervention, but also wants to encourage Don to give generously to whichever intervention he chooses.[10] His credibility on these topics, however, is very different. Don knows that Ben has no direct incentive to lie about *which* kind of help he prefers. He does have a direct incentive to mislead Don about the effectiveness of this intervention, since he would always prefer that Don give more, while Don trades off this help against his private benefits of consumption.

Formally, it follows immediately from inspection of (56) that any equilibrium must be action-equivalent to an equilibrium in which Ben chooses at most one message that induces Don to donate to each intervention. The reason is simply that if two messages m, m' both induced intervention a (say) and $d^*(m) < d^*(m')$ then Ben would always prefer to send message m'. Hence we can without loss of generality restrict attention to equilibria in which Ben sends at most two messages with positive probability, m^a inducing a or m^b inducing b. This in turn lets us characterize a unique recipient-optimal equilibrium. To do so define $\overline{\theta}^i = \max\{\theta^i\}$ as the most optimistic view about intervention i given priors π. Then we have

Observation 5. *There exists a unique equilibrium in which Don gives $d^*(\overline{\theta}^a)$ to a if $\theta^a d^*(\overline{\theta}^a) \geq \theta^b d^*(\overline{\theta}^b)$ and gives $d^*(\overline{\theta}^b)$ to b otherwise.*

Proof. By the argument above, in any equilibrium strategy Don either gives $d^*(m^a)$ to a or $d^*(m^b)$ to b. Ben's problem thus amounts to choosing between the payoffs $\theta^a d^*(m^a)$ and $\theta^b d^*(m^b)$. It follows that in any equilibrium Ben sends message m^a if and only if

$$\frac{\theta^a}{\theta^b} \geq \frac{d^*(m^b)}{d^*(m^a)} \tag{57}$$

[10] Provided $\theta^i \geq 0$. Consider this case for now.

Given this, Don's optimal donation level d^a on observing m^a must satisfy

$$u'(y - d^*(m^a)) = \max\left\{\theta^a : \exists \theta^b \text{ such that } \pi(\theta^a, \theta^b) > 0 \text{ and } \frac{\theta^a}{\theta^b} \geq \frac{d^*(m^b)}{d^*(m^a)}\right\} \tag{58}$$

$$= \overline{\theta}^a \tag{59}$$

where the second step follows from the assumption that π has full support on an interval in $\mathbb{R}^2$. Similarly, Don's donation on observing m^b is given by $u'(y - d^*(m^b)) = \overline{\theta}^b$. This uniquely determines $\frac{d^*(m^b)}{d^*(m^a)}$. If this quantity lies within $\left[\frac{\underline{\theta}^a}{\overline{\theta}^b}, \frac{\overline{\theta}^a}{\underline{\theta}^b}\right]$ then it defines a unique interior equilibrium; in this case there is some communication in equilibrium. If on the other hand it is greater than $\frac{\overline{\theta}^a}{\underline{\theta}^b}$ then Ben only sends m^b, while if it is less than $\frac{\underline{\theta}^a}{\overline{\theta}^b}$ then Ben only sends m^a; in these cases nothing is communicated in equilibrium. □

This equilibrium generically features a distortion away from the most effective intervention. To see this, consider the most interesting case in which there is non-trivial communication in equilibrium. In order to maximize effectiveness Ben would like to recommend intervention a if and only if $\theta^a \geq \theta^b$. In equilibrium, however, he gets intervention a when $\theta^a d(\overline{\theta}^a) > \theta^b d(\overline{\theta}^b)$. These conditions coincide only if $\theta^a = \theta^b$; otherwise they diverge, and Ben is either too likely to get one or the other intervention.

The basic issue here is intuitive. For any given amount Don spends, he and Ben would both prefer that he spend it on the most effective intervention. This motivates Ben to inform Don if the intervention he is considering is not in fact the best. Ben also realizes, however, that if Don is excited about the potential of one intervention then disillusioning him may not only affect *how* he helps but also *how much*. He may therefore optimally allow Don to retain a mistakenly optimistic view of some "pet" intervention, preferring a lot of somewhat useful help to a smaller amount of more impactful giving.[11]

The result indicates that the size of this distortion depends on the relative magnitude of $\overline{\theta}^a$ and $\overline{\theta}^b$. If the two interventions allow similar scope for optimism or have similar "upside potential" then distortions will be minimized. For example, there should be little bias in conversations about the best way to achieve some fixed goal. If not then there will be a bias towards the intervention with more upside potential at the expense of the one with the higher expected return; in extreme cases where $\underline{\theta}^a d(\overline{\theta}^a) > \overline{\theta}^b d(\overline{\theta}^b)$ communication breaks down entirely. Note that because bias is driven by upside this implies that donors will tend to be biased towards relatively new, untested interventions whose potential upside is still very high at the expense of older, more tested interventions whose effects are well-known – a bias which gives rise in a natural way to "fads."

[11] While the details differ, the basic tension here parallels that in Che et al. (2013). They study a model in which an agent advises a decision-maker on which of several discrete projects to implement. Given perfect information the decision-maker and agent have identical preferences over these projects, but the decision-maker also places positive value on an "outside option" which is worthless to the agent. This tension introduces distortions in communication, with the better-informed agent sometimes recommending inferior projects in order to prevent the decision-maker from exercising his outside option.