A Theory of Good Intentions*

Paul Niehaus
UC San Diego

August 8, 2013

Abstract

Why is other-regarding behavior so 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, especially when helping remotely (as for example in international development projects). Among other things the model helps explain why donors have a limited interest in learning about effectiveness, why charities market based on need rather than effectiveness, 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.

JEL codes: D64, O1

*I thank Nageeb Ali and Navin Kartik for helpful comments, Adam Szeidl for encouragement and advice, and Microsoft Research New England for their 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 have done any research comparing the effectiveness of alternatives.[2] This pattern is in fact 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?

The predominant interpretation in the literature has been that funders want to be effective, but struggle to learn how because of market failures. Information about effectiveness is a public good (Duflo and Kremer, 2003; Levine, 2006; Ravallion, 2008; Krasteva and Yildirim, 2011), and communication from practitioners to funders is often distorted by strategic considerations (Pritchett, 2002; Duflo and Kremer, 2003; Levine, 2006). Addressing such market failures was one stated purpose for creating many of the institutions that today produce and disseminate effectiveness research – the Center for Global Development, the Jameel Poverty Action Lab, Innovations for Poverty Action, and the Center for Effective Global Action, among others.

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. To underscore how distinct these concepts can be, consider donating to a charity that feeds malnourished African children. This induces agreeable thoughts of children eating nutritious meals. Now suppose you learn that the charity is ineffective – perhaps an exposé reveals that management committed serious fraud. Presumably this reduces your satisfaction. What is more interesting is that, 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*. Put bluntly, your altruistic preferences cannot literally be over childrens' outcomes; these occur on another continent, outside of your experience.

I formalize this idea in a model of a single benefactor whose actions affect a beneficiary.[3] The state of the world is uncertain, so that the benefactor does not know ex ante how his decision will affect the beneficiary ex post. The unusual feature of the model is that this uncertainty persists ex post with positive probability. For example, a donor may never learn whether the charity he gave to is honest. As a result the benefactor faces *ex post ambiguity*: he may observe information that is insufficient to reveal the state and have to interpret it. This is an interesting problem precisely because he has no way of learning the correct interpretation over time, even if the game repeats, since the true state remains unobserved. I therefore examine the case in which the benefactor interprets ambiguity in the way that maximizes his expected utility. I find that he optimally holds empirically correct beliefs about observable quantities, but interprets ambiguity optimistically. For example, a donor correctly forecasts

[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.

[3] The model thus abstracts from public goods issues.

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.

Section 3 applies this framework to an otherwise standard model of "pure" altruism. I find that the benefactor's interest in learning is ambiguous. On the one hand, information's primal role is to constrain his beliefs. A donor who learns ex post that his donation was stolen can no longer believe that it was effective, for example, and is thus worse off than if he had remained ignorant. Yet the donor wishes to avoid such disappointments, and this induces a positive demand for ex ante information. Before deciding how much to give, for example, the donor would like to know how likely it is that an unpleasant scandal will subsequently break. More generally, the benefactor prefers to do enough research ex ante to accurately forecast the feedback he will receive ex post, but no more. This result contrasts with the standard decision-theoretic view in which information always has non-negative instrumental value. Yet it is broadly consistent with the way charitable givers describe their own research. As one survey respondent put it, "I just want to make sure my charities 'hurdle the bar,' I don't care by how much." (Hope Consulting, 2012, p. 40)

Incentives for fundraising intermediaries (e.g. charities) are also mixed. Interestingly, an intermediary's expected revenue falls when it commits to conducting an impact evaluation (formally, disclosing information about a parameter that complements the benefactor's action). The intuition is that the benefactor's and intermediaries' interests are already aligned: the benefactor wants to believe the best about impact, and so further information is more likely to hurt than to help. This result also illustrates a more general theme, namely that the model exhibits a generic tradeoff between the *quality* and the *quantity* of giving. Even the beneficiary may prefer not to disillusion a well-intentioned benefactor, preferring to receive large amounts of help in less-than-optimal ways.

The intermediary does benefit in expectation, however, from disclosing information about need (formally, disclosing information about a parameter that substitutes for the benefactor's action). Information is effective here because of a conflict of interest between the parties: the benefactor wants to believe the beneficiary is doing well, but the intermediary wants him to believe that the need is great. Together these results may help explain nonprofit organizations' tendency to market based on need rather than evidence of effectiveness (and in an extreme example, their use of so-called "poverty pornography" images). They also help rationalize the prevalence of "awareness-raising" campaigns as devices for social change – an otherwise puzzling pattern, since better information need not generically dispose benefactors towards greater engagement.

The model also naturally organizes a set of issues related to the salience of altruistic acts. Because the benefactor derives utility from his thoughts, anything that brings those thoughts to mind tends to raise the return on giving. Among other things, this helps explain why donors often support work on problems that have directly affected loved ones – for example, funding cancer research after losing a relative to cancer; why charities spend money to thank donors repeatedly for past gifts; and why charities encourage donors to think of their gifts as buying discrete, memorable in-kind goods (e.g. cows) even when in reality (and in the legal fine print) they have no influence over how the charity allocates funds.

The paper draws on and extends two strands of research. First, it takes quite literally Andreoni's (1989) idea that altruists benefit from the "warm glow" that their acts induce. While this construct has yielded valuable insights about strategic interaction among donors, it also begs the question why some acts generate more warm glow than others. The present paper offers one potential answer: warm glow is a function of the altruists' *perceptions* of outcomes, which may diverge widely (and predictably) from reality.

Second, the analysis draws inspiration at several points from Brunnermeier and Parker's (2005) theory of optimal expectations. The key technical difference is that, unlike in their model, there is no tradeoff between anticipatory and flow utility; instead the decision-maker's 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 of work that emphasizes the importance of beliefs for overall well-being (e.g. Akerlof and Dickens (1982)). This literature focuses on self-regard; the argument here is that beliefs must be at least as important for understanding other-regard.

2 The Good Intentions Framework

2.1 Timing

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

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 benefactor's decision d and on the state of the world 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 purely private benefits the benefactor obtains as a function of his decision. For example, if $d \in [0, y]$ is a donation to a charitable cause then $u(y-d)$ would be the benefactor's consumption utility. The second term represents the utility the benefactor obtains

from the beneficiary's outcome; note that this specification thus implies that the benefactor must be *aware* of the ex-post realization of v. To allow for the possibility of ex-post ambiguity, I study the case in which the benefactor's payoff depends 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 the parameter $\hat{\pi} \in \Delta(\Theta)$ which is the benefactor's ex-post subjective belief about the state of the world.[4] This specification thus embodies the idea that the benefactor's 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)). Finally, a benefactor might care not about the beneficiary's outcome per se but about this outcome relative to some reference point; for example, Duncan (2004) examines a donor who cares about the "impact" of his actions relative to the outcome if he had not helped. I abstract from these frictions here to focus on issues of perception.

2.3 Optimization

Given beliefs, the benefactor's decision-making process is entirely standard: he chooses a decision d to maximize his subjective expected utility at the moment he decides. 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 modelling the evolution of the benefactor's beliefs and their effects on his behavior through (3). I begin by placing mild restrictions on the subjective beliefs $\hat{\pi}$ he may hold.

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

[4] Uncertainty about u can be incorporated with additional notation but without further insight.

problem does not go away with learning (Kalai and Lehrer, 1993). Because the benefactor does not observe θ even ex post, he cannot learn anything new about $\pi(\theta|s_2, s_1)$ no matter how many times he observes i.i.d. draws of (s_2, s_1).

To resolve this indeterminancy, I study beliefs that are optimal in the sense that they maximize the benefactor's 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. Before stating results, I characterize and interpret the differences here.

First, the benefactor holds preferences over beliefs as well as over outcomes. This modeling approach 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. While this literature has focused on self-regarding beliefs, thoughts or beliefs are surely at least as important for understanding other-regard. When giving to help children in Africa, for example, it is hard to see how anything *other* than beliefs could matter.

Second, the model explicitly endogenizes beliefs. This contrasts with usual practice, which is to exogenously specify beliefs that match the empirical distributions of unknown quantities (i.e. $\pi = \hat{\pi}$). Such an approach seems hard to justify in this setting given that the benefactor does not observe θ, even if the game is arbitrarily repeated. Yet might the argument not at least apply to the benefactor's beliefs about observables (s_2, s_1)?

As it turns out, optimal beliefs *endogenously* co-incide with empirical distributions wherever the latter are observable. One can show this by further characterizing optimal beliefs. First, note that the benefactor's ex post belief $\hat{\pi}(\theta|d, s_2, s_1)$ affects his payoffs only through $\mathbb{E}_{\hat{\pi}(\theta|d,s_2,s_1)}[\theta]$. 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 chooses ex ante

beliefs that lead him to act in the way that is optimal, given that he will ultimately hold the optimistic view $\overline{\theta}$ of the beneficiary's situation. Having established this point we can now prove that optimal beliefs are, without any loss of generality, consistent with Bayes rule.

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)*.*

Proof. See Appendix A for all proofs. □

The proof 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 his *interpretation* of this feedback in order to view it as indicative of an appealing state of the world $\overline{\theta}$. This has five noteworthy implications.

First, optimal beliefs have all the mathematical properties we typically expect of beliefs: for example, they behave as martingales. As an empirical corollary, a researcher cannot distinguish between the beliefs of a "well intentioned" agent and those of a standard agent without using ancillary data such as observed behaviors, the empirical distribution π, etc.

Second, as claimed above optimal beliefs are consistent with the observable data precisely for those variables for which the usual empirical approach has bite. Formally, the distribution over (s_2, s_1) implied by (7) is simply 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.

Third, optimal beliefs differ from the actual distribution only in the way they describe the parts of the model that are *unobservable*, and thus pin down the interpretation of ambiguous data where the standard techniques must simply assume that agents know the truth about quantities they never observe. Formally, the only differences between the benefactor's beliefs and the empirical distribution lie in the conditional distribution of θ given (s_2, s_1).

It is worth noting that assuming optimistic interpretations of *ambiguous* information is conservative in relation to recent theoretical and empirical work. Recent theoretical advances model beliefs as inconsistent even with the distribution of observables; 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). Similarly, recent experimental work has documented optimistic (mis)interpretation even in cases with little or no ambiguity. Formalizing suggestive earlier work from social psychology, Mobius et al. (2012) document self-serving updating biases in a rigorously controlled and incentivized laboratory experiment. They track the evolution of

subjects' beliefs in response to noisy feedback from a known data generating process and show that subjects skew their interpretation of feedback, responding more to positive than to negative information. In comparison to these precisely defined signals, feedback in the settings considered below is typically far more ambiguous. There are no objective measures, for example, of the likelihood that a nonprofit executive is corrupt conditional on the absence of scandal. Such settings only provide greater scope for the imagination.

Fourth, optimal beliefs are self-consistent (and uniquely so within the class of beliefs that are empirically consistent with observables). To see this note that (7) continues to hold if we redefine π as the right-hand side. This implies that a benefactor holding these beliefs has no desire to deviate from them, and hence that they are optimal in the usual sense. This property need not hold for the empirical distribution π. It also distinguishes the model from theories of optimal self-regarding beliefs, in which the tension between between utility from actions and from beliefs typically leads agents to hold self-inconsistent beliefs. (c.f. Brunnermeier and Parker (2005)) Here there is no such tension as the benefactor cares exclusively about optimizing his beliefs.

Fifth and finally, the good intentions framework nests the benchmark model of preferences over outcomes. To see this, consider what happens when a piece of evidence (s_2, s_1) is consistent with only a single state $\theta : \pi(\theta|s_2, s_1) > 0$. In this case the result says that the unbiased interpretation $\hat{\pi}(\theta|s_2, s_1) = \pi(\theta|s_2, s_1)$ is part of an optimal belief structure. To elaborate further, say that feedback is *fully revealing* if $\{\theta \in \Theta : \pi(\theta|s_2, s_1) > 0\}$ 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 Loops)**.** *Beliefs derived via Bayesian updating from the prior $\pi(\theta, s_2, s_1)$ are optimal if feedback is fully revealing.*

Put another way, the standard model and the good intentions framework coincide if the benefactor expects no ex-post ambiguity about θ.[5] To see the intuition, consider the case in which the benefactor makes a decision affecting himself only. Because he directly experiences the consequences, on can think of this as a case in which he necessarily "learns" the true value of θ. If we formalize the analogy by letting $s_2 = \theta$ in the current model we recapture this case, and the Lemma tells us that Bayesian updating is optimal. Lemma 2 thus provides a convenient device for contextualizing the model's predictions.

3 Effective Giving

This section applies the good intentions framework to the question of learning: how much will the benefactor know in equilibrium when choosing how to help? For concreteness I develop the main ideas using simple functional forms; Section 3.5 generalizes the main ideas to more general functional forms. I motivate the analysis using charitable giving as a leading case, but note that the issues at stake are broader. Unwanted Christmas gifts, for example, are so common that there are websites devoted to displaying bad examples: knick-knacks, ugly sweaters, and so on.[6] Waldfogel (2009) argues that holiday gift-giving is so wasteful that

[5] The antecedent can be made both necessary and sufficient by adding appropriate sensitivity conditions.
[6] See for example `www.badgiftemporium.com` or `whydidyoubuymethat.com`.

people should stop it entirely. On a grander scale, critics of foreign aid argue that much of it is ineffective or even counterproductive. Bill Easterly puts it bluntly: "poor people die not only because of the world's indifference to their poverty, but also because of ineffective efforts by those who do care" (Easterly, 2006).

3.1 An Example

Consider the following parable: Don, a mid-career marketing executive living in Manhattan, contemplates a donation to an NGO working to improve the lot of Ben, a subsistence farmer in Africa. Don can donate any amount d up to his total income y. Ben's quality of life depends both on Don's donation and on factors such as the level of rainfall and the effectiveness of the NGO's work; for simplicity, say that 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 any donation d. Don's prior assessment is that $\pi(\theta = \theta^g) \equiv \gamma \in (0, 1)$.

Don is well-intentioned in the sense that he genuinely prefers to see Ben better-off. He has no paternalistic motives and does not seek to impress anyone with his generosity. Because Ben is thousands of miles away, however, his level of satisfaction depends on what he *thinks* is happening in Africa, which may differ from what is actually happening. Formally, Don's preferences are represented by

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

where $\hat{\gamma}_2$ is Don's subjective assessment, after donating, of the likelihood that the situation in Africa is good. This assessment is shaped by what he learns both before and after choosing how much to donate. Suppose for simplicity that 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 the growing conditions in Africa, he might or might not read a story in the news about the state of African agriculture. Let p be the probability that he learns the truth before donating, and q the probability that he learns it after donating *conditional on not having learned it before*.

If Don learns the state of affairs before donating then this pins down his beliefs and simplifies his decision to

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

with solution $d^*(\theta)$. In the more interesting case where he does not learn before donating, he must consider how he will likely view the situation in the future. With probability q he will learn the true state of affairs. With probability $1 - q$, on the other hand, he will obtain ambiguous information which he can interpret as meaning that all is well ($\theta = \theta^g$). For example, if he hears no scandals surrounding the NGO he has donated to he will interpret this as meaning that the NGO is well-managed. His future opinion 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] 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. Given these beliefs, Don's charitable donation solves

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

with solution $d^*(\emptyset)$. Intuitively, Don interprets no news as good news and makes his charitable giving plans accordingly.

3.2 Learning to Help

Don's tendency to take an optimistic view of Ben's situation shapes his motives for learning in unusual ways. Consider first how Don's expected payoff changes when he learns the truth about Ben's situation 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 reflects information's role in the model as a *constraint*. New information rules out hypotheses that formerly were plausible, and thus limits the views Don can take. Yet somewhat paradoxically, by constraining beliefs information also creates its own endogenous value. To see this, suppose momentarily that Don knew with certainty that he would learn the truth ex post, and consider his demand for 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. In fact, when Don expects to confront the truth eventually then his decision problem has the same structure as a standard, single-agent problem. But suppose alternatively 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

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

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

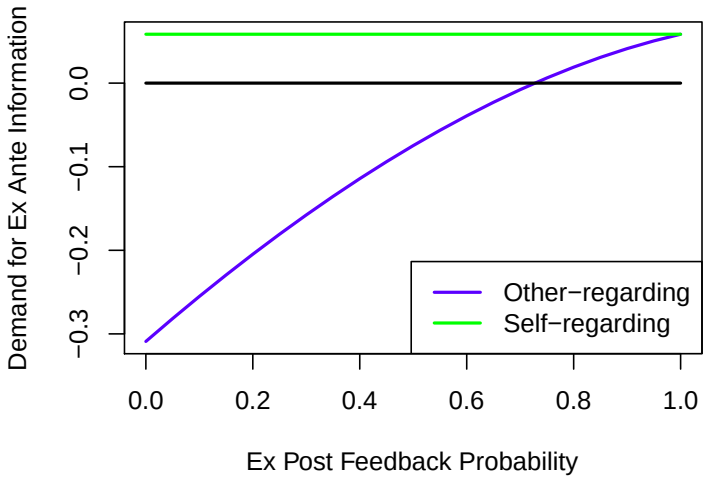
$$\arg\max_d y - d + \mathbb{E}_{\hat{\gamma}_1}[v(\theta, d)] = \arg\max_d y - d + \mathbb{E}_{(1-q(1-\gamma))}[v(\theta, d)] \tag{10}$$

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: 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.

Figure 1 illustrates the basic tension between the costs and benefits of learning for 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 faced the same decision but as a matter of self- instead of other-regard.

Observations 1 and 2 show that well-intentioned givers will typically do little research before giving. How well does this explain real-world donor ignorance? One alternative explanation is that donors do in fact want to learn but that the *supply* of information about effectiveness is limited. GiveWell, one of the few organizations that does provide in-depth analysis of

nonprofit performance, makes this point part of their raison d'etre: "The issues charities address – from fighting disease in Africa to improving education in the U.S. – are extremely complex, and useful information about what different charities do and whether it works isn't publicly available anywhere."[8] Yet as always it is difficult to tell whether the absence of a good reflects lack of supply or demand.

Existing work on this issue emphasizes the supply side, positing that donors want to learn how to be effective but face incentive problems. Information about effectiveness is a public good (Duflo and Kremer, 2003; Levine, 2006; Ravallion, 2008; Krasteva and Yildirim, 2011), and communication from practitioners to givers is often strategic and intended to persuade (Pritchett, 2002; Duflo and Kremer, 2003; Levine, 2006). The desire to redress these market failures was a key part of the rationale for creating the ecosystem of institutions that today produce and disseminate research on effectiveness: the Center for Global Development, the Jameel Poverty Action Lab, Innovations for Poverty Action, and the Center for Effective Global Action, among others. In contrast, the "good intentions" framework describes donors who fundamentally do *not* want to learn about effectiveness and will prefer to ignore new sources of information. It is more closely aligned with Lant Pritchett's observation that funders "have too little doubt" (quoted in Dugger (2004)); the model provides an explanation rooted in the idea that having "too little doubt" may be *optimal* for maximizing expected utility.

While the supply side undoubtedly matters, there are at least three reasons to believe that a purely supply-side story cannot fully explain the data. First, small private donors are not the only actors to exhibit surprisingly limited interest in effectiveness research. In "Building Learning into the Global Aid Industry," for example, David Levine argues that "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). Second, organizations that compete in markets serving smaller donors could presumably gain a competitive edge over their rivals by providing evidence on effectiveness if this were something donors sought. Third, donors themselves describe their research behavior in terms that closely mirror Don's ambivalence. Several of Hope's respondents explicitly stated that there is a limit to what they wanted to know:

> "With known nonprofits, unless there is a scandal, you assume they are doing well with your money." (p. 38)
>
> "I just want to ensure that I'm not throwing my money away." (p. 40)
>
> "I just want to make sure my charities 'hurdle the bar,' I don't care by how much." (p. 40)
>
> "I don't research, but I am sure that the nonprofits to which I donate are doing a great job." (p. 42)

These strategies closely parallel Don's approach – both his desire to learn enough to avoid disappointment and his aversion to learning any more. As the second respondent puts it, there is no need to delve into the details when one can safely "assume" a good outcome. This

[8] `http://www.givewell.org/about/story`, accessed 9 February 2013.

attitude – naive from the perspective of standard theory – is exactly what good intentions predicts. As Hope concludes, "this creates a big challenge to getting people to do more research – they see no need to do so." (p. 44)

Economists have argued that similar attitudes prevail even among development professionals. David Levine laments that "practitioners are almost always convinced their programs are both useful and cost-effective" (Levine, 2006). Bill Easterly puts it colorfully: "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*" (Easterly (2006), emphasis added). It is also striking that the recent increase in the use of randomized impact evaluation has been driven first by academics rather than by funders, a point Chris Udry makes in commemorating Esther Duflo's John Bates Clark award. "Over J-PALs almost decade of existence, and I believe in large measure due to the example of its members and its institutional support, the tools of randomized evaluations have become much more common and prominent in development economics" (Udry, 2011). Moroever, the demand for RCT outputs has been mixed: for example, Brigham et al. (2013) 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.

If as Hope Consulting argues donors simply see little "need" to do research, it is intriguing that the good intentions framework highlights ex post feedback as the driver of this need. By forcing donors to confront reality, it can stimulate demand for ex-ante information about what works. This idea is absent in standard models in which payoffs depend only on material consequences, not on beliefs, so that ex post feedback is irrelevant. Yet some development economists argue that it is a practical necessity. 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 lead to improvement, it will serve to focus the energies of the education system on the outcome that actually matters to millions of first-generation learners, which is functional literacy and numeracy (that the system is currently not delivering)."

One of the best-known examples of feedback mechanisms in play may be the World Bank's "Doing Business" reports, which document simple – but often striking – facts about the business climate in developing countries. For example, the original report found that creating a new business took at least 47 business days and 47% of per capita income on average across 85 countries (Djankov et al., 2002). Interpreted through the lens of rational expectations it is not obvious how these data would affect policy-making, as they would be equally likely to be a positive or a negative surprise. Interpreted through the lens of good intentions, the fact that such data will be collected regularly is powerful simply because it forces policy-makers to confront an inconvenient truth they would otherwise have willingly ignored.

3.3 Intermediaries

Given Don's ambivalence about research, a natural next question is whether other market players have an incentive to proactively inform him. For example, nonprofit organizations consciously design marketing strategies in order to attract charitable donations. What marketing strategies maximize these donations, and in particular what information does a fundraising intermediary benefit from providing?

To address this question I next characterize how information affects Don's expected *action*, as opposed to his payoff. For a fundraising intermediary, this is the comparative static that determines the returns to undertaking research. For simplicity I focus here on the case in which Don and the intermediary are symmetrically informed and the researcher can commit to disclosing her results; this corresponds, for example, to commissioning an academic study by J-PAL with the expectation that the results, good or bad, will be made public.

The answer, as it turns out, is that research can either increase or decrease Don's expected generosity depending on the nature of the research. This principle applies to both ex ante and ex post research, but it will be helpful for technical reasons to state the results separately:

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

To see this, note that the probability of feedback after a donation matters for Don's decision-making only in the case where he is uninformed when donating, 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 $v_d(\theta^b, d) - v_d(\theta^g, d)$. One can also prove an analogous statement about ex ante information, with the one complication that one must first purge the model of standard neoclassical effects of information:

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 intuition for both of these results is straightforward: Don prefers to believe that things are going well for Ben, and so information generally forces him to revise those beliefs downward. What this implies for his donation d then depends on whether giving is more or less impactful on the margin when the situation θ is bad. Suppose the state θ complements donations – for example, let θ measure something about the effectiveness of giving. Then forcing Don to confront the truth about effectiveness will lower his donation. In these cases an intermediary has no incentive to inform him. Suppose on the other hand that the state θ substitutes for donations d – for example, let θ measure Ben's baseline level of income. Then forcing Don to confront the truth about Ben's poverty raises his perception of the *marginal* impact of his donation, and hence motivates him to give more. An intermediary has strong incentives to force Don to absorb this sort of information.

These results cast nonprofit marketing practices in an interesting light. Consider first that information about the effectiveness of giving is information about a parameter that complements donations, since higher effectiveness is associated with a higher marginal return to giving. The model predicts that nonprofits will communicate little to donors about their effectiveness for the simple reason that they do not need to: donors and nonprofits both prefer for the donor to believe that effectiveness is high. This directly increases the donor's payoff and also indirectly benefits the nonprofit by increasing donations.

Now consider information about the need of the beneficiary. is information about a parameter (well-being) that typically substitutes for donations. Intuitively, the marginal return on donations is highest when the beneficiary is worst-off. As a result donors and non-profits have conflicting preferences: the donor prefers to believe that the beneficiary is doing well, but the nonprofit wishes to force the donor to confront the reality of poverty. This may help explain why reminders of need are ubiquitous in nonprofit marketing at the same time as basic information about what exactly nonprofits *do* with funds (let along how cost-effective they are) is scarce.

This result may help explain the prevalence of "awareness-raising" campaigns as a strategy for achieving altruistic ends. In a standard model, the impacts of heightened awareness are ambiguous: new information may reveal that needs are either greater or less than previously thought. In the good intentions framework, on the other hand, altruists have a generic bias towards believing that others are doing *better* than they really are. This creates scope for sophisticated social entrepreneurs to mobilize them by forcing them to confront unpleasant realities that they would otherwise have ignored.

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 the point, consider a trivial extension of the model in which Don thinks about Ben in the period after donating with probability $\rho \in (0, 1)$. 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). For example, a donor who has lost a loved one to cancer is more likely to support anti-cancer research since, through the associate property of memory, he is more likely to be reminded of this gift.

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. Indeed, frequent thank-you notes are a mainstay of nonprofit marketing practice.[9] Less obviously, charities can increase the memorability of a gift by associating it with something memorable. The use of "gift catalogues" is a leading example: large nonprofits often allow donors to chose specific items (e.g. a goat) to give to recipients. From a pure decision-making point of view this approach makes little sense, for two reasons. First, donors are unlikely to have good information about which interventions will best serve particular recipients. Second, donors' "choices" are in fact not legally binding, and the accompanying fine print typically makes clear that the nonprofit will do whatever it wants with the donation. From a salience point of view, however, gift catalogues and in-kind transfers generally make perfect sense, as goats are far more memorable than dollar figures.

3.5 General Results

I now state and prove general versions of the observations made above about Don and Ben. Doing so requires some additional terminology that lets us compare the information content of signals. First, I define a sense in which two abstract signals are "the same," i.e. convey the same information:

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

Second, I adopt the standard Blackwell concept to rank which (if either) of two signals is more informative about some third random variable:

[9] Note that Don's taste for reminders is ambiguous due to the absence of any absolute unit with which to measure v: 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.

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.*

Finally, I will use the shorthand $X \succsim Y$ to indicate that the benefactor's expected payoff is weakly greater when he observers the random variable X than when he observers Y. We can now state the following proposition about the value of ex post feedback:

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

This generalizes Observation 2 and states that the benefactor generally prefers as little ex post feedback as possible. As above, the intuition is that feedback constrains his beliefs without providing any decision-relevant information.

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.

To generalize Observation 4 we first need a general statement 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 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 benfactor 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. For example, the fact that information acts as a constraint on beliefs suggests that the benefactor may begin life with an optimistic outlook but become "jaded" over time.

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.

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.

Djankov, Simeon, Rafael La Porta, Florencio Lopez-De-Silanes, and Andrei Shleifer, "The Regulation Of Entry," *The Quarterly Journal of Economics*, February 2002, *117* (1), 1–37.

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

Dugger, Celia, "World Bank Challenged: Are the Poor Really Helped?," *New York Times*, July 2004.

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.

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.

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," Technical Report 2011.

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," Policy Research Working Paper Series 4547, The World Bank March 2008.

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

Udry, Christopher, "Esther Duflo: 2010 John Bates Clark Medalist," *Journal of Economic Perspectives*, September 2011, *25* (3), 197–216.

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:

$$\hat{\pi}(\theta, s_2, s_1) = 1(\theta = \overline{\theta}(d^*(s_1), s_2, s_1))\pi(s_2, s_1) \tag{20}$$

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

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

where

$$d^*(s_1) = \arg\max_d \mathbb{E}_{\pi(s_2|s_1)}[u(d) + \mathbb{E}_{\hat{\pi}(\theta|d,s_2,s_1)}[v(d,\theta)]] \tag{23}$$

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

$$\max_{\tilde{\pi}\in\Theta(s_2,s_1)} \mathbb{E}_{\tilde{\pi}}[v(d,\theta)] \tag{24}$$

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.

$$\arg\max_d \left[u(d) + \mathbb{E}_{\hat{\pi}(s_2|s_1)}[v(d,\theta)]\right] = \arg\max_d \left[u(d) + \mathbb{E}_{\pi(\theta,s_2|s_1)}\mathbb{E}_{\hat{\pi}(\theta|d,s_2,s_1)}[v(d,\theta)]\right] \tag{25}$$

This condition holds if

$$\hat{\pi}(\theta|s_1) = \mathbb{E}_{\pi(s_2|s_1)}[\hat{\pi}(\theta|d, s_2, s_1)] \tag{26}$$

$$= \mathbb{E}_{\pi(s_2|s_1)}[1(\theta = \overline{\theta}(d, s_2, s_1))] \tag{27}$$

$$= \sum_{s_2} 1(\theta = \overline{\theta}(d, s_2, s_1))\pi(s_2|s_1) \tag{28}$$

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

$$u(d^*) + \sum_{s_2} \left[\max_{\theta \in \Theta(s_2, s_1)} \{v(d^*, \theta)\} \right] \pi(s_2|s_1) \tag{29}$$

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

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

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 (30) is at least as great as that in (29) for any particular (s_2', s_2) and hence (30) is also greater in expectation. Since (30) 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

$$U(s_2, s_1) \equiv \max_{d, \theta \in \Theta(s_2, s_1)} u(d) + v(d, \theta) \tag{31}$$

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

$$\begin{aligned} \Theta(s_2, s_1) &= \{\theta \in \Theta : \pi(\theta|s_2, s_1) > 0\} && (32) \\ &\subseteq \{\theta \in \Theta : \pi(\theta|s_2) > 0\} && (33) \\ &= \Theta(s_2) && (34) \end{aligned}$$

and hence the constraint in (31) 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

$$u(d^*) + \sum_{s_2} v(d^*, \overline{\theta}(d^*, s_2, s_1)\pi(s_2|s_1) \tag{35}$$

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

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

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 (36) as

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

which must by definition be weakly less than (35) 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

$$f(d, \{x(s_2', s_2, s_1)\}) \equiv u(d) + \sum_{s_2} \sum_{s_2'} v(d, x(s_2', s_2, s_1))\pi(s_2'|s_2)\pi(s_2|s_1) \tag{38}$$

where

$$x(s_2', s_2, s_1) = \max\{\theta : \pi(\theta, s_2, s_1) > 0\} \tag{39}$$

in the case where he observes S_2 and

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

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

$$\sum_{s_1} \left[\arg\max_d u(d) + \sum_{s_2} v(d, \overline{\theta}(s_2, s_1)) \pi(s_2|s_1) \right] \pi(s_1|s_1') \\ - \arg\max_d u(d) + \sum_{s_2} v(d, \overline{\theta}(s_2, s_1')) \pi(s_2|s_1') \tag{41}$$

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:

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

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

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

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

$$\mathbb{E}\left[\arg\max_d u(d) + \mathbb{E}[v(d,\overline{\theta})|S_1]\right] - \arg\max_d u(d) + \mathbb{E}\left[v(d,\overline{\theta})\right] \tag{44}$$

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

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

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

$$\mathbb{E}\left[\arg\max_d u(d) + v(d,\overline{\theta})|s_1\right] = \arg\max_d u(d) + \mathbb{E}[v(d,\overline{\theta})|s_1] \tag{46}$$

Taking expectations of both sides over S_1 yields

$$\mathbb{E}\left[\arg\max_d u(d) + v(d,\overline{\theta})\right] = \mathbb{E}\left[\arg\max_d u(d) + \mathbb{E}[v(d,\overline{\theta})|S_1]\right] \tag{47}$$

which together with (45) implies that (44) is zero.

□

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

$$v(\theta, d) = \theta^a d^a + \theta^b d^b \tag{48}$$

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

$$u(y - d^a - d^b) + \mathbb{E}_{\hat{\pi}}[\theta^a d^a + \theta^b d^b] \tag{49}$$

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

$$\hat{\pi}(\theta^i = x|m) = 1(x = \max\{\theta^i : \mathbb{P}(m|\theta^i) > 0\}) \tag{50}$$

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

$$\overline{\theta}^i(m) = \max\{\theta^i : \mathbb{P}(m|\theta^i) > 0\} \tag{51}$$

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

$$i^*(m) = \arg\max_{i\in\{a,b\}} \{\overline{\theta}^i(m)\} \tag{52}$$

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

$$u'(y - d^*(m)) = \overline{\theta}^{i^*(m)}(m) \tag{53}$$

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

$$\max_{m\in M} d^*(m)\theta^{i^*(m)} \tag{54}$$

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 (54) 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

$$\frac{\theta^a}{\theta^b} \geq \frac{d^*(m^b)}{d^*(m^a)} \tag{55}$$

[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

$$u'(y - d^*(m^a)) = \max\left\{\theta^a : \exists \theta^b \text{ such that } \pi(\theta^a, \theta^b) > 0 \text{ and } \frac{\theta^a}{\theta^b} \geq \frac{d^*(m^b)}{d^*(m^a)}\right\} \tag{56}$$

$$= \overline{\theta}^a \tag{57}$$

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.