A Theory of Good Intentions*

Paul Niehaus
UC San Diego & NBER

June 21, 2014

Abstract

Altruists often seem to produce results that fall short of their intentions. I examine this tension in a model where altruists derive warm glow from their *perceptions*, as opposed to actual outcomes. Perceptions and reality can diverge when feedback is incomplete: a donor who supports an international development project, for example, may learn little about its results. The base model predicts that altruists in such settings prefer not to learn everything they could before acting, and that market intermediaries promote their services based on need rather than effectiveness. Policy-makers and beneficiaries, meanwhile, face a tradeoff between the quantity and quality of altruistic behavior. The framework can readily accommodate alternative motives such as impact and guilt.

*I thank Nageeb Ali, Jim Andreoni, Navin Kartik, Joel Sobel, Adam Szeidl, Lise Vesterlund, and seminar participants at Microsoft Research New England, Columbia, UCLA, NEUDC, and the Duke Mini-conference on Charitable Giving for helpful comments. Microsoft Research New England provided generous hospitality.

1 Introduction

Altruists often seem to produce results that fall short of their intentions. This perception is so widespread that the term "well-intentioned" has become a euphemism for "poorly informed." Consider charitable giving, for example: Americans give about 2% of GDP to charity each year, suggesting they care deeply about others, yet only 3% of donors even *claim* to have done any research comparing the effectiveness of alternatives.[1] Such figures beg the question: if people really are well-intentioned, why don't they *become* well-informed?

Economists have generally taken the view that they want to, but find it costly or difficult. Krasteva and Yildirim (2013) emphasize that the costs of learning may exceed the benefits in the context of small charitable donations. In international development, market failures are seen as a major culprit: 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 such as J-PAL, IPA, and CEGA were created in part to address such concerns.

I examine an alternative (and complementary) interpretation: altruists do not want to achieve a better outcome. Instead, they want to *believe* they have done so. This creates tension in settings where perception and reality can easily diverge. Consider, for example, making a donation to help feed malnourished African children. Thinking about those children eating generates a "warm glow" (Andreoni, 1989). But now suppose you learn that your donation was wasted or stolen. Presumably this dampens the glow. Yet if you had *not* learned of the waste, you would have continued to experience warm glow thinking about your impact *even though in reality no such impact existed*. This suggests that the preferences that motivated your gift cannot literally be over children's outcomes, which occur far outside of your experience. Instead, perceptions count.

This paper studies how learning works in a market where perceptions are the product. It focuses on a single benefactor and beneficiary, thus abstracting from issues of public goods. 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 effective and have to decide what this implies. He cannot learn the correct interpretation through repeated experience, precisely because the true state remains unobserved. I therefore consider the case where he interprets the evidence in the way maximizes his expected utility. This approach builds on evidence from psychology and economics that people tend to interpret information in a self-serving manner.[2]

The beliefs that result have an innocuous structure: they are (endogenously) Bayesian and consistent with the distribution of observable data, and hence not falsifiable without ancillary data. For example, a well-intentioned donor correctly forecasts the probability that he will

[1] Giving statistics: 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). Research statistics: 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.

[2] See for example Eil and Rao (2011) and Mobius et al. (2013) and the references therein.

hear bad news about the cause he supported. On learning nothing, however, the donor assumes that "no news is good news" and views the cause as definitely effective. Because this effect appears only in the presence of ambiguity, the model predicts relatively normal behavior when outcomes are easily observable (such as helping a neighbor) but relatively distorted behavior when outcomes are unobserved (such as helping internationally).

The well-intentioned benefactor has a limited desire to learn. He always prefers to avoid ex-post feedback as this constrains his beliefs. A donor who learns that his donation was stolen, for example, finds it harder to believe that it was effective. He does want to obtain a limited amount of information ex ante, however, 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 enough research ex ante to accurately forecast the feedback he will receive ex post, but no more. Limited feedback is thus directly linked to disinterest in learning.

These motives then shape the marketing strategies of revenue-maximize intermediaries such as charities. Pundits have raised the concern that "useful information about what different charities do and whether it works isn't publicly available anywhere."[3] This can be good marketing in the model, however. Expected revenue falls when an intermediary commits to generating information on effectiveness – by commissioning an impact evaluation, for example. The reason is that the benefactor's interests are already aligned with those of the intermediary: he *wants* to believe the best about impact, and so further information is more likely to disappoint than excite him. Conversely, the intermediary benefits from marketing based on need. Need-based marketing works because of a conflict of interest: 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 marketing strategies that emphasize "awareness-raising" and graphic depictions of need ("poverty pornography") as opposed to evidence of cost-effectiveness.

More generally the model implies a tradeoff between the *quality* and the *quantity* of giving. From the point of view a policy-maker, wishful thinking is problematic as it 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. Sponsoring rigorous evaluation could address this; if results do not live up to the hype, funders will turn to alternatives. But funders will be less excited about these alternatives than they were about their first approach. Total giving will thus tend to fall. For the policy-maker (and the beneficiary himself) the smaller flow of better-informed giving may or may not be a better outcome than the larger flow of poorly-informed giving.

Because it construes altruism as a preference over perceptions, the model accommodates salience effects relatively easily. Anything that brings thoughts of the beneficiary to mind tends to raise the return on altruistic acts. This is consistent with the fact that donors often support work on problems that have affected their loved ones (Small and Simonsohn, 2008). It may also help explain the fact that charities spend money to thank donors for earlier gifts, and encourage donors to think of those gifts as buying discrete, memorable items (e.g. cows, sponsoring a specific child) even when they are in fact fungible.

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

After applying the framework to the benchmark case of "pure" altruism, I illustrate how it can accommodate some more nuanced motives discussed in recent work. Duncan (2004) argues, for example, that some donors care not about outcomes per se but about the *impact* of their actions. More recently, Andreoni et al. (2012) present evidence of the role played by *guilt*. I examine a class of preferences that nest these motives, depending on the reference point against which the benefactor evaluates outcomes. This turns out to have no new qualitative implications for the benefactor, who continues to avoid feedback ex post and do limited research ex ante. Its major implications are rather for intermediaries. Impact philanthropists are ideal customers, as they want to believe that their actions have a large marginal impact. Guilty givers are the opposite extreme: they tend to assuage their guilt by convincing themselves that needs are exaggerated and that nothing they could do would make a difference. Intermediaries such as nonprofits thus do best to leave impact philanthropists uninformed, but must convince guilty givers of both need *and* effectiveness.

The model's formalisms could potentially be applied to any sort of other-regarding preference.[4] Empirically the link seems tightest to individual charitable giving where, as mentioned earlier, donors overwhelmingly give without researching alternatives. Respondents told interviewers 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). Citing these data, the Hewlett Foundation recently ended an 8-year, $12M initiative to promote evidence-based giving, saying that "the initiative assumed that donors would use this information if they could find it... [but] most donors aren't even looking."[5] Dictators in laboratory games appear to behave similarly: Dana et al. (2007) find, for example, that only 56% of dictators choose to observe *free* information on the relationship between their actions and the recipient's payoffs.[6] Of course, donors' motives for giving (or not) vary widely and no one model is likely to capture them all. This model best describes donors who give proactively. Other donors give only when asked and may be driven more by social pressure than by a desire to believe they are "making a difference."[7]

Some observers see good intentions as an issue in the institutional aid industry. Easterly (2006) emphasizes the role played by 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 that affect the creation and use of information.[8] But it is consistent with the idea that there is something

[4]Or even some self-regarding preferences. Consider for example acting to improve your social image: unless you have unusually frank peers, there is a good chance you will never really know what they think.

[5]Video interview with Lucy Bernholz, `http://www.hewlett.org/programs/effective-philanthropy-group`, accessed 18 May 2014.

[6]See also Fong and Oberholzer-Gee (2011) who find low willingness to pay for information about recipient identify, and Grossman and van der Weele (2013) who find willingness to pay for ignorance.

[7]The model 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 `www.badgiftemporium.com` or `whydidyoubuymethat.com`). Waldfogel (2009) argues that holiday gifts are so wasteful that in many cases it would be better not to buy them.

[8]Industry veterans often lament the historically limited role of evidence in aid decision-making. Pritchett (2002) describes the process as "ignorant armies clashing by night," with "very rarely any firm evidence presented

fundamentally different about spending money on others' behalf. It may help explain why a new approach such as micro-lending can capture the imagination of practitioners and grow into a large industry before any rigorous evidence of its impact is available (Duflo et al., 2013). The model predicts that this is most likely precisely when little research exists to check the imagination.[9]

Conceptually the paper draws on and extends three strands of theoretical research. First, it takes very 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)). This literature has focused on self-regard; I argue that its tenets are at least as relevant for understanding other-regard.

Third, it provides an alternative view of persuasive activity. Economists have shown how persuasion is possible when a sender can exploit information asymmetries (e.g. Crawford and Sobel (1982)) or non-linearities in the receiver's mapping from beliefs to actions (e.g. Kamenica and Gentzkow (2011)). In this model persuasive marketing is possible without either of these mechanisms. Here, persuasion works by exploiting the receiver's (predictably) wishful thinking. Wishful beliefs that serve the sender's interests are left alone, while those that harm his interests are confronted with data.

The rest of the paper is organized as follows. Section 2 presents the framework and characterizes optimal interpretations. Section 3 expresses the main ideas of the paper in the context of a simple example, which Section 4 then generalizes. Section 5 extends the analysis to alternative motives for giving, and Section 6 concludes with a discussion of open questions.

2 The Good Intentions Framework

2.1 Timing

There are two players, a benefactor and a beneficiary. The timing of play is as follows:

1. Nature determines the value of a finite-valued parameter $\theta \in \Theta$
2. A signal $s_1 \in S_1$ is revealed and the benefactor forms subjective ex ante beliefs $\hat{\pi}(\theta, s_2|s_1)$

and considered about the likely impact of... proposed actions."

[9] On this note see Brigham et al. (2013) who find, intriguingly, 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.

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

Let $\pi(\theta, s_2, s_1)$ be 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 marginal distribution $\pi(s_2, s_1|\theta)$ 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 a standard model of "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. This specification implies that the benefactor is *aware* of the ex-post realization of v, however. To allow for ex-post ambiguity, I allow the benefactor's payoff to 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 the distortions considered in earlier work. 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)). To clarify the source of distortions in the model I begin by analyzing the benchmark pure-altruism case, but then show in Section 5 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) 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. This is an extreme, dichotomic form of the more general idea that there are higher cognitive costs to convincing oneself of things that are less plausible given the data. I use this particular form purely for analytic simplicity.

Which of many possible states is most plausible remains ambiguous. In particular 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. 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.[10]

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 and memories of the past 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.

[10]The usual argument that agents who have played the same game many times should hold empirical priors is not useful here precisely because the benefactor does not observe θ ex post. Given this, 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.

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 I next characterize optimal beliefs.

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 $\overline{\theta}$. Given this one can show that optimal beliefs are, without loss of generality, Bayesian.

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

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 a few 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. A corollary is 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. (2013) 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 fixed-point property is one distinction between the model and models 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.[11] This highlights the fact that it is the absence of experienced consequences that makes the model distinctive. In the applications discussed here this is a natural assumption because the consequences are experienced by other people. In principle the same formalisms could be applied to purely selfish activities, however.

3 An Example

How informed will a well-intentioned benefactor become? This section illustrates the main ideas in the context of a simple example, using convenient functional forms and taking the decision about how much to give to a specific charity as a leading case. Section 4 then extends the results to arbitrary functional forms and decisions.

3.1 Learning to Give

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 preferences satisfy $v(\theta^g, d) > v(\theta^b, d)$ for all d. Don's prior is that $\pi(\theta = \theta^g) \equiv \gamma \in (0, 1)$. Don maximizes

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

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

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 *conditional* probability that he learns it after donating if he had not learned it before.

If Don observes 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 p he will learn the true state, while with probability $1-p$ he will obtain ambiguous information which he will optimally interpret as meaning that all is well ($\theta = \theta^g$). His future perception is thus $\hat{\gamma}_2 = 0$ with probability $p(1-\gamma)$ and $\hat{\gamma}_2 = 1$ with probability $1-p(1-\gamma)$. Given this, he optimally interprets the absence of news at time $t=1$ to mean that matters in Africa are good with probability $\hat{\gamma}_1 = 1-p(1-\gamma)$[12] and gives

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

Don's tendency to take a self-serving view of things shapes his motives for learning. Consider first what happens if he learns the truth ex post, after giving. 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 ($\theta = \theta^g$). Becoming informed thus cannot help and may hurt, forcing him to confront unpleasant realities ($\theta = \theta^b$).

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

This makes explicit the idea that information is primarily a constraint, rather than a resource. Information rules out possibilities that Don might otherwise have been able to believe in. This might seem to suggest that he will *never* want to learn. This turns out to be true in the limit case where Don is sure to *not* learn the truth ex post ($q = 0$). In that case his payoff when also informed ex ante is

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

while when he is not informed it is

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

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

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

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.

He thus obtains a benefit from being uninformed proportional to

$$\max_d [y - d + v(\theta^g, d)] - \max_d \left[y - d + v(\theta^b, d)\right] \geq \max_d \left(v(\theta^g, d) - v(\theta^b, d)\right) > 0 \tag{14}$$

The intuition here is the same, that information constrains the imagination. Absent any threat of real consequences, Don prefers maximum scope to "think positive." Yet things are less clear-cut when Don faces some real chance of ex post feedback ($p > 0$). To see this, consider the extreme case $p = 1$. Don's payoff when informed ex ante is again given by (12), but his payoff when uninformed ex ante is now

$$\max_d y - d + (\gamma)v(\theta^g, d) + (1 - \gamma)v(\theta^b, d) \tag{15}$$

and standard arguments can be used to show that this is less than his informed payoff, so that Don strictly values information. We thus have

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.*

Moreover, linearity in p implies that Don's motives for learning before giving are strictly greater the more likely it is that he will eventually learn the truth ex post. Figure 1 illustrates this graphically: Don's willingness to learn the state ex ante is increasing in the probability that he will learn it ex post, positive only when that probability is sufficiently high, and strictly lower at all interior points than his demand would be if he were making the decision for himself (which is equivalent to $p = 1$).

This pattern suggests a novel way to think about the value of outcome measurement. Economic analyses often highlight the value of measuring outcomes either in order to tie incentives to them or to enable learning for the future. In the good intentions framework there is an additional effect: measuring outcomes forces altruists to worry about them ex ante, rather than simply act on a plausible hypothesis and hope for the best.[13]

3.2 Nonprofit Marketing

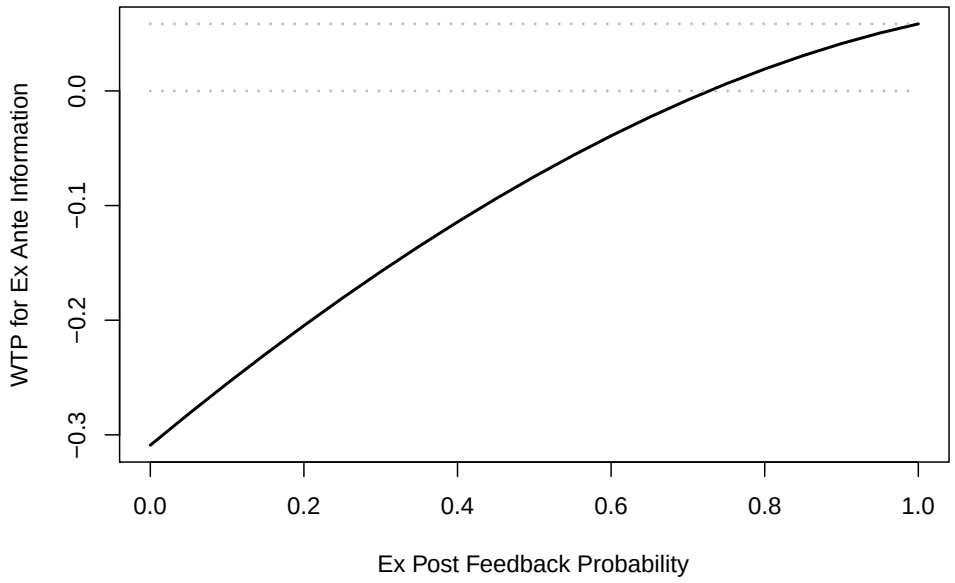
I next examine how Don's attitude towards learning shapes the marketing practices of an intermediary seeking to maximize donations. I focus in particular on the expected returns to generating public information – for example, commissioning a pre-announced randomized controlled trial. This lets me abstract from issues of strategic communication, which introduce a second impediment to learning, and focus exclusively on the role of donor demand.

Observation 3. *Ex post feedback (higher p) 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

[13] This is one interpretation, for example, of Muralidharan's (2012) argument that "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..."

Figure 1: Demand for Ex Ante Information on Effectiveness



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

uninformed ex ante, so that his donation is given by (11). The comparative static is

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

which shares the sign of $v_d(\theta^b, d) - v_d(\theta^g, d)$.

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.*

The statement of this result is complicated slightly by the fact that information will generically tend to affect the *expectation* of giving d even in a standard model ($p = 1$). This is the mechanism for persuasion studied in Kamenica and Gentzkow (2011), for example. To suppress this effect and isolate the *relative* effect of good intentions, define

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

Ex ante information thus has no average effect in the $p = 1$ case if $d^*(\gamma) = \gamma d^*(1) + (1-\gamma)d^*(0)$. Suppose this holds, and consider the case $p < 1$. 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 p if v is supermodular (submodular), and hence Don gives less (more) than $d^*(\gamma)$ when uninformed.

These two results both share the same underlying mechanism. Because Don prefers to believe that things are going well for Ben, 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. Both results seem broadly consistent with nonprofit marketing practices, which seem to emphasize "awareness-raising," broad aspirations, and depictions of need heavily relative to concrete information about what will be done with donations and how effective it is.

3.3 Beneficiary-Optimal Policy

The results above show that revenue-maximizing intermediaries have little incentive to generate evidence on effectiveness. One might expect incentives to be stronger for a policy-maker focused solely on the well-being of the beneficiary (or for the beneficiary himself). Interestingly, in the example above this is not the case. Because Don's only choice is *how much* to help, Ben is completely aligned with the intermediary in seeking to maximize total donations d. To break this result we need to introduce an additional dimension into Don's choice problem, letting him choose both how much and also how to give. Even then, however, there may exist a tradeoff between the quantity and quality of giving.

To illustrate this, suppose Don can now direct his donation to one of two causes $c \in \{s, r\}$; s represents a safe cause with certain returns $v(d)$, while r represents a risky one with returns equal to $\theta v(d)$. Don's expected payoff is

$$y - d + \begin{cases} (\hat{\gamma}_1 \theta^g + (1 - \hat{\gamma}_1)\theta^b) v(d) & c = r \\ v(d) & c = s \end{cases} \tag{18}$$

To make the choice of c non-trivial, assume that $\theta^g > 1 > \theta^b$. Don optimally chooses to give to the risky cause if $\hat{\gamma}_1 \theta^g + (1 - \hat{\gamma}_1)\theta^b > 1$ and to the safe cause otherwise. This immediately implies that research which reveals the true state (and thus forces Don to believe $\hat{\gamma}_1 = 1(\theta = \theta^g)$) will weakly increase the likelihood that Don gives to the more effective cause. To make this stark, suppose that $p = 0$, so that Don believes $\hat{\gamma}_1 = 1$ and gives to the risky cause unless he learns up-front that $\theta = \theta^b$. Revealing θ in this case strictly increases the probability that Don supports the more effective cause.

Revealing θ also affects the total amount he gives, however, which is defined implicitly by the first-order condition

$$v'(d^*(\hat{\gamma}_1)) = \max\{\hat{\gamma}_1 \theta^g + (1 - \hat{\gamma}_1)\theta^b, 1\} \tag{19}$$

If uninformed, Don gives a total of $v'^{-1}(\theta^g)$ to the risky cause, while if informed he does the same with probability γ or gives $v'^{-1}(1)$ to the safe cause with probability $1-\gamma$. The net effect of information on Ben's payoff is thus

$$(1-\gamma)[v(v'^{-1}(1)) - \theta^b v(v'^{-1}(\theta^g))] \tag{20}$$

The sign of this expression is ambiguous, as $\theta^b < 1$ but $v'^{-1}(\theta^g) > v'^{-1}(1)$. Intuitively, research that reveals the risky cause to be less effective than Don had hoped leads him to give to a more effective cause, but also leads him to give less overall. On net Ben could be either better or worse-off.[14]

This example highlights a basic tradeoff in the model between the *quantity* and *quality* of altruistic activity; similar results can be obtained for the effect of ex-post feedback. This raises the interesting question whether, for example, it might be better to allow funders to pursue a popular new idea such as micro-lending rather than risk "bursting their bubble" with disappointing evidence. Of course, the model says only that this is possible, not necessarily probable. The quantity/quality tradeoff may also help explain why misguided altruism can persist even in settings where the benefactor and the beneficiary can communicate directly. The beneficiary may find it optimal not to reveal his true needs or preferences if the benefactor is very excited about some other approach. Appendix B illustrates this in a cheap-talk setting.

3.4 Salience and Charitable Giving

Because it emphasizes thoughts rather than outcomes, the framework also helps rationalize some salience-related features of charitable marketing and giving. To see this, consider a simple extension in which Don thinks about Ben ex post with probability ρ. His expected payoff is

$$y - d + \rho\left[\hat{\gamma}_2 v(\theta^g, d) + (1-\hat{\gamma}_2)v(\theta^b, d)\right] \tag{21}$$

While the analysis above fixed $\rho = 1$, endogenizing it has several implications. First, 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)). 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.[15] 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

[14] To verify that the latter is possible, note that we can pick θ^b sufficiently close to 1 as to make (20) strictly negative.

[15] 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.

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."[16]

4 General Results

This section presents general results for arbitrary functional forms and choice sets. To articulate these I first define comparisons between 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.*

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).

[16] 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/`.

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

$$\arg\max_d \mathbb{E}_\mu[u(d) + v(d,\theta)] = \mathbb{E}_\mu[\arg\max_d u(d) + v(d,\theta)] \tag{22}$$

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. It also establishes the generality of the tradeoff between the quality and quantity of generosity discussed in Section 3.3

5 Alternative Motives

The results above describe an altruist with good intentions but whose altruism is otherwise "pure" in the sense that she cares about (her perceptions of) the beneficiary's welfare. I next examine how the good intentions framework interacts with other motives considered in recent work. I focus on motives that admit the following reference-dependent representation:

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

This specification extends (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 of preferences nests at least two cases of interest. 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 correspond to $\overline{d} = \arg\max_d u(d) \in D$, i.e. setting as reference the decision that maximizes the benefactor's own well-being. A second case is that of guilt-driven altruism. 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. One can then think of the benefactor as experiencing pride when she does "more" than $\overline{d}$ but guilt when she does "less." To isolate the guilt motive, let $\overline{d}$ represent a maximally generous action. If D is real-valued, for example, let $\overline{d} = \max D$.[17] This describes the opposite extreme to the impact philanthropy model; together the two cases thus bookend the set of possible reference points.

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

Although it admits a wide range of interpretations, (2b') turns out to have the same qualitative implications for benefactor behavior as the base "pure altruism" model. The proof is by simple redefinition: let $\tilde{v}(d, \theta) \equiv v(d, \theta) - v(\overline{d}, \theta)$ 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. This result suggests that the distinctive behavioral traits of a well-intentioned altruist – avoiding feedback, and conducting research ex ante only to avoid ex post regret – are quite general, irrespective of the details that underpin her altruism. Of course, there may still be important quantitative differences in how different altruists behave.

Donor motives do turn out to matter for market intermediaries, however. For a revenue-maximizing intermediary facing an impact philanthropist, the optimal strategy is to provide *no information at all*, including (ironically) information about impact. The intuition is that impact philanthropists want to believe in exactly those things that make the marginal return on their giving (and hence, the level of their giving) high. For example, a charitable donor motivated by impact wants to believe that the charity he supports is extremely effective and serves a disparately needy population. The intermediary is best-off facilitating this wishful thinking by providing no information, leaving the donor free to hold the beliefs most conducive to giving.

The reverse is true for guilty givers. These donors want to believe ex-post that even taking the most generous possible action $\overline{d}$ would have made little difference, so that they need not feel guilt. Such donors may seek to convince themselves, for example, that the need is not very great, or that all foreign aid is corrupt and never actually reaches people in need. This lets them give very little without experiencing guilt over missed opportunities. A fundraiser pitching such a donor thus benefits in expectation from the release of evidence of both need and efficacy. The donor would of course want to avoid this pitch.

The following Proposition formalizes these points.[18]

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 $\overline{d} = \min D$ ($\overline{d} = \max D$).*

[17] More generally, there may not be a uniquely most generous decision d independent of θ.

[18] Note that this result is consistent with Proposition 3, which cannot be applied here via substitution as in the proof of Proposition 4 since $v(d, \theta) - v(\overline{d}, \theta)$ need not be increasing in θ even if v itself is.

- *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.

6 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.

Crawford, Vincent P and Joel Sobel, "Strategic Information Transmission," *Econometrica*, November 1982, *50* (6), 1431–51.

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, Abhijit Banerjee, Rachel Glennerster, and Cynthia G. Kinnan, "The Miracle of Microfinance? Evidence from a Randomized Evaluation," Working Paper 18950, National Bureau of Economic Research May 2013.

_ 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.

Eil, David and Justin M. Rao, "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself," *American Economic Journal: Microeconomics*, 2011, *3* (2), 114–38.

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.

Grossman, Zachary and Joël van der Weele, "Self-Image and Strategic Ignorance in Moral Dilemmas," University of California at Santa Barbara, Economics Working Paper Series, Department of Economics, UC Santa Barbara 2013.

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.

Kamenica, Emir and Matthew Gentzkow, "Bayesian Persuasion," *American Economic Review*, October 2011, *101* (6), 2590–2615.

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 November 2013.

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

Beliefs consistent with Bayes' rule must satisfy

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

for all (θ, s_2, s_1). Consider the following family of history-contingent subjective beliefs:

$$\hat{\pi}(\theta, s_2, s_1) = 1(\theta = \overline{\theta}(d^*(s_1), s_2, s_1))\pi(s_2, s_1) \tag{23}$$
$$\hat{\pi}(\theta, s_2|s_1) = 1(\theta = \overline{\theta}(d^*(s_1), s_2, s_1))\pi(s_2|s_1) \tag{24}$$
$$\hat{\pi}(\theta|d, s_2, s_1) = 1(\theta = \overline{\theta}(d, s_2, s_1)) \tag{25}$$

where

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

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

$$\max_{\tilde{\pi}\in\Theta(s_2,s_1)} \mathbb{E}_{\tilde{\pi}}[v(d, \theta)] \tag{27}$$

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.

$$\arg\max_d \left[u(d) + \mathbb{E}_{\hat{\pi}(s_2|s_1)}[v(d, \theta)]\right] = \arg\max_d \left[u(d) + \mathbb{E}_{\pi(\theta,s_2|s_1)}\mathbb{E}_{\hat{\pi}(\theta|d,s_2,s_1)}[v(d, \theta)]\right] \tag{28}$$

This condition holds if

$$\hat{\pi}(\theta|s_1) = \mathbb{E}_{\pi(s_2|s_1)}[\hat{\pi}(\theta|d, s_2, s_1)] \tag{29}$$
$$= \mathbb{E}_{\pi(s_2|s_1)}[1(\theta = \overline{\theta}(d, s_2, s_1))] \tag{30}$$
$$= \sum_{s_2} 1(\theta = \overline{\theta}(d, s_2, s_1))\pi(s_2|s_1) \tag{31}$$

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

$$u(d^*) + \sum_{s_2} \left[\max_{\theta \in \Theta(s_2, s_1)} \{v(d^*, \theta)\} \right] \pi(s_2|s_1) \tag{32}$$

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

$$u(d^*) + \sum_{s_2} \sum_{s_2'} \left[\max_{\theta \in \Theta(s_2', s_1)} v(d^*, \theta) \right] \pi(s_2'|s_2, s_1)\pi(s_2|s_1) \tag{33}$$

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 (33) is at least as great as that in (32) for any particular (s_2', s_2) and hence (33) is also greater in expectation. Since (33) 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

$$U(s_2, s_1) \equiv \max_{d, \theta \in \Theta(s_2, s_1)} u(d) + v(d, \theta) \tag{34}$$

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

$$\begin{aligned} \Theta(s_2, s_1) &= \{\theta \in \Theta : \pi(\theta|s_2, s_1) > 0\} && (35) \\ &\subseteq \{\theta \in \Theta : \pi(\theta|s_2) > 0\} && (36) \\ &= \Theta(s_2) && (37) \end{aligned}$$

and hence the constraint in (34) 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 improves 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

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

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

$$u(d(s_1')) + \sum_{s_2} v(d(s_1'), \overline{\theta}(d(s_1'), s_2, s_1')\pi(s_2|s_1', s_1) \tag{39}$$

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 (39) as

$$u(d(s_1')) + \sum_{s_2} v(d(s_1'), \overline{\theta}(d(s_1'), s_2, s_1)\pi(s_2|s_1) \tag{40}$$

which must by definition be weakly less than (38) 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

$$f(d, \{x(s_2', s_2, s_1)\}) \equiv u(d) + \sum_{s_2} \sum_{s_2'} v(d, x(s_2', s_2, s_1))\pi(s_2'|s_2)\pi(s_2|s_1) \tag{41}$$

where

$$x(s_2', s_2, s_1) = \max\{\theta : \pi(\theta, s_2, s_1) > 0\} \tag{42}$$

in the case where he observes S_2 and

$$x(s_2', s_2, s_1) = \max\{\theta : \pi(\theta, s_2', s_1) > 0\} \tag{43}$$

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

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

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:

$$\sum_{s_1} \left[\arg\max_d u(d) + \sum_{s_2} v(d, \overline{\theta}(s_2, s_1))\pi(s_2|s_1) \right] \pi(s_1|s_1') - \sum_{s_1} \left[\arg\max_d u(d) + \sum_{s_2} v(d, \overline{\theta}(s_2, s_1'))\pi(s_2|s_1) \right] \pi(s_1|s_1') \\ + \sum_{s_1} \left[\arg\max_d u(d) + \sum_{s_2} v(d, \overline{\theta}(s_2, s_1'))\pi(s_2|s_1) \right] \pi(s_1|s_1') - \arg\max_d u(d) + \sum_{s_2} v(d, \overline{\theta}(s_2, s_1'))\pi(s_2|s_1') \tag{45}$$

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

$$\arg\max_d u(d) + \sum_{s_2} v(d, \overline{\theta}(s_2, s_1))\pi(s_2|s_1) - \arg\max_d u(d) + \sum_{s_2} v(d, \overline{\theta}(s_2, s_1'))\pi(s_2|s_1) \quad (46)$$

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

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

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

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

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

$$\mathbb{E}\left[\arg\max_d u(d) + v(d, \overline{\theta})|s_1\right] = \arg\max_d u(d) + \mathbb{E}[v(d, \overline{\theta})|s_1] \quad (49)$$

Taking expectations of both sides over S_1 yields

$$\mathbb{E}\left[\arg\max_d u(d) + v(d, \overline{\theta})\right] = \mathbb{E}\left[\arg\max_d u(d) + \mathbb{E}[v(d, \overline{\theta})|S_1]\right] \quad (50)$$

which together with (48) implies that (47) 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

$$\max_{\theta \in \Theta(s_2, s_1)} v(d, \theta) - v(\overline{d}, \theta) \quad (51)$$

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

A core result in the model is that benefactors may not want information about the impact of their actions, since new information may limit the extent to which they can believe in good outcomes. In some settings this result is decisive, but in others the beneficiary may also have opportunities to communicate information to the benefactor. For example, givers and receivers of holiday gifts may talk beforehand about the kinds of things the receiver likes. It is therefore worth understanding whether such communication will tend to eliminate information asymmetries between the two. A full analysis of this issue is beyond the scope of the paper, but I provide her an example to illustrate that the beneficiary may find it in her best interest to conceal information from the benefactor.

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

$$v(\theta, d) = \theta^a d^a + \theta^b d^b \tag{52}$$

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

$$u(y - d^a - d^b) + \mathbb{E}_{\hat{\pi}}[\theta^a d^a + \theta^b d^b] \tag{53}$$

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

$$\hat{\pi}(\theta^i = x|m) = 1(x = \max\{\theta^i : \mathbb{P}(m|\theta^i) > 0\}) \tag{54}$$

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

$$\overline{\theta}^i(m) = \max\{\theta^i : \mathbb{P}(m|\theta^i) > 0\} \tag{55}$$

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

$$i^*(m) = \arg\max_{i \in \{a,b\}} \{\overline{\theta}^i(m)\} \tag{56}$$

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

$$u'(y - d^*(m)) = \overline{\theta}^{i^*(m)}(m) \tag{57}$$

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

$$\max_{m \in M} d^*(m)\theta^{i^*(m)} \tag{58}$$

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.[19] 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 (58) 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

$$\frac{\theta^a}{\theta^b} \geq \frac{d^*(m^b)}{d^*(m^a)} \tag{59}$$

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

$$u'(y - d^*(m^a)) = \max\left\{\theta^a : \exists \theta^b \text{ such that } \pi(\theta^a, \theta^b) > 0 \text{ and } \frac{\theta^a}{\theta^b} \geq \frac{d^*(m^b)}{d^*(m^a)}\right\} \tag{60}$$

$$= \overline{\theta}^a \tag{61}$$

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

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.[20]

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."

[20] 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.