False Positives and Transparency in Scientific Research

JONATHAN LIBGOBER

Department of Economics, Harvard University

First version: March 25, 2015
This version: April 26, 2018

ABSTRACT. This paper develops a model of costly information acquisition, focusing on an application to scientific research. When dimensions of an experiment that may bias a scientific result are not verified, scientists are incentivized to make their experiments more susceptible to false positives, even though they obtain higher surplus from more informative experiments. On the other hand, non-transparency can induce a scientist to undertake a costlier but more informative experiment if it also enables her to commit to acting scrupulously. Using this insight, this paper demonstrates the general existence of non-degenerate experiment costs such that greater transparency in scientific methodology results in research choices that are worse for those interested in the results.

KEYWORDS. False positives, sender-receiver games, information acquisition, experimentation, transparency.

JEL Codes. D82, D83.

Any analysis that relies upon statistical inference inevitably risks arriving at an incorrect conclusion. Nevertheless, as argued in Ioannidis (2005), there are compelling reasons to believe that mistakes in published research cannot be explained by statistical error alone, and arise due to decisions in experimental design leading to bias (i.e., positive results becoming more likely,

CONTACT. jlibgober@fas.harvard.edu. This paper supercedes my earlier draft, “False Positives in Scientific Research,” and comes from the second chapter of my PhD thesis. I thank my committee, Drew Fudenberg, Eric Maskin, Jerry Green and Ben Golub, for their encouragement and advice. I am also especially grateful to Johannes Hörner for guidance on this project in its early stages. Finally, I thank Philippe Aghion, Isaiah Andrews, Vivek Bhattacharya, Kirill Borusyak, Gonzalo Cisternas, Mira Frick, Oliver Hart, Scott Kominers, Harry Di Pei, Neil Thakral, Kathryn Spier, Tomasz Strzalecki, Juuso Toikka, Carl Veller and Heidi Williams for very helpful conversations, and seminar participants at Harvard for excellent feedback. Any remaining errors are my own.
irrespective of hypothesis validity). A natural question arises as to how to evaluate policies designed to combat false positives and improve research quality.

Toward that end, the medical research journal *The Lancet* published a series of articles in January 2014 discussing guidelines to improve the efficiency of scientific research.\(^1\) The prevalence of false positives was one particular focus, and partially attributed by Ioannidis et. al. (2014) to the lack of documentation requirements for experimental conduct. During any experiment, researchers make a number of decisions that could lead to bias, and not all of them will be described in the resulting publication. Examples discussed by Ioannidis et. al. (2014) include selecting samples for high-risk patients, failing to adapt significance levels to the number of tests run, selective reporting of results, or insufficient training of clinical researchers.

Ioannidis et. al. (2014) suggest implementing registration requirements to combat these issues, with the understanding that scientists typically need some kind of external certification to make credible claims regarding research activity.\(^2\) But given the varying degree of difficulty in verifying distinct activities that could lead to bias, these requirements are not a simple yes-or-no matter. Instead, implementation would occur on specific dimensions of an experiment. It may be easy for some outside authority to verify properties of certain samples. But it may be difficult to verify that research assistants were well-trained, or to distinguish a genuine need to restart an experiment from disappointment with a negative result. Answering whether a particular registration policy is beneficial requires an analysis of how the properties of experiments will change in response.

Stepping back, however, followers of the applied mechanism design literature could be skeptical that transparency requirements necessarily lead to improved research output. To see why, note that limited contractibility is the rule, not the exception, for many types of scientific research. A researcher studying whether a particular gene can lead to a particular kind of cancer may not know which pharmaceutical company would be able to use that insight in order to develop a treatment. And fundamental research is often left to universities this research requires a time horizon much longer than private entities would be capable of providing.

In settings with limits on contractibility, several authors have cast doubt on the optimality of transparency requirements, even *without* resorting to costs of monitoring.\(^3\) As discussed in the literature review, Prat (2005) notes that transparency requirements are eschewed in a variety of settings, such as corporate governance, where one may expect them to be beneficial. His model

\(^1\)The biologist Ed Wilson wrote a letter in response that, while praising the series in general, specifically lamented the lack of involvement from economists.

\(^2\)Some researchers have even considered the effects of similar policies in economics and social science (Miguel et al. (2014), Coffman and Niederle (2015), Olken (2015)).

\(^3\)Of course, if one truly thought transparency requirements contributed to false positives, one could always assert that false positives are a priori problematic and should be minimized. But as Glaeser (2006) notes, this is unlikely a natural objective. This paper adopts his view that, if false positives were widely known to be prevalent (evidenced by the number of citations of Ioannadis (2005)), observers should read results skeptically and debias accordingly.
shows that agents can be more willing to act according to their private information under non-
transparency in a career concerns model. Similarly, Cremer (1994) ties transparency requirements
to the ability to provide high powered incentives within organizations. His model explains some
documented features regarding firm boundaries.

This paper comments on the efficacy of transparency requirements in a simple framework
that can shed light on the aforementioned debate in scientific research communities. In our
model, a scientist (she) chooses an experiment that is characterized by a vector of research
activities, and produces an observable outcome (success or failure) that is seen by a developer (he).
One dimension could reflect the number of samples collected; another, the number of times
the experiment is repeated; or whether they are disingenuous and decide to directly alter their
data to increase the probability of a positive result. The experiment imposes a cost on the scientist,
but provides information as to whether or not the developer will be able to successfully develop a
drug (which would yield a benefit to both players) by exerting costly effort.

This paper focuses on the question of whether the inability of the developer to observe
experimental methods (due to lack of transparency) makes him better off or worse off. The
main result highlights that whether transparency requirements are advantageous depends on the
complementarity in costs between different kinds of research actions.

To illustrate the intuition, consider the discovery of the Higgs Boson in 2012 using the Large
Hadron Collider (LHC) at CERN. Discovery actually meant that five-sigma confidence (a chance of
roughly 1 in 3.5 million) had been reached. In other words, the existence of the Higgs Boson was
a statistical finding, inevitably short of mathematical certainty (though, depending on opinion,
perhaps not by much). The discovery was a remarkably high profile event that garnered widespread
celebration, and culminated in a Nobel Prize for Englert and Higgs the following year (Wired
(2015)). However, the LHC was shut down for two years after the discovery, preventing replication.
Furthermore, the data from the experiment were not released for another four years, and hence
the algorithms that had gone into the particle’s discovery went unexamined.

Why was the physics community apparently unconcerned by the lack of public data and the
inability to replicate the experiment? Presumably the lack of transparency was not a concern due
to the difficulty in being able to bias an experiment with a five-sigma significance threshold. The
five-sigma threshold is common in physics due to the feasibility of collecting large amounts of
data (particularly at CERN). This does not apply to other empirical disciplines like biology. Still,
it is often the case that a dataset’s size is left to the discretion of researchers. But due to limited
contractibility over experiments, scientists typically cannot be compelled to incur the costs of

---

4 We use “developer” here in order to make our story more concrete. For interpretation purposes, we can think of this
person as someone intrinsically interested in the results of the scientist’s experiment, for whatever reason.
5 Actually, three sigma confidence had already been obtained. That level of confidence is uncommon in other disciplines.
additional data collection directly.

This paper highlights that transparency requirements have the benefit of discouraging bias, but might also discourage scientists from undertaking costlier experiments to counteract the perception of bias. Imagine a researcher deciding to collect a large dataset (e.g., the equivalent of an experiment at CERN) or a small dataset. Large dataset experiments may be much more difficult to undertake than small dataset experiments. If the scientist is required to make all data and regression specifications public, then they could very well prefer the small dataset experiment, believing it is informative enough and finding the large dataset too costly. Without transparency requirements, however, an outsider observing a positive result would assume that some kind of biasing activity (e.g., regression fishing) took place and therefore not adapt his beliefs to the outcome of the experiment. Since the large dataset experiment is less susceptible to this kind of bias (as appeared to be the case for the Higgs Boson discovery), this loss in credibility would then induce the scientist to compensate by using the larger dataset. This compensation, in turn, can overwhelm the loss of informativeness due to biasing.

Framed in this way, a lack of transparency requirements encourage the use of large dataset experiments (such as those performed at CERN) by instead discouraging the use of less informative experiments that were susceptible to bias. In general, the strength of these complementarities determine whether non-transparency simply adds bias or can be compensated for more effectively along other dimensions. This result is similar to those on the optimality of “money burning” that can arise in delegation settings (for instance, as in Ambrus and Egorov (2014)). However, this conclusion is only true if the scientist’s preference for developer success is sufficiently important, since otherwise non-transparency will not have this beneficial effect.

This paper contributes to the aforementioned policy debate by clarifying the connection between preferences of scientists and the merits of transparency requirements. There are two key features of the model which drive our conclusions: First, scientists care about follow-on research, and second, difficulty or costs associated with experiments influence experiment choice. Debates over transparency requirements should take these incentives into account. For instance, new researchers may be preoccupied with making tenure or developing a reputation in the discipline, while established researchers may be influencing drug companies directly and want to ensure that their research is useful. Depending on the form these career concerns take, our analysis in Section 4.2 suggests that it may be undesirable to apply the same transparency requirements for early-career grants versus late-career grants.

Finally, despite our focus, there is no reason the model should only be applicable to scientific research. But the justification of preferences and limits to contractibility seem appropriate for this application. And as mentioned above, transparency over differing research activities is an active policy question. While the model could certainly describe other kinds of information acquisition,
we leave a more thorough analysis of these applications to future work.

We proceed as follows. We first discuss the literature and introduce the model in Section 2. Readers interested in the mechanism highlighted but not the general model are encouraged to skip to Section 3.1, which show the key forces at work. In Section 3.2, we consider the scientist’s equilibrium behavior, and in Section 3.3, provide a cost function for any arbitrary experiment set such that drug developers are better off without transparency requirements. We use the model to comment on policy in Section 4. We proceed to consider a number of alternate specifications for the analysis in Section 5, and conclude in Section 6. Most proofs are in the Appendix.

1. RELATED LITERATURE

The non-transparency result provided in this paper is similar to a number of others that have been derived in the literature. Results of this form can be found in Cremer (1994), Prat (2005) and Bergemann and Hege (2005). In these papers, the intuition behind the optimality of non-transparency is that it gives the principal additional commitment power which would not be credible under full transparency. The results of this paper are in a similar spirit, but allow for a variety of research actions, and also provide a central role for the incentives to induce follow-on work. A more direct contrast is that the present paper obtains non-transparency as a way of burning surplus in a way that aligns the incentives of the principal and the agent. The presence of another action that is incentivized by this “money burning” is crucial for our result, and does not have a direct counterpart in these papers. Importantly, this feature relies upon the high degree of alignment between preferences of scientist and developer.

When framed as a result of money burning, our results are therefore actually closer to Angelucci (2014) or Szalay (2005), whereby the incentives are aligned when the principal takes actions which seem to harm both players. But in these papers, the distortions take other forms. Ambrus and Egorov (2014) consider cases where money burning is part of an optimal contract in a delegation setting, though our counterfactual of non-transparency does not nest in their framework.

The model itself is reminiscent of multitasking, Bayesian Persuasion (a la Kamenica and Gentzkow (2011)) and career concerns. Multitasking arises due to the variety of research actions the scientist undertakes, resembling the literature following Holmström and Milgrom (1991). This literature has showed that transparency over different dimensions can distort an agent’s effort choice. Still, as these papers typically involve full observability of the agent’s actions as achieving the first best, the contribution in this paper seems distinct.

The literature on communication games (and Bayesian Persuasion in particular) has flourished recently and been used for many applications. In fact, the idea of using these models to study

---

6 Other papers give conditions under which the principal is better off when the agent is not able to perfectly observe a state variable, for example, Jehiel (2014) and Ederer, Holden and Meyer (2014).
scientific research is not in itself novel, as it is also done by Kolotilin (2015). The main difference with these settings is (1) we do not allow all information structures to be feasible, (2) we impose costs that are parameterized with the information structures, and (3) we consider a case where the sender can only commit to part of the information structure. This third point is perhaps the most significant departure. Hoffmann, Inderst and Ottaviani (2014) are also interested in the lack of commitment, but view it as arising from a disclosure problem. Several papers have consider cases where distortions of information can take particular forms, such as fraud as in Lacetera and Ziruliu (2008), or selective disclosure as in Henry (2009) and Felgenhauer and Schulte (2014). Other papers have studied information aquisition and communication in cheap talk models; see Argenziano, Severinov, and Squintani (2014), as well as Pei (2014).

This introduction of signal distortion is reminiscent of the career concerns literature (Holmström (1999) and Dewatripont, Jewitt and Tirole (1999)). While we accomodate these preferences for the scientist, our model provides a novel channel for preferences over informational content (as opposed to just the outsider’s posterior). We distinguish these incentives from one another. Incentives for information acquisition is represented by the convexity of the scientist’s expected payoffs as a function of the developer’s beliefs. On the other hand, the marginal benefit from distorting can be most clearly seen by studying the slope of the expected payoff conditional on the state (i.e. the truth of the hypothesis).

Lastly, this paper relates to the literature on academic publication (as in Aghion, Dewatripont and Stein (2005), for example). Azoulay, Bonatti and Krieger (2015) empirically study the effect of a retraction on a scientist’s reputation, documenting that retractions lead to a drop in citations consistent with reputation loss. Andrews and Kasy (2017) provide methods for determining when publication bias may be important, and suggest a way to debias taking this into account. Yoder (2016) develops a principal-agent model with transfers, similarly motivated by Bayesian Persuasion, to describe the optimal incentives for research institutions. His conclusion that negative results should be rewarded is consistent with this paper, since doing so can align incentives.

Several papers have cautioned against associating false positives with problems in scientific conduct. Glaeser (2006) studies the incentives behind false positives, and argues that eliminating them may be socially harmful. His reasons for this do not directly relate to our overcompensation effect, instead focusing hypothesis choice. Kiri, Lacetera and Ziruliu (2015) consider the incentives for fact-checking, arguing that failure to observe mistakes would suggest the lack of verification activities. Furukawa (2017) develops a vote-counting model to show that publication bias may arise naturally when results are only coarsely interpreted by practitioners, and suggests a way to correct for it. Hopefully these ideas will be useful in the design of research guidelines, and will further improve our understanding of how to effectively structure scientific endeavors.
2. MODEL

A scientist is endowed with a hypothesis whose validity is given by $\theta \in \{T, F\}$, drawn by nature. The scientist is able to conduct an experiment, the results of which are of interest to a drug developer. All players share a common prior on $\theta$, with $P[\theta = T] = p_0$. We will always take $p_0$ to be interior, and we will think of $T$ as being the “good” state, and $F$ as the “bad” state. The hypothesis could be whether a particular gene is associated to a specific disease, or whether a certain particle can safely target a specific biological pathway.

The scientist chooses an experiment from some set of possible experiments. We assume that the set of experiments are parameterized by an $n$-tuple $(a_1, \ldots, a_n)$, with each $a_i \in [a_i, \overline{a}_i] \subset \mathbb{R}$, yielding an experiment $I(a_1, \ldots, a_n) : \Theta \to \Delta\{0, 1\}$ at cost $c_S(a_1, \ldots, a_n)$. For now, we only assume that the cost is increasing in each coordinate, and we highlight that we will not assume that higher actions yield more informative experiments. As mentioned in the introduction, each dimension is meant to capture a different kind of research activity; collecting data, $p$-hacking activity, selecting samples, and so on.

A restriction of the above setting is that the experiment outcome can only take one of two values, so that the developer only observes experiment produces an outcome $y \in \{0, 1\}$ according to a distribution that depends on $\theta$ and $a \in A = \prod_{i=1}^{n}[a_i, \overline{a}_i]$. This outcome $y \in \{0, 1\}$ is observable to the developer, and we will refer to the event $y = 1$ as a “positive result,” and $y = 0$ as a “negative result.” A false positive occurs when $y = 1$ and $\theta = F$. Define:

$$h_\theta(a) := P[y = 1 \mid \theta, a],$$

and assume this function is continuous with bounded derivatives. In order to interpret positive results as evidence for hypothesis validity, we assume that $h_T(a) > h_F(a)$, for all $a$.

The developer does not necessarily observe the entire profile of research activities $a$. Instead, we assume that there is some third party (e.g., a journal or a funding agency) that can dictate what the developer observes. Formally, the third party chooses a set of indices, $M \subset \{1, \ldots, n\}$ with the interpretation that the developer will observe all $a_i$ such that $i \in M$, in addition to the outcome $y$ itself. The interpretation is that the third party has the ability to make various dimensions, such as the number of specifications run, observable to the developer. In principle, one could imagine that there are costs associated with different choices of $M$, but since the implication of those costs is straightforward we do not model them explicitly. However, one could imagine that in practice these costs are prohibitive for the scientist. We are interested in how the developer’s

---

7The scientist can be thought of as a sender, or an agent, and the developer as being a receiver, or a principal.

8Continuity ensures the existence of a pure strategy equi in order to ensure the existence of a pure strategy equilibrium with unobserved coordinates.
payoff changes with different choices of $M$.

After observing the action choice and $y$, the developer updates his prior from $p_0$ to $\hat{p}(y)$. Since the developer’s beliefs will depend on equilibrium behavior, we will typically suppress the dependence of the posterior $\hat{p}(y)$ on the experiment choice of the scientist, but occasionally we will make this dependence explicit when the choice is known to be $a = (a_1, \ldots, a_n)$ by denoting the posterior $\hat{p}_a(y)$.

The developer chooses a level of effort $e \in [0, 1]$ at cost $c_R(e)$, where $c_R(e)$ is an increasing and convex function. The choice of $e$ determines the realization of a random variable $x \in \{0, 1\}$, the distribution of which is given by:

$$P[x = 1 \mid \theta = T, e] = e, \quad P[x = 1 \mid \theta = F] = 0.$$ 

We think of $x = 1$ as the event that a drug is developed, and $x = 0$ as the event that it is not. For example, the drug could attempt to cure the disease by deactivating the gene studied in the scientist’s experiment. Increasing effort in drug development makes it more likely to succeed, but only in the event where the gene is in fact associated with disease incidence—that is, if the scientist’s hypothesis is true.

If the drug is developed, the developer obtains a payoff of $b > 0$, and the scientist obtains a payoff of $\lambda$. As mentioned, for the baseline model we imagine there is limited contractibility between scientist and developer, so we take these payoffs as being exogenous. In that case, $\lambda$ might reflect profit, whereas $b$ may reflect prestige or pride associated with having cured a disease. We also suppose the scientist also receives a benefit of $g(\hat{p})$ when the public belief is $\hat{p}$ at the end of the game, for some increasing function $g$. We refer to this as payoffs coming from career concerns; insofar as scientists may (be thought to) have hypothesis validity correlated across time, then they may place some premium on having future drug developers believe that their past hypotheses were true. We use this interpretation to distinguish our specification from other papers and to discuss policy recommendations. In an appendix, we discuss a version of the model where the scientist also has preferences over $y$ itself.

Hence final payoffs for the scientist are

$$\lambda \cdot x + g(\hat{p}) - c_S(a),$$

whereas for the developer, they are

$$b \cdot x - c_D(e).$$

We assume that the parameters are such that the developer’s optimal effort choice always
2.1. Measuring Uncertainty

In this section we place some restrictions on how we parameterize information structures. We also introduce the following terminology:

**Definition 1.** A dimension \( a_i \) is biasing if \( h_T(a_i, a_{-i}) \) and \( h_F(a_i, a_{-i}) \) are both increasing in \( a_i \), for all \( a_{-i} \). A dimension \( a_i \) is informative if \( h_T(a_i, a_{-i}) \) is increasing in \( a_i \) and \( h_F(a_i, a_{-i}) \) is decreasing in \( a_i \), for all \( a_{-i} \).

For the main results of the paper, we will assume that experiments are Blackwell ordered along each dimension (increasing along informative dimensions and decreasing along biasing dimensions). This assumption is stronger than what is necessary, although without it, we would need stronger assumptions on preferences in order to ensure that actions could be classified as biasing or informative. For example, our results would continue to hold if we assumed quadratic developer effort costs and considered experiments ordered by the reduction in posterior variance (which is implied by Blackwell ordering of experiments but not conversely). We adopt this specification for analysis of Examples 1 and 3 below.

Note that the developer’s expected payoff, given effort \( e \) and belief \( p \), is given by:

\[
\pi_R(\hat{p}) := \max_e b\hat{p}e - c_R(e).
\]

We adopt the terminology of Ely, Frankel and Kamenica (2015) to call a measure of uncertainty to be a strictly concave function that is 0 at degenerate beliefs. Define the developer’s measure of uncertainty from an experiment \( I \) to be \( p_0\pi_R(1) - \mathbb{E}_{y \sim I}[\pi_R(\hat{p}(y))] \).

**Proposition 1.** The payoff gain of the developer from observing the scientist’s experiment is a measure of uncertainty.

It is immediate from the definition that the developer’s payoff is higher for experiments that are more informative according to his measure of uncertainty. The proposition, together with our ordering on experiments, implies that the optimal scientist experiment for the developer would involve the maximal action on informative dimensions and the minimal action on biasing dimensions.

2.2. Examples of Information Acquisition Technologies

To illustrate the analysis, we provide two stories behind information acquisition technologies which satisfy the assumptions of the model. These are meant to demonstrate a tighter link between

\[^9\text{A sufficient condition that ensures this, for example, would be that } c_R(e) = \frac{k}{2}e^2 \text{ and } b < k.\]
what the model captures and the kinds of scientist behavior practitioners tend to be concerned about.

**Example 1** (p-hacking). First consider the case of \( c_R = \frac{k}{2} e^2 \), so that we can assume experiments are ordered by posterior variance (as opposed to Blackwell ordered). Suppose the scientist chooses a number of times, \( a_2 \in [3, 15] \) to run an (iid) experiment and a precision \( a_1 \), where:

\[
\mathbb{P}[\text{Success on an experiment } | \theta = T] = \mathbb{P}[\text{Failure on an experiment } | \theta = F] = a_1.
\]

The resulting informativeness (as measured by the posterior variance) is plotted for \( a_1 = 2/3 \) and \( p_0 = 1/2 \) in Figure 1. While it increases at first, eventually it decreases until the experiment is basically uninformative given a high enough number of experiments. We can therefore think of one dimension as being the amount of p-hacking and the other to be the informativeness of the underlying experiment (which are biasing and informative, respectively, on some appropriate range).

**Example 2** (Lying or Fudging Data). Again suppose that \( a_1 \) parameterizes the underlying experiment as in the p-hacking example. However, now suppose that \( a_2 \) is the probability that the scientist changes a result of \( y = 0 \) to a result of \( y = 1 \)—for example, as a result of either direct falsification or altering the data. This is a Blackwell garbling of the underlying signal technology, and hence decreases the informativeness for any informativeness measure, not just the relevant one for the developer (i.e. the posterior variance).
Example 3 (Quality Decisions). Now suppose that $a_1$ is investment in research equipment, and that $a_2$ measures the quality of lab technicians or research assistants (supposing it is as easy to hire good research assistants as bad ones). Suppose the resulting information structure is

$$
P_{I(a_1, a_2)}[y = 1 \mid T] = (4/5)a_1 + (1/5)a_2, \quad P_{I(a_1, a_2)}[y = 1 \mid F] = (1/5)a_1 + a_2,
$$

where $a_2 \in [a, \bar{a}]$ reflects employee quality and $a_1 \in [q, \bar{q}]$ reflects equipment quality, with parameters chosen so that probabilities are within $[0, 1]$ and $P_{I(a_1, a_2)}[y = 1 \mid T] > P_{I(a_1, a_2)}[y = 1 \mid F]$.

Of course, our model would also accommodate hybrids of these technologies, and most of the results turns out to not depend on the particular functional forms of the set of experiment space. The underlying point is that the model can accommodate natural qualitative features of experimentation, treating different kinds of research activities as observationally distinct.

3. MAIN RESULTS

3.1. A Simple Example

In this section, we illustrate our construction and walk through the intuition in a numerical example. We take $p_0 = 1/4$, and specialize to the quality example of Example 3. Take we take $c_R(e) = \frac{1}{2}e^2$ and $b = \lambda = 1$. Suppose the experiments the scientist has access to are as follows, for $a_1, a_2 \in \{0, 1\}$:

$$
P_{I(0, a_2)}[y = 1 \mid T] = 2/5 + (1/10)a_2, \quad P_{I(0, a_2)}[y = 1 \mid F] = a_2/6,$$

$$
P_{I(1, a_2)}[y = 1 \mid T] = 4/5, \quad P_{I(1, a_2)}[y = 1 \mid F] = 0,
$$

where $P_e$ denotes the probability measure when experiment $e$ is chosen. One can check that as $a_2$ increases, the experiment $I(0, a_2)$ becomes less informative in the sense of Blackwell. Further note that $a_2$ does not affect the informativeness of the experiment $I(a_1, a_2)$.

An interpretation of this parameterization comes from Example 3, where $a_1$ may reflect equipment quality and $a_2$ may reflect training or quality of research assistants. The idea is that experiments with “low quality equipment” are both less informative, and susceptible to bias depending on the research assistants. In contrast, “high quality equipment” gives an experiment which does not require the help of research assistant, and is also less informative no matter what. The important feature of this example is that while these actions might make an experiment susceptible to false positives, a positive result is already very likely if the equipment is of high quality, provided the hypothesis is true. We suppose experiments with low quality equipment are costless, but high quality equipment is costly. Research assistant quality, in contrast, is costless.
Suppose the developer sees the complete experiment chosen by the scientist—both equipment quality $a_1$ and research assistant quality $a_2$. For instance, suppose the funding agency requires documentation of equipment and can require research assistants to satisfy certain prerequisites. Discreteness is helpful in that we can compute payoffs experiment-by-experiment: For $i \in \{ S, R \}$,

$$
\pi_i(0, 0) = 1/8 \quad \pi_i(0, 1) = 1/12 \\
\pi_i(1, 0) \approx 1/5 \quad \pi_i(1, 1) \approx 1/5,
$$

where $\pi_S(a_1, a_2)$ and $\pi_R(a_1, a_2)$ are the benefits to the scientist and developer, respectively, from an experiment of type $(a_1, a_2)$.

Now suppose that the developer can observe equipment quality $a_1$, but not research assistant quality $a_2$, leaving research assistant quality unverfied. In this case, if the scientist picks $a_1 = 0$, then $a_2 = 0$ will not be chosen when the choice of $a_2$ is costless. To see this, simply note that the scientist’s expected payoff is always higher following a signal of $y = 1$ than a signal of $y = 0$. On the other hand, the developer cannot distinguish $a_2 = 0$ and $a_2 = 1$, and so in equilibrium his belief will not change with the choice of $a_2$. Since higher $a_2$ generates a higher probability of $y = 1$ when $\theta = T$, the scientist opts for lower quality research assistants.

Does this mean that the experiment the scientist chooses is less informative when $d$ is unobserved? Not necessarily, due to costs. If the cost of high quality equipment is either sufficiently small or sufficiently large, the developer does best if both $a_1$ and $a_2$ are observed. If the costs of high quality equipment are small, then the scientist would choose this experiment in either regime and hence bias due to research assistants would not be a concern. On the other hand, if the cost of high quality are sufficiently large, then the scientist will never choose these experiments and hence all non-transparency does is induce added bias. For an intermediate range of costs\(^{10}\), however, the scientist switches choice of $a_1$ across the regimes. Under observability, low quality equipment is

The example highlights three main messages which are useful for understanding our main results. First, the scientist exhibits a strict preference for more informative experiments.\(^{11}\) Specifically, the above calculations verified that the scientist would always prefer a more informative experiment if the extra information were free. Our analysis in Section 3.2 shows that this is a general feature for a wide variety of settings. Notice, however, that the scientist does not gain anything if the developer’s beliefs are unduly optimistic when $\theta = F$, and in fact strictly prefers that the developer know the state when $\theta = T$. Furthermore, the more informative the experiment, the

\(^{10}\) Specifically, costs between $3/40$ and $7/60$ (approximately).\(^{11}\) This does not follow immediately from Blackwell’s Theorem, since the scientist is not the decision maker who uses the information. See Kim (1995) for a discussion of the use of Blackwell informativeness in a principal-agent model with moral hazard.
higher beliefs are (in expectation) when $\theta = \mathcal{T}$. We will show how the state dependent nature of the scientist’s payoff as a function of beliefs translates into an incentive for information acquisition, which will allow us to distinguish these incentives from the incentives for distortion.

Second, despite liking informativeness, when $a_2$ is unobserved the scientist loses credibility for scrupulousness. Since beliefs do not respond to the choice of distortion, and since the posterior rises if and only if $y = 1$, increasing $a_2$ increases the probability that $\hat{p}$ is higher when $\theta = \mathcal{T}$. Hence even though the scientist would like to set $a_2 = 0$, she cannot commit to doing so when unobserved, since the developer will realize that it is profitable for her to deviate to $a_2 = 1$. This is also studied in Section 3.2.

Third, despite the loss of credibility for scrupulousness, the scientist compensates by exerting costly effort which ultimately increases the informativeness of the chosen experiment. Since the scientist has a preference for more information, the only reason a scientist would choose a less informative experiment over a more informative one would be on account of costs. So by making it impossible to commit to $a_2 = 0$, the scientist is induced to take a costly action which proves her scrupulousness. Here, this takes the form of choosing an even more informative experiment. In other words, the scientist needs to exert more costly effort (in this case, acquire higher quality equipment) in order to prove that the experiment is actually informative.

We are able to use the intuition gained from this example to demonstrate the optimality of partial transparency without having to resort to particular specifications of the informational environment. Despite this, the example is still somewhat restrictive. It rules out, for instance, higher $a_2$ having any influence on the experiment with higher $a_1$. It is also ad hoc in that it rules out multiplicity and the possibility of mixed strategies. Under the generality of the main model, these main not be ensured, and in these cases, the intuition from above cannot be applied directly. Our main results show that these can be ignored under certain specifications for cost functions. The generality of the full model is also useful in explaining why our result could not occur in other common benchmarks studied in the literature (that is, without both costly communication and limited commitment). While this may not be obvious from this specialized setting, the comparison is more direct in our general model.

### 3.2. Influence of Transparency on Experiment Choice

This section describes the how observability of $a_i$ influences the experiment choice of behavior. We demonstrate how to adopt the belief-based approach to this setting. Gentzkow and Kamenica (2014) remarked that this is less straightforward when there are costs, since the payoff may depend on the signal structure outside of the beliefs they induce. Nevertheless, we are able to use this for our specification.

The scientist takes the effort choice of the developer as given, and the expected payoff from
an experiment is a function of the posterior belief of the developer following $y$, denoted by $\hat{p}(y)$. Payoffs are therefore:

$$p_0 \mathbb{E}[\lambda e(\hat{p}(y)) \mid \theta = T] + \mathbb{E}[g(\hat{p}(y))] - c_S(a)$$  \hspace{1cm} (1)

The following Lemma rewrites the payoffs of the scientist without conditioning on the state. It shows that we can changes the prior belief $p_0$ in (1) into a posterior belief by removing the conditioning event $\theta = T$:

**Lemma 1.** In any pure strategy equilibrium where the experiment is correctly inferred as $\mathcal{I}(a)$, the scientist’s payoffs can be written as:

$$\mathbb{E}_{y \sim \mathcal{I}(a)}[\lambda \hat{p}(y)e(\hat{p}(y)) + g(\hat{p}(y))] - c_S(a)$$  \hspace{1cm} (2)

This lemma is similar in spirit to many arguments that have utilized the belief-based approach in the persuasion literature (and is not particularly complicated), though we are not aware of (2)\textsuperscript{12} having been explicitly stated or utilized directly. While Kamenica and Gentzkow (2011) do allow for state dependence on the Sender’s utility function, their characterization of the Sender’s value function does not require them to explicitly state Receiver’s preferences, meaning this lemma would have limited use for their exercise. However, see Section 4 of their paper for a discussion of how preference alignment influences the solution.

The Lemma can intuitively be thought of as “flipping the order of integration”; rather than first taking an expectation over the signal realization $y$ and then the state realization $\theta$, it takes the expectation of the state first, and signals second. That said, the result relies upon the developer and scientist updating their beliefs in the same way in response to the experimental outcome. Equation (2) also clarifies that $\lambda$ generates preference for information acquisition. It is well understood from the persuasion literature that convexity of payoffs in beliefs can generate incentives for information acquisition, but it is hard to see how to apply this intuition directly from looking at equation (1). In contrast, the Lemma implies shows that the term driving the scientist’s added incentive for information acquisition from the developer’s actions is given by the term $\hat{p}e(\hat{p})$, meaning there are additional incentives for information acquisition whenever this term is concave.

The lemma allows us to adopt the belief-based approach in cases where the experiment choice is observed by the developer, but does not do so under different transparency regimes. In order to accomplish that task, we need to ensure that the experiment choice is deterministic and unique. If the scientist were randomizing actions, then this would imply that the scientist’s beliefs following any given outcome would not coincide with the developer’s, and hence we could not use a

\textsuperscript{12}The appendix uses a more general identity, which is also novel to our knowledge.
Figure 2: Graphical illustration of Lemma 1. All graphs express the scientist’s payoff as a function of the developer’s belief. The top row considers the model where $\lambda = 0$, and the bottom row considers the model when $\lambda > 0$, where we normalize the ex ante payoff when $\hat{p} = 1$ to 1 for both cases. The left column displays payoff of the scientist from the ex-ante perspective, taking an expectation over the state. The right column writes the scientist’s payoff conditional on each state.
single \( \hat{p}(y) \). In principle uniqueness could be dispensed of if we were willing to make a selection argument (as in Kamenica-Gentzkow (2011) or Lipnowski-Ravid (2017)), although given our focus on developer (receiver)-optimality it is not clear that these selection arguments would be desireable. However, using the following lemma, it follows that we can indeed adopt the belief-based approach in this setting:

**Lemma 2.** Suppose that \( c(a_M, a_{-M}) \) is weakly convex in \( a_{-M} \) for all \( a_M \), and furthermore, that \( h_T(a_M, a_{-M}), h_F(a_M, a_{-M}) \) are weakly concave in \( a_{-M} \) for all \( a_M \).

(1) A Perfect Bayesian equilibrium in pure strategies exists when \( a_{-M} \) is unobserved. In this equilibrium, if \( a_i \) is interior and unobserved, we have:

\[
\frac{\partial c_S(a)}{\partial a_i} = \lambda \cdot (e(\hat{p}(1)) - e(\hat{p}(0))) p_0 \frac{\partial h_T}{\partial a_i} + \left( g(\hat{p}(1)) - g(\hat{p}(0)) \right) \left( \frac{\partial h_T}{\partial a_i} + (1 - p_0) \frac{\partial h_F}{\partial a_i} \right).
\]

(3)

If \( a_i \) is observed, the first order condition corresponds to the derivative of (2) with respect to \( a_i \).

(2) If either convexity of \( c \) or concavity of \( h_T \) is strict, there is no equilibrium in mixed strategies.

While the proof of this lemma is mostly straightforward, one assumption that turns out to be important is that \( h_T(a) > h_F(a) \) for all \( a \). This ensures that \( \mathbb{P}[y = 1] \) and \( \mathbb{P}[y = 0] \) are both interior, meaning that the receiver always places positive weight on observing any given signal. This is in contrast to other papers in the persuasion literature, which often do not restrict the set of signals that a receiver may observe a priori.

The lemma indicates that the scientist’s equilibrium behavior can be thought of as equating marginal benefits and marginal costs along each dimension. While the former simply depends on \( c_S(a) \), the marginal benefit is the sum of the gain due to the career concerns term as well as due to the change in the developer’s effort. While such a condition is intuitive, the complication is that the marginal benefit depends on beliefs, which are endogenous. Additionally, as mentioned above, the intuition requires the validity of the belief-based approach.

Another implication of the lemma is a subtle (and potentially empirically relevant) difference between the scientist’s career concerns and investment in developer outcome. Lemma 1 implies that under observable behavior, the case of \( \lambda = 0 \) and \( g(p) = \beta p e(p) \) results in an identical preference over experiments and hence identical experiment choice compared to \( \lambda = \beta \) and \( g(p) = 0 \). However, this does not hold when dimension \( a_i \) is unobservable; in that case, the false

---

13In Section 5.1, we show that mixed strategies will generally arise with the presence of certain kinds of private information for the scientist.

14Admittedly this specification rules out the discrete example from Section 3.1, although we find it easier to parameterize the experiments more generally by resorting to the specification of the main model. Furthermore, the example is useful for the purposes of intuition in our general specification.
positive rate $h_F(a_M, a_{-M})$ does matter for the career concerns payoff, but does not matter for the developer outcome payoff. We state this as follows:

**Corollary 1.** When $g(p) = 0$, the false positive rate does not influence the experiment choice when action $a_i$ is observed.

Finally, we note that the lemma illustrates the difference between the first order condition as transparency changes. When an action changes from being unobserved to being observed (by the developer), a term equal to:

$$E\left[\left(\lambda \hat{p}(y)e'\left(\hat{p}(y)\right) + g'\left(\hat{p}(y)\right)\right)\frac{\partial \hat{p}(y)}{\partial a_i}\right]$$

is added to the right hand side of (3). This term can be positive or negative, even when information structures are monotonic in $a_i$ in the Blackwell order; in that case $\frac{\partial \hat{p}(1)}{\partial a_i}$ and $\frac{\partial \hat{p}(0)}{\partial a_i}$ will have opposite signs, and hence (4) would be convex sum of a negative term and a positive term. Since we can find Blackwell ordered information structures that also hold $\hat{p}(1)$ and $\hat{p}(0)$ constant, it follows this additional term can either be positive or negative.

**Corollary 2.** A sufficient condition to ensure that the scientist’s payoffs is $c''''(e) \geq 0$, or if $c_R(e) = e^n/k$ for any $n > 1$ and $g(p)$ is not too concave. If action $i$ is distortive, then $a_i = a_i$ in equilibrium whenever it is observed, provided $g(p)$ is not too concave.

To summarize: the incentives for information acquisition arise due to the state dependence of the scientist’s payoff as a function of the developer’s posterior, since the convexity of this line is what generates incentives for information acquisition. But the loss of credibility occurs due to the positive slope of the scientist’s payoff conditional on $\theta$—that is, because the payoff is still higher when the developer’s belief is higher. By developing this model and comparing the influence of $\lambda_p$ and $\lambda_x$, we have also shown why the forces highlighted are distinct from others that have been proposed, most notably in the career concerns literature.

### 3.3. Optimal Transparency

Having characterized scientist behavior as a function of the transparency requirements, we are now in a position to present results on the optimality of different transparency regimes. First, we point out that greater transparency is always better for the scientist; if an action $a$ arises in the equilibrium where only some subset of the coordinates are observable, then the scientist could always guarantee this outcome when $a$ is totally observable, and could in fact achieve a higher payoff potentially via some other action. This argument proves that:

**Proposition 2.** Full transparency always achieves the scientist-optimal payoff.
Our main result is that, under the payoff specifications of this setting, there generally exist payoff specifications such that full transparency is not optimal.

**Theorem 1.** Suppose \( c''_R(e) \leq 0 \) or \( c_R = e^n/k, \lambda > 0 \) and \( g(p) \) is not too concave. Let \( J \) denote the indices that are informative and \( N \setminus J \) denote the indices that are distortive, and let \( K \subset N \setminus J \). Then there exists a cost function (increasing in all coordinates) such that the developer does better when \( a_i \) is not observed for all \( i \in K \).

The theorem is proved by constructing cost functions inspired by the example in Section 3.1. We construct cost functions with the property that distortion is “cheap” for uninformative experiments but “expensive” for informative ones. The conditions of the theorem ensure that the scientist does obtain a higher payoff from conducting experiments that are more informative. Because of this, it is costly for them to be perceived as adding bias. The cost functions constructed have the property that higher actions along the informative dimensions make biasing more costly. For example, it is hard to run an experiment many times if collecting a new data set each time is difficult. While other incentives may be present, we view this result as adding an important, subtle caution to the debate on transparency requirements.

### 4. POLICY ANALYSIS

#### 4.1. Costless Communication

From the theoretical perspective, the main novelty of the model is in its use of communication costs as well as limited commitment by the sender (scientist). If any experiment is feasible and costs are not present, then the model reduces to Bayesian Persuasion (as in Kamenica and Gentzkow (2011)) when \( a \) is observable,\(^{15}\) and reduces to cheap talk (as in Crawford and Sobel (1982) or Lipnowski and Ravid (2017)) in the case where \( a \) is not observable at all.\(^{16}\) In this sense, the model provides an “intermediate commitment” benchmark.

In fact, we can show that without costs, the developer’s payoff is increasing in the level of scientist commitment:

**Proposition 3.** If the scientist’s cost for all experiments are known to be zero, then making any biasing dimension unobservable strictly lowers the developer’s payoffs.

---

\(^{15}\)Ichihashi (2017) studies the case of Bayesian Persuasion when the sender’s choice set can be limited.

\(^{16}\)Technically speaking, Kamenica and Gentzkow (2011) and Lipnowski and Ravid (2017) do not restrict to \( y \in \{0, 1\} \), although it follows from their results in this setting that the sender would not benefit from a richer signal space since the state is binary.
4.2. Career Concerns versus Follow-on Interest

Recall the earlier point (highlighted in Corollary 1) that even when career concerns replicate the preferences over experiments as interest in follow-on outcomes, there may still be differences in how the scientist responds to transparency changes. We use this insight to generate testable predictions of the model, in terms of the comparative statics of how researchers with the same access to experiments but different weight placed on immediate outcomes (as opposed to their career outcomes) would respond to transparency changes differently.

Our first observation is that, unless the payoffs derived from career concerns are strictly convex, there is no benefit to making biasing actions unobservable. This is the case since the developer has no incentive for information acquisition but does have incentives for adding bias the case; we immediately obtain.

**Proposition 4.** Let \( \lambda = 0 \) and \( g''(p) \leq 0 \). Then for any cost function, full transparency is developer-optimal.

Whether career concerns incentives should be convex or concave in general seems difficult to answer a priori. Risk aversion over long-term career outcomes would suggest concavity is appropriate, though high-power incentives for “superstar” researchers would generate convexity.

To highlight the differences between career concerns (given either convexity or concavity) and follow-on interest, we focus on the case where biasing actions do not increase the true positive rate, but only increase the false positive rate. In this case, it similarly follows immediately from Corollary 1 that:

**Proposition 5.** Suppose \( \frac{\partial h}{\partial a_i} = 0 \) for any biasing action \( a_i \). Then partial non-transparency only results in more informative experiments suppose \( g''(p) > 0 \).

These comparisons are significant since it is reasonable to assert that the significance of follow-on research versus the beliefs in the validity of hypotheses would vary according to the stage of the career of the scientist. For younger researchers, follow-on research may be less significant than the belief in the validity of the hypothesis. For older researchers, legacy may be more significant, in which case the importance of generating follow-on research would matter more. In this sense, the model suggests that transparency requirements would have different impacts across these different researchers. Policymakers may consider, for example, imposing transparency requirements on early stage researchers, but not late stage researcher, if they believed that most scientists were risk-averse over long term outcomes.
Figure 3: Graphical explanation for mixed strategies in Section 5.1. The points $C$ and $E$ represent expected payoffs if distortions are feasible (hence to the right of the prior), and points $B$ and $D$ represent expected payoffs if distortions are infeasible (hence to the left of the prior). If the cost of the fully informative experiment moves the payoff from $A$ to a point lower than $C$ but higher than $E$, then the equilibrium behavior will involve mixed strategies.

5. EXTENSIONS

5.1. Private Information on Distortability

So far, the set of experiments available to the researcher has been taken to be common knowledge. This is a sensible assumption if, for example, the set of possible experiments is well-understood and could be characterized in advance. On the other hand, if the scientist has specialized knowledge about the experiment in the first place, then one may also be interested in what would happen in case this assumption were relaxed.

In this section, we illustrate that the presence of this kind of private information may result in the scientist’s equilibrium behavior involving mixed strategies. In order to minimize notation, we demonstrate this in the context of the example from Section 3.1, rather than the general model, and assert that similar intuition applies for that setting as well. Recall that in this setting, the scientist can choose a perfectly informative experiment at a cost, or can choose an imperfectly informative experiment for free (but is susceptible to bias).

Suppose instead that the scientist is only able to add positive bias to experiment with $a_1 = 0$ (low quality equipment) with probability $t$. This situation is illustrated in Figure 3, focusing again on the case with costless choice of $a_2$ (research assistant quality, the biasing dimension). There are two features which distinguish this version from the previous analysis. First, under
non-transparency, the developer’s beliefs need not form a martingale from the perspective of the scientist, when the experiment \( a_1 = 0 \) is chosen—that is, for the scientist, \( \mathbb{E}[\hat{p}(y)] \neq p_0 \). To see this, note that with probability \( t \), the scientist chooses \( a_2 = 1 \) when choosing \( a_1 = 0 \), and with probability \( 1 - t \) the scientist is forced to set \( a_1 = 0 \). Hence the probability that \( y = 1 \) is larger for the scientist who is allowed to bias than it is for the scientist who cannot. But the developer cannot observe whether the scientist is able to bias, and his beliefs are a martingale from his perspective. The result is that if the scientist can pick \( a_2 > 0 \), then \( \mathbb{E}[\hat{p}(y)] > p_0 \), but if the scientist cannot, then \( \mathbb{E}[\hat{p}(y)] < p_0 \).\(^{17}\)

To see the second difference, notice that when considering the informativeness of an experiment, we were able to treat the experiment choice as given and then ask which level of distortions would be picked. In this case, however, the informativeness of the signal \( a_1 = 0 \) will depend on whether a player who is able to distort would prefer to choose \( a_1 = 1 \) or \( a_1 = 0 \). If the scientist chooses the proposal \( a_1 = 0 \) when distortions are available, then the experiment is less informative than it would be if she were to choose \( a_1 = 1 \) when distortions are available. When \( t = 1 \), these two cases are the same, but they are not in general.

This observation explains why we may have difficulty with ensuring the existence of a pure strategy equilibrium. Indeed, for certain values of \( t \), the scientist who can choose \( a_1 = 0 \) and bias might mix between \( a_2 = 1 \) and \( a_1 = 0 = 1 - a_2 \). Indeed, it may be the case that, without mixing, if the developer thought the scientist would always choose \( a_1 = 1 \) if available, the scientist would prefer to choose \( a_1 = 0 \) and set \( a_2 = 1 \), whereas if the developer thought the scientist would only choose \( a_1 = 1 \) if distortions were impossible, then the scientist would always prefer to choose \( a_1 = 0 \). We remark that several other papers consider principal agent problems where the agent follows a mixed strategy due to a lack of commitment at the time of contracting; see, for example, Fudenberg and Tirole (1990). Still, to the best of our knowledge, we believe the mechanism isolated here for mixed strategies in this setting is new.

We briefly comment that other forms of scientist private information could be prevalent, and multiplicity may be an issue. For example, if the scientist has private information on \( \theta \), the setting becomes a signalling game and the choice of experiment has an additional impact of conveying this private type. The complications arising with these settings, while interesting, are left to future work.

### 5.2. Contractibility

There are many possible interpretations of the benefits to the scientist, and in this paper we prefer an interpretation where these benefits are non-monetary. For example, if drug development is

\(^{17}\)We again emphasize again that \( \hat{p}(y) \) is the probability the developer, and not the scientist, assigns to the event that \( \theta = T \).
successful, other researchers may view this work as a valuable contribution, and follow-on by working on similar problems. These kinds of payoffs are not easily contractible, and we believe this is a good approximation to many kinds of research activity, particularly research that is too costly or long term to be undertaken by private entities. On the other hand, for research where the follow-on work is very quick, it may be more feasible to introduce transfer as a function of the chosen experiment. We point out that some kinds of contractibility would make the model degenerate. For instance, if the scientist’s experiments could be restricted a priori, then there would be no benefits from changing the transparency regime.\footnote{This would be difficult if no third party could determine the set of feasible experiments before they were chosen, as in the previous extension.} Alternatively, if the developer could commit to an effort profile, they could set $e = 0$ unless a given experiment were chosen. Without career concerns, under this policy, the scientist could only possibly obtain non-negative payoff from choosing that experiment. In that case, any experiment that is individually rational for the scientist could be implemented.

However, these comments do not imply that our insights are sensitive to an extreme lack of contractibility. Suppose, for instance, it were possible to pay the scientist an additional amount if they undertake some given experiment, say $I(\bar{\pi})$, which is preferred by the developer to the experiment the scientist would undertake with non-transparency. In this case, the scientist would be compelled to undertake $I(\bar{\pi})$ if payment for doing so was greater than the difference in the scientist’s payoff. Ultimately, the conclusion of Theorem 1 remains valid, even with transfers (though there are cases under which transfers with transparency outperforms non-transparency), since the cost of compelling the scientist to make this change may be larger than the gain that the developer obtains from the more informative experiment.

6. CONCLUSION

6.1. Discussion of Model Assumptions

The abstraction of the model is meant to focus on intuition, rather than explicitly mapping to a given setting. Still, one may wonder whether the assumptions of the model make sense. Perhaps most important is the assumption that the transparency regime is set by an outsider over whom the scientist has no control.\footnote{We mentioned this possibility in Section 3.1.} It could be that the only way of documenting experimental methods is through a pre-registration database, and that the scientist is unable to credibly do this on their own. Why this does appear to be the case in a variety of disciplines is beyond the scope of this paper. It is not difficult to imagine, though, that without a formal mechanism for doing so, credibly demonstrating that the protocols followed were scrupulous might impose prohibitive costs on the
scientist.

Our specification that the scientist’s benefit comes from the follow-on work of a developer is simply to provide a concrete story within which our results can be presented. We could have alternatively just been concerned about informativeness, and taken as a reduced form the societal objective. In other settings, it may make more sense to think about follow-on researchers extending or elaborating on the scientist’s results (with further citations). In that case, one may be interested in modeling the citation process explicitly or describing the “feedback process” through which the scientist, in turn, learns from follow-on researchers. Our analysis is relevant to these settings, but accommodating them explicitly is left to future work.

We imposed some technical assumptions on the space of information structures (e.g., continuity) to ensure equilibrium existence. We also imposed parameterizations on the set of information structures (and restrictions on costs) in order to allow for limited commitment while still being able to interpret the results. While these assumptions are restrictive, we believe that the model formalizes the basic intuition represented in the introduction. Hence the model seems to be an appropriate formalization of the debate over transparency referenced in the introduction. From that perspective, this paper echoes Wilson (2014), on the importance of careful economic reasoning in the design of research environments.

6.2. Final Comments

This paper has shown why a sender’s inability to fully commit to an information structure may make a receiver better off. While we illustrated the main forces at work through a simple example, our general model clarified that complementarity (which is natural, at least in some settings) between the different kinds of research actions drove the result. As an application, we have considered whether transparency requirements should be more widespread in academic disciplines. The key insight is that non-transparency on one dimension can induce scientists to exert more effort or incur more costs on another dimension, in a way that ultimately makes those interested in the results better off. In assessing this conclusion, we have primarily been concerned with the interest of those who use the results downstream. This is, of course, a simplified view of welfare, but it is motivated by the idea that society invests in scientific research so that it produces results which will be useful for others downstream.

In the main model we assumed that the only policy lever under the designer’s control is whether the developer observes the scientist’s choice of protocols. For example, a funding agency may be able to implement pre-registration requirements, but might not have the authority to dictate specific steps that scientists follow. In the extensions we considered what might happen if transfers were also available at their disposal, and demonstrated that the main conclusion would not be changed. Still, the main model highlighted the relevant tradeoffs of transparency. On
the other hand, transparency also might have an advantage of allowing for richer punishments, something we do not consider here, though they are natural places for future work.

Though our model is theoretical, it still allows us to comment on the active debate on the costs and benefits of transparency mentioned in the introduction. Our main model has highlighted that this debate should consider the extent to which follow-on research influences which experiments scientist choose to perform. In cases where it is significant, our model shows that scientists have a natural incentive to both add informational content to their experiments and bias their experiments. In this case, non-transparency encourages scientists to choose experiments which are inherently harder to bias. If these experiments are particularly difficult as well, then scientists may not be sufficient motivated to undertake them under transparency, but if they are more informative then presumably society is better off if they do.

Finally, we remark that the intuition for many of our results would apply to the research process more generally, and not just academic publication. While we called the person who acts after the scientist the developer, in general this could be any individual who will use the scientist’s research, and whose actions the scientist would be interested in influencing. We have shown that in general, a sender might be incentivized to acquire information but not distort a signal, or conversely. We provided a framework which describes when each will happen, so long as there is some sense in which these two types of actions can be distinguished. The observation that non-transparency can be used as a mechanism for money burning will likely have applications beyond the focus of this paper, as these concerns are similar to those that have been raised in a wide variety of contexts in organizational economics.

References


7. APPENDIX

7.1. Omitted Proofs

Proof of Proposition 1. Since the developer’s cost function is convex, the developer’s payoff function is a Fenchel transformation of a convex function and is hence convex itself. It follows that \( \hat{p}(y) \pi_R(1) - \pi_R(\hat{p}) \) is a measure of uncertainty according to Ely, Frankel and Kamenica (2015), since it is 0 at degenerate beliefs and concave for all interior beliefs (since it is a linear function minus a convex function), as desired.

Proof of Lemma 1. We find it useful to demonstrate the following, more general equation, for arbitrary function \( t(p) \) and \( f(p) \), noting that Lemma 1 follows immediately by setting \( t(p) = \lambda e(p) + g(p) \) and \( f(p) = g(p) \): For functions \( t, f : [0, 1] \to \mathbb{R} \),

\[
p_0 \mathbb{E}_y [t(\hat{p}(y)) \mid T] + (1 - p_0) \mathbb{E}_y [f(\hat{p}(y)) \mid F] = \mathbb{E}_y [\hat{p}(y)t(\hat{p}(y)) + (1 - \hat{p}(y))f(\hat{p}(y))] \tag{5}
\]

By writing out the definition of the conditional expectation, we have:

\[
p_0 \mathbb{E}_y [t(\hat{p}(y)) \mid T] = p_0 \sum_{y \in Y} t(\hat{p}(y)) \mathbb{P}[y \mid T]
\]
\[
= \sum_{y \in Y} t(\hat{p}(y)) \left( p_0 \frac{\mathbb{P}[y \mid T]}{\mathbb{P}[y]} \right) \mathbb{P}[y]
\]
\[
= \sum_{y \in Y} t(\hat{p}(y)) \hat{p}(y) \mathbb{P}[y] = \mathbb{E}_y [\hat{p}(y)t(\hat{p}(y))].
\]
An almost identical argument can be used for the other term in (5). In fact, this generalizes for any number of states, not just \( \theta \in \{ T, F \} \).

**Proof of Lemma 2.** Proof of (1) Suppose \( M \) is the index set of observable indices, and partition the scientist’s action into \( a = (a_M, a_{-M}) \). We show that there is some \( a^n_{-M} \) such that when the developer conjectures that \( a^n_{-M} \) are the unobserved actions of the scientist, the scientist’s best response is to follow action \( a^n_{-M} \). Since \( p_0 \) is interior and \( h_T(a) > h_F(a) \) for all \( a \), the developer always puts non-negative probability on observing \( y = 0 \) or \( y = 1 \), for any conjecture regarding the scientist’s behavior. Therefore, there are unique beliefs \((\hat{p}(0), \hat{p}(1))\) formed after observing a signal \( y = 0 \) or \( y = 1 \), respectively, for any equilibrium strategy of the scientist. In fact, given a conjecture, since \( A \) is compact, we have that \( \mathbb{P}[y = 0] \) and \( \mathbb{P}[y = 1] \) are both bounded away from 0. This implies that beliefs are a continuous function of actions, and well-defined given any conjecture.

Let \( t(p) = \lambda e(p) + g(p) \) and \( f(p) = g(p) \), and define the function \( \phi \) as follows:

\[
\phi(a_{-M}) = \arg \max_{a_{-M} \in A_{-M}} p_0[t(\hat{p}(a_M, a_{-M})(0)) + h_T(a_M, \tilde{a}_{-M})(t(\hat{p}(a_M, a_{-M})(1)) - t(\hat{p}(a_M, a_{-M})(0)))]
+ (1 - p_0)[f(\hat{p}(a_M, a_{-M})(0)) + h_F(a_M, \tilde{a}_{-M})(f(\hat{p}(a_M, a_{-M})(1)) - f(\hat{p}(a_M, a_{-M})(0)))] - c(a_M, a_{-M})
\]

Note that \( \phi(a_{-M}) \) gives the payoff maximizing response, assuming (observable) actions \( a_M \) are chosen and a conjecture of \( a_{-M} \). Taking \( a^n_{-M} \to a_{-M} \), and \( b_n \in \phi(a_{-M}) \) with \( b_n \to b \), since beliefs are continuous in \( a \) and \( f(p), t(p) \) are continuous as well (by continuity of \( e(p) \)), we have

\[
t(\hat{p}(a_M, a^n_{-M})(1)) - t(\hat{p}(a_M, a_{-M})(0)) \to t(\hat{p}(a_M, a_{-M})(1)) - t(\hat{p}(a_M, a_{-M})(0)),
\]

and similarly for \( f \). If \( b \notin \phi(a_{-M}) \), then there exists some value \( \delta \) such that:

\[
p_0(h_T(a_M, \delta) - h_T(a_M, b))(t(\hat{p}(a_M, a_{-M})(1)) - t(\hat{p}(a_M, a_{-M})(0)))
+ (1 - p_0)(h_F(a_M, \delta) - h_F(a_M, b))(f(\hat{p}(a_M, a_{-M})(1)) - f(\hat{p}(a_M, a_{-M})(0))) > c(a_M, \delta) - c(a_M, b).
\]

But since \( a^n_{-M} \to a_{-M} \) and \( b_n \to b \), by continuity we can find some \( n \) sufficiently large such that

\[
p_0(h_T(a_M, \delta) - h_T(a_M, b_n))(t(\hat{p}(a_M, a^n_{-M})(1)) - t(\hat{p}(a_M, a^n_{-M})(0)))
+ (1 - p_0)(h_F(a_M, \delta) - h_F(a_M, b_n))(f(\hat{p}(a_M, a^n_{-M})(1)) - f(\hat{p}(a_M, a^n_{-M})(0))) > c(a_M, \delta) - c(a_M, b_N),
\]

contradicting that \( b_n \) is a maximizer of \( \phi(a^n_{-M}) \). Hence the map \( \phi \) is upper-hemicontinuous.

Furthermore, \( \phi(a_{-M}) \) is nonempty and closed because \( A_{-M} \) is compact (being the product of intervals) and the objective function in the expression for \( \phi(a_{-M}) \) is continuous. Finally, to
see that it is convex, notice that if \( a'_{-M}, a''_{-M} \) are both in \( \phi(a_{-M}) \) the convexity of \( c_S(a) \) and the concavity of \( h_T(a), h_F(a) \) means that we must have that

\[
p_0 h_T(\bar{a}_{-M}) (t(\hat{p}(a_M, a_{-M})(1)) - t(\hat{p}(a_M, a_{-M})(0))) + (1 - p_0) h_F(a_M, \bar{a}_{-M}) (f(\hat{p}(a_M, a_{-M})(1)) - f(\hat{p}(a_M, a_{-M})(0))) - c_\alpha(\bar{a}_{-M}).
\]

is constant for all \( \bar{a}_{-M} \) with \( \bar{a}_{-M} = \alpha a'_{-M} + (1 - \alpha) a''_{-M} \) with \( \alpha \in [0, 1] \), so that \( \phi(a_{-M}) \) is convex. Hence by Kakutani’s fixed point theorem, an equilibrium exists when \( a_M \) is observed, for any choice of \( a_M \).

That the first order condition is given by Equation (3) follows from observing that \( \hat{p}(a_M, a_{-M}) \) does not respond to the choice of \( a_i \) for \( i \notin M \). In that case, we see that the marginal benefit is given by:

\[
p_0 \frac{\partial h_T(a_M, a_{-M})}{\partial a_i} (t(\hat{p}(a_M, a_{-M})(1)) - t(\hat{p}(a_M, a_{-M})(0))) + (1 - p_0) \frac{\partial h_F(a_M, a_{-M})}{\partial a_i} (f(\hat{p}(a_M, a_{-M})(1)) - f(\hat{p}(a_M, a_{-M})(0))),
\]

which reduces to the stated condition when equal to marginal cost, as desired.

**Proof of (2)** Suppose to the contrary that there is a mixed strategy equilibrium. Then the first order condition in equation (3) must hold for two values of \( a_{-M} \), say \( a^1_{-M} < a^2_{-M} \). On the other hand, the developer’s beliefs do not depend on the choice of \( a_{-M} \). Keeping the notation from the previous we have:

\[
\nabla_{a_{-M}} c(a_M, a^i_{-M}) = p_0 (t(\hat{p}(1)) - t(\hat{p}(0))) \nabla_{a_{-M}} h_T(a_M, a^i_{-M}) + (1 - p_0) (f(\hat{p}(1)) - f(\hat{p}(0))) \nabla_{a_{-M}} h_F(a_M, a^i_{-M}),
\]

and hence subtracting the equation for \( i = 1 \) from the equation for \( i = 2 \), and taking the dot product for some arbitrary \( \alpha \) with \( ||\alpha|| = 1 \), we have:

\[
\alpha \cdot \nabla_{a_{-M}} (c(a_M, a^2_{-M}) - c(a_M, a^1_{-M})) = p_0 (t(\hat{p}(1)) - t(\hat{p}(0))) \alpha \cdot \nabla_{a_{-M}} (h_T(a_M, a^2_{-M}) - h_T(a_M, a^1_{-M})) + (1 - p_0) (f(\hat{p}(1)) - f(\hat{p}(0))) \alpha \cdot \nabla_{a_{-M}} (h_F(a_M, a^2_{-M}) - h_F(a_M, a^1_{-M})),
\]

By the multivariate mean value theorem, applied to \( h_T, h_F \) and \( c \), we have, for some \( a_t, a_f \) and \( a_c \) which are convex combinations of \( a^1_{-M} \) and \( a^2_{-M} \) such that:

\[
\alpha \cdot \nabla_{a_{-M}}^2 c(a_M, a_t) (a^2_{-M} - a^1_{-M}) = p_0 (t(\hat{p}(1)) - t(\hat{p}(0))) \alpha \cdot \nabla_{a_{-M}}^2 h_T(a_M, a_t) (a^2_{-M} - a^1_{-M}) + (1 - p_0) (f(\hat{p}(1)) - f(\hat{p}(0))) \alpha \cdot \nabla_{a_{-M}}^2 h_F(a_M, a_f) (a^2_{-M} - a^1_{-M}).
\]
But since either $h_T$ or $h_F$ is strictly concave or $c$ is strictly convex, either the left hand side is strictly positive or the right hand side is strictly negative, with both being at least weakly so, a contradiction. Hence in equilibrium, there can only be pure strategies. \qed

Proof of Corollary 2. This follows from demonstrating that the scientist’s payoffs, as a function of the developer’s beliefs, are convex. This is immediate in the case of polynomial effort costs. Indeed, we take the second derivative of $pe(p)$ (the benefit from follow-on effort) and observe that it is equal to:

$$
\lambda(2e'(p) + e''(p)).
$$

Since $c_R(e)$ is strictly convex, $e'(p) > 0$, since the first order condition is:

$$
bp = c'_R(e(p)),
$$

and differentiating with respect to $p$ gives:

$$
b = c''_R(e(p))e'(p),
$$

and differentiating again gives:

$$
0 = c''_R(e'(p))^2 + c''_R(e)e''(p).
$$

Since $e(p)$ is strictly increasing, the assumptions on $c''_R(e)$ ensure that $e''(p) \geq 0$, and hence the objective is convex. \qed

In general, convexity of developer effort by itself is not a strong enough assumption in order to ensure that $pe(p)$ is convex. To see this, suppose that:

$$
c_R(e) = 1 - \sqrt{1 - e} \Rightarrow c'_R(e) = \frac{1}{2\sqrt{1 - e}} > 0 \Rightarrow c''_R(e) = \frac{1}{4(1 - e)^{3/2}} > 0.
$$

In that case, we have:

$$
e(p) = \max\{0, -\frac{1}{4p^2} + 1\},
$$

and we observe that $pe(p)$ is concave whenever $e(p) > 0$.

Proof of Theorem 1. Consider the following family of cost functions:

$$
c_{\eta, \gamma, \tilde{a}}(a) = \eta \sum_{j \in J} (a_j - \tilde{a}_j)^2 + \gamma \sum_{k \in K} (a_k - \tilde{a}_k)^2 (a_j - \tilde{a}_j)1[a_j \geq \tilde{a}_j],
$$
for $\eta, \gamma > 0$ and $\tilde{a}_j > a_j$. Note that these are strictly convex in $a_K$ for all parameters, since it is a quadratic function when we treat $a_j$ as a constant. By Corollary 2, when distortive action $a_k$ is observed, it is set equal to $a_k$ in equilibrium. Let $\eta^*$ denote the smallest value of $\eta$ such that $a_j = \tilde{a}_j$ for all informative actions. Since all distortive actions are set equal to $a_k$ under complete observability, $\eta^*$ does not depend on the choice of $\gamma$ or $\tilde{a}$.

Now suppose that the $K$ coordinates are unobserved. Note that for all $k \in K$, we have:

$$\frac{\partial c_{\eta, \gamma}}{\partial a_k} = 2\gamma(a_k - \tilde{a}_k) + \sum_{j \in J} (a_j - \tilde{a}_j)1[a_j \geq \tilde{a}_j].$$

Note that at $a_i = a_i$ for all actions $i$ (as is the scientist’s choice under observability, by construction), the first order condition in (3) cannot be satisfied for any $\gamma > 0$. It follows that some distortive dimension $a_k$ must be in the interior (or possibly at $\tilde{a}_k$) whenever $\gamma$ is positive.

Now let $\gamma \to \infty$ and set $\tilde{a}_j = a_j + \epsilon$. We claim that for all $j$, $a_j \to a^* \geq \tilde{a}_j > a_j$ (passing to a subsequence if necessary, we can ensured that the limit always exists since the action space is compact) when $\epsilon$ is sufficiently small. Suppose otherwise; in that case, $a_k \to \tilde{a}_k$ by Lemma 2, since the marginal cost is 0. It follows that the loss to the scientist due to distortion does not vanish as $\gamma \to \infty$, for every $a_k \in [a_k, a_k + \epsilon]$ and for any $\epsilon$. On the other hand, by taking $\epsilon$ sufficiently small, we can ensure that the cost of raising the action for $a_j$ to $\tilde{a}_j$ is negligible, in which case the scientist would find it to be an equilibrium to choose $a_k = \tilde{a}_k$. It follows that there exists a pair $(\gamma, \epsilon)$ such that making the $K$ coordinates unobserved results in a more informative experiment chosen than under full transparency.

Proof of Proposition 3. Without costs, then under the restrictions of preferences studied in this paper, the scientist’s payoff is increasing in the informativeness of the experiment. Hence the scientist chooses the maximally informative experiment under observability. On the other hand, making any biasing dimension unobservable results in $a_i = \pi_i$ being chosen by the scientist in equilibrium, by Lemma 2. The result follows from observing that by definition, the most informative experiment under $a_{-M} = a_{-M}$ is less informative than the most informative experiment under $a_{-M} = \tilde{a}_{-M}$.

7.2. Scientist Preferences over $y$

In this appendix, we comment on a modification to the model where we allow for the scientist to have preferences over $y$ itself. While this could take several different forms in general, for simplicity we will comment on the case where the payoffs are separable, and the scientist obtains an added benefit of $\lambda_y \cdot y$ from a positive result.

In general this does not interfere with our application of the belief-based approach, noting that
any positive result leads to a higher belief and any negative result leads to a lower belief. Hence this setting is as if there were a jump in the scientist’s payoff function at the prior. That said, it is simplest to comment on this case simply by inspection. In this case, it is immediate that the scientist is incentivized to maximize the biasing action in this case (whether higher informative actions will be taken depends on the prior):

**Proposition 6.** There exists $\overline{\lambda}_y$ such that when $\lambda_y > \overline{\lambda}_y$, biasing actions are all maximized in equilibrium.

More to the point of the paper, however, by itself transparency does not interact with the experiment choice when these kinds of considerations are dominant:

**Proposition 7.** Suppose $\lambda = g(p) = 0$ and $\lambda_y > 0$. Then the scientist’s experiment choice does not differ depending on transparency or not.

This is immediate since the prior does not depend on experiment choices in this setting.

While immediate, the result may be counterintuitive given that it is natural to feel inclined to prefer positive results rather than negative results. One may be that negative results may be harder to publish than positive results, in which case there should actually be an interaction between positive results and the other payoff terms. One could accommodate this and obtain similar results, as this is similar to imposing even greater convexity in the payoffs. Alternatively, it may be that positive results that are obtained “cheaply” (via bias) are less meaningful, but those that are achieved “scrupulously” (via informativeness) are more meaningful. This would suggest greater interdependence between the cost function and the benefit than what we have here. Since this would take us too far afield, we do not pursue such specifications further.