Problems and Prescriptions in Psychiatric Explanation

A normative analysis of explanatory framing effects

Sera Schwarz

Yale University
sera.schwarz@yale.edu

Abstract: A growing body of psychological research suggests that different kinds of explanations of mental illness can have striking and distinctive effects on their audiences' attitudes and inferences. But it is surprisingly difficult to account for why this is. In this paper, I present a "normative model" of explanatory framing effects, which I claim does a better job of capturing the empirical data than do intuitive alternatives. On this model, different explanations will tend to differently affect their audience's reasoning because each encodes a different picture of the kind of *problem* represented by the explanandum, and therefore the kinds of responses to it that are normatively apt to pursue. For example, a biological explanation of depression will convey to its audience that depression is a specifically biological problem, and therefore that appropriate responses to it should be directed at biological facts and norms. The communication of this normative information is, I argue, importantly different from communicating that depression has biological causes. For example, although it seems plausible that most causal explanations can be viewed additively, different characterizations of a problem cannot be so easily combined. This might explain why philosophers and mental health experts sometimes seem to regard different explanations of mental illness as competing or mutually incompatible, despite their appreciation for the causal complexity of these conditions.

1. Framing the problem

There's a lot we don't understand about mental illness. But one thing almost everyone does understand is that there typically isn't a single explanation, much less a *simple* explanation, for why someone develops psychiatric symptoms. Mental illness is a very complicated kind of phenomenon, with many very complicated kinds of causes. And you don't need a clinical license or a philosophy degree to recognize that, in view of this complexity, many different kinds of facts are going to be relevant to whether and how a person develops a psychiatric condition. For example, most of us would agree that, if a person's genes had been very different, they would probably have had a very different kind of psychological life. But most of us think that the same would be true if a person had been systematically abused, or were constantly hopped up on cortisol, or had tended towards an obsessive kind of perfectionism about their lives.

Intuitively, then, we understand that many factors can make real differences to people's psychological outcomes. We also understand that these factors don't necessarily compete. A

person's psychological condition isn't caused by their genetics *rather than* their neurochemistry or cognitive traits, just as an election isn't won by individual ballots *rather than* a politician's campaign platform. Even if we don't have much of a philosophical vocabulary for defending it, most of us know that things are more complicated than "your genes made you do it." Clinicians and researchers clearly recognize this, as evidenced by their increasingly impassioned calls for "holistic" or "biopsychosocial" approaches to mental illness. But even people who know much less than experts do often talk about the importance of different explanatory factors—say, psychological trauma, neurotransmitter levels, and personality types—in a single breath.

A substantial body of empirical literature, however, seems to tell a strikingly different story. Across a range of correlational, experimental, and meta-analytic studies, researchers have found that providing people with information about one particular explanatory factor (say, genetics), rather than some other factor (say, trauma), tends to influence their reasoning about mental illness in startling and systematic ways. In other words, we now have strong evidence for the significance of "explanatory framing effects" in psychiatry: the particular kinds of explanation of symptoms people focus on seem to really *matter* for many of their downstream beliefs, attitudes, and behaviors, however broad-minded about mental illness they might otherwise appear to be. What is even stranger is that these effects don't seem to affect only the unwashed (or at least under-educated) masses. They also emerge in studies of expert psychiatrists and psychologists, as well as people with first-hand experience of psychiatric symptoms.

In this paper, I'll argue that these effects are much more puzzling and philosophically interesting than first meets the eye. I'll also argue that the best way to make sense of them is to take seriously the idea that explanations have intrinsically *normative* features and functions. The key to understanding why people reason differently across explanatory contexts, I will suggest, is to understand that explanations don't simply describe some number of facts about causal history. They also serve to characterize an outcome as representing a particular kind of *problem* or *issue.* Once we appreciate that explanations do this "problem-defining" work, our otherwise puzzling data will fall into place: they will reflect people's natural responses to new normative information.

This paper is organized into six sections. In the first two sections, I'll provide a selective overview of some of the recent literature on explanatory framing effects in psychiatry, and make a case for the strangeness and the philosophical significance of its findings. I'll then present, in section 4, what I take to be a novel framework for interpreting these data, which I will call the "normative model" of explanatory framing effects. On this model, different explanations will tend to differently affect their audience's reasoning because each encodes a different picture of the "real problem" from which the explanandum diverges (i.e., the kind of normative divergence it represents). For example, a biological explanation of depression will convey to its audience that depression is a specifically biological problem, and therefore that apt or appropriate responses would be directed at biological facts and norms. But a psychological explanation would suggest that depression is a cognitive problem, and so call for correspondingly cognitive solutions.

I'll argue that this normative model is a powerful alternative—or at least a powerful supplement—to more familiar ways of accounting for explanatory framing effects in psychiatry,

including those that appeal to judgments about causal relevance, non-rational intuitions, or good old-fashioned ignorance. The normative model, I will argue, edges out these alternatives in several important ways. It more neatly predicts and illuminates the precise effects we see in the literature. It also makes these predictions in a way that captures the potential *reasonability* of the variation in beliefs, inferences, and behaviors associated with different explanations. Perhaps most importantly, it presents a highly plausible general mechanism, consistent with recent work on causal reasoning, by which explanations would seem to perform intrinsically normative work.

2. Explanatory framing effects in psychiatry: a review of the evidence

In the last several decades, researchers have started to observe some surprising trends associated with giving people different kinds of explanations—for example, broadly biological, psychological, or environmental explanations—of even the very same psychiatric symptoms. In this section, I'll provide an overview of some of the most striking and robust kinds of effects to have emerged in this literature. These can, for present purposes, be grouped into three basic categories. First, there are studies that supply evidence for changes to *prognostic* reasoning associated with receiving different explanations of mental illness—for example, feelings of pessimism or hopefulness about its future course, or predictions about the likelihood of recovery. Second, there are studies that investigate the different inferences that people make about appropriate *interventions* when presented with different kinds of explanatory information. Third, there are studies that supply evidence for changes to the valence and strength of various *interpersonal attitudes*, including stigmatizing attitudes, empathy, and ascriptions of agency.

After I introduce some of the evidence for these effects, I'll present a few reasons for thinking that these data are more philosophically interesting and important than immediately meets the eye. At a first pass, the bare fact that people respond differently to different explanatory information might seem unsurprising. But I'll argue that there are both empirical and theoretical grounds for finding their particular responses puzzling, and for trying to understand what might account for them.

A. Influence on prognostic beliefs and attitudes

Some of the clearest and most consistent evidence for the differential effects of explanatory framings concerns people's thinking about psychiatric *prognoses.* A substantial body of research now suggests that, when people are given broadly biological explanations of mental illness, they tend to have bleaker views about the future course of these conditions than when they receive psychological or environmental explanations: they think that episodes of illness will last longer, recur more often, involve more severe symptoms, be less responsive to interventions, and require more extended treatment (for a review of much of this evidence, see Lebowitz & Appelbaum, 2019). In an important meta-analysis by Kvaale et al. (2013), for example, a review of 28 experimental studies yielded evidence for a significant association between what the authors call "biogenetic explanations" of mental illness—that is, explanations that invoke facts about genes,

brains, or biochemistry—and various forms of "prognostic pessimism".[1] A number of studies conducted since have further corroborated these results (see, e.g., Lebowitz et al., 2013; Haslam & Kvaale, 2015; Loughman & Haslam, 2018; Lebowitz & Ahn, 2018; Zimmerman et al., 2020).

One particularly striking feature of this literature, and one to which I will return at length below, is that it has produced evidence for the association between biological explanations of mental illness and prognostic pessimism across very different demographics. Intuitively, you might not expect members of the general public, people actively struggling with psychiatric symptoms, and clinical experts to react in the same way to information about the biological bases of mental illness. You might even think that—to the extent that this information would affect prognostic thinking at all—it would encourage *optimism* at least among clinicians and people experiencing psychiatric symptoms. After all, being able to identify the biological causes of mental illness seems like an important step towards engaging with it as a medical problem much like any other, for which we typically have targeted, evidence-based treatments. In this way, biological explanations could precisely brighten our outlooks on the course of mental illness.

But recent research tells a different story. Prognostic pessimism emerges as either an effect or correlate of biological explanations not just with ordinary audiences (see, e.g., Phelan, 2005; Bennett et al., 2008), but also with people actively experiencing psychiatric symptoms (Lebowitz et al., 2014; Gershkovich et al., 2018; Lebowitz et al., 2021). For example, Lam & Salkovskis (2007) found that, when people with anxiety symptoms were given genetic or neurobiological explanations of panic disorder, they were more likely to think that a person with the disorder would need an extended course of treatment, would be unlikely to recover, and were more likely to harm themselves or others, relative to both participants who received psychological explanations and a control group. Lebowitz et al. (2014) observed similar effects among people with a diagnosis of generalized anxiety disorder given biological explanations of GAD. In another striking study by Lebowitz & Ahn (2018), people experiencing symptoms of depression given sham "evidence" for a genetic predisposition to MDD were less confident that they would be able to recover.[2] (A similar design was employed by Kemp et al., 2014, with similar results.)

Experts probably aren't immune from the pessimism effect, either. Although there has not yet been much research directed at assessing prognostic pessimism among mental health professionals, Magliano et al. (2019) found that medical doctors who explained schizophrenia by reference to biogenetic causes were more skeptical about the likelihood of patients' recovery, and more convinced of the need for lifelong pharmacological interventions, than those who explained it by appeal to psychosocial causes.[3] Ahn et al. (2009) and Lebowitz & Ahn (2014) also found that psychiatrists, psychologists, and social workers who endorsed biological explanations for a

---

[1] A meta-analysis of correlational studies by the same authors produced similar results: see Kvaale et al. (2012). I borrow the term "prognostic pessimism" from Lebowitz & Appelbaum (2019).

[2] Lebowitz & Ahn 2017 conducted a similar "sham genetics" test on asymptomatic participants, and, stunningly, found that they were more likely to believe that they had experienced depressive symptoms in the past. See also Schroder et al., 2020 for correlational evidence of this association in a sample of inpatients.

[3] These results map neatly onto evidence collected from lay populations for the relationship between biomedical explanations and prognostic pessimism about schizophrenia (Bennett et al., 2008), which is at least a preliminary basis for suspecting that the same kinds of effects we find in the general population might carry over to experts.

mental illness were more likely to believe that recovery would require medication, and were less optimistic about the potential efficacy of psychotherapy. Given that most psychiatric medications are taken for significant periods of time, and are increasingly prescribed for long-term or "maintenance" therapy, there is reason to suspect that these inferences track more overtly pessimistic judgments about the likely persistence or chronicity of illness. In any case, it clearly suggests that biological explanations can increase clinicians' pessimism about at least some routes to recovery. This reflects another dimension across which different explanations of mental illness seem to have a significant differential effect: namely, people's judgments about which kinds of *interventions* are appropriate to pursue.

B. Influence on reasoning about interventions

Converging lines of evidence suggest that members of the general public, people experiencing symptoms of mental illness, and even expert clinicians tend to reason differently about treatment options for a given mental health problem in response to different explanatory information. In particular, people seem to consistently prefer interventions that are categorically congruent with the kinds of explanations of mental illness they accept. For example, when people are given broadly biological explanations of psychological symptoms, they are more likely to prefer treatment by medication over psychotherapy than when the same or similar symptoms are given psychosocial or environmental explanations (Proctor, 2008; Deacon & Baird, 2009; Marsh & Romano, 2016; Magliano et al., 2019).

Importantly, this preference doesn't seem to be just a "brute" preference, which might be fully explained in terms of some implicit intuition that explanations and interventions should track phenomena of similar categorical kinds. When people reason about some set of symptoms in light of, say, biological explanations, they don't seem to simply assume that pharmacological interventions "make more intuitive sense" than psychotherapy: they also predict that medication will be more *effective* relative to psychotherapy, and that psychotherapy will be *less effective in general*. Lebowitz & Appelbaum (2017), for instance, found that participants presented with genetic explanations for either alcohol use disorder or gambling disorder judged that medication was significantly more likely to be clinically helpful, and that psychotherapy was significantly less likely to be helpful, relative to people who received non-genetic explanations (see also Lebowitz et al., 2021). A similar pattern emerges when people are provided psychological explanations of a clinical vignette: they tend to predict, in such cases, that psychotherapy will be a more effective or more credible intervention than medication (Iselin & Addis, 2003).[4]

Crucially, this preference for "explanation-congruent interventions" does not seem to be limited to specific populations. They recur in studies of lay audiences (Marsh & Romano, 2016;

[4] There is also evidence for subtler distinctions within the domain of broadly non-biological explanations. For example, Kim & LoSavio (2009) found that psychological symptoms that were explained in terms of a person's internal psychological makeup or dispositions — e.g., their individual choices or behaviors — were judged to be in greater need of professional psychological treatment than people with the same symptoms whose behaviors were explained in terms of environmental causes (e.g., their childhood environment, or even other people's behaviors).

Deacon & Baird, 2009; Iselin & Addis, 2003), people with clinical symptoms (Lebowitz et al., 2021; Iselin & Addis, 2003), and mental health professionals (Ahn et al., 2009; Lebowitz & Ahn, 2014). So they reflect inferences that appear to be surprisingly pervasive and robust.

C. Influence on personal and interpersonal ascriptions

Different explanations of mental illness also seem to influence people's judgements about persons who experience mental illness, as well as—and perhaps relatedly—their interpersonal attitudes towards them. One especially consistent finding in this domain is that biological explanations of clinical symptoms tend to be associated not only with diminished ascriptions of blameworthiness for a person's having those symptoms, but also with weakened ascriptions of agential capacity more generally. In an influential study by Miresco & Kirmeyer (2006), for example, psychiatrists' ratings of the "neurobiological etiology" of mental illness symptoms were negatively correlated with their judgments of a person's "responsibility" for them (where these encompassed a wide range of judgements about blameworthiness, agential control, intention, capacity for change, and so on). Responsibility judgments were, however, positively correlated with ratings of "psychological" etiology.[5]

A similar effect has been reproduced in clinical samples by Deacon and Baird (2009) and Kemp et al. (2014), both of which found that explaining depression to people with depressive symptoms by appeal to "chemical imbalances" diminished their sense of self-blame, but also weakened their perceptions of their own agency with respect to managing or recovering from these symptoms, and even regulating their negative moods. This effect has also been observed in samples from the general public. In a particularly striking study by Dar-Nimrod, Zuckerman, & Duberstein (2013), participants rated themselves as less able to control their drinking when they were told—baselessly—that they had a genetic predisposition to alcoholism. In a similar study by Lebowitz and Appelbaum (2017), people provided genetic rather than non-genetic explanations of a person's psychological symptoms reduced both their ascriptions of blame and their more general ascriptions of agency and self-control.

Many other broadly interpersonal judgments seem to be modulated by different explanations of mental illness. In an unsettling study by Lebowitz & Ahn (2014), mental health clinicians reported feeling less empathy for hypothetical patients when their symptoms were explained biologically than when explained psychosocially. The effect persisted even when *both* biological and psychosocial explanations were provided, so long as the biological information was foregrounded. Other lines of research suggest that, when people's psychiatric symptoms are framed in terms of stressful life events, both laypeople and clinicians judge them to be less psychologically "abnormal" than when these explanatory contexts are not provided (Ahn, Novick, & Kim, 2003; Kim, Paulus, Gonzalez, & Khalife, 2012; Weine and Kim, 2018).[6]

---

[5] Intriguingly, judgments of psychological and neurobiological etiology were inversely correlated. I'll return to this finding, which is further bolstered by evidence from more recent studies, in section 5 below.

[6] Although these results don't bear directly on more specific contrasts between biological and psychological or environmental explanations, they are suggestive—especially once we consider that biological explanations tend to

Finally, a number of experimental and correlational studies have found that biological explanations of mental illness are associated with greater endorsement of negative stereotypes about people with psychiatric symptoms, including heightened perceptions of them as potentially unpredictable or dangerous (for review, see Haslam and Kvaale, 2015; Angermeyer et al., 2018; Baek et al., 2022). Some research has also produced evidence for an association between biological explanations of mental illness and a desire for social distance, although the literature on this is most mixed.[7]

3. What's the problem?

Considered individually, the studies I've reviewed above might seem straightforward enough. Each supplies evidence that people respond differently to different explanatory information. But this, you might think, is just what we should expect. Explanations tell us about causes, and different kinds of explanations tell us about different kinds of causes. So it's not surprising that people's beliefs about mental illness often change in concert with the kinds of explanations they accept. If I were to tell you that depression is explained by heightened levels of cortisol, it would only be natural for you to infer that dysregulated cortisol *causes* depression, and perhaps even that it is the *most potent* or *most relevant* cause of depression. But if I instead told you that depression is explained by maladaptive cognitive styles, you are likely to think that it is instead people's habits of thought—for example, habits of ruminating or catastrophizing—that is the causal factor most relevant to predicting and intervening in their being depressed.

At a first pass, this seems like a neat explanation. But I want to suggest that the empirical data, when taken together, present a picture that is much stranger and more confusing than this simple analysis would suggest. One way to get a sense for this is to notice that the effects that have emerged in the empirical literature should seem overtly *unreasonable*. To put the point more bluntly: people come out of these studies looking exceptionally stupid. Whether they are laymen or experts, they seem to consistently make the same extremely rudimentary mistakes in their reasoning. For example, we've seen that participants consistently judge that a psychiatric condition is likely to be especially severe, or that it can only be managed by medication, when they learn that it has some broadly biological causes. But these inferences are pretty clearly unfounded. For one thing, it seems plausible that *most* human outcomes can be understood, in principle, in terms of biology. Surely no one believes that there are facts about human beings that somehow float free of biological underpinnings. Perhaps more importantly, many conditions that have clear biological causes—such as gum disease, obesity, or diabetes—aren't especially severe. And they can often be managed by behavioral or environmental interventions (e.g., by people making changes to their diet), rather than strictly biological ones (e.g., surgery).

---

cite "internal problems", which are those more often perceived as evidence for clinically significant abnormality and the need for treatment (Kim and LoSavio, 2009).

[7] For example, a meta-analysis by Angermeyer et al. (2018) suggests that many of the effects on stigma might be modulated by the diagnosis in question, such that, e.g., biological explanations of schizophrenia, but not alcoholism, and only inconsistently depression, are associated with desire for social distance.

Of course, it's not impossible that ordinary people sometimes forget to think through these complexities, or even that they are pervasively ignorant about them. But I think it's hard to accept that people simply don't know that mental illness has many complicated causes, or that there are always going to be *some* biological factors relevant to its development. It should, however, be even harder to accept that mental health professionals—highly trained psychiatrists, psychologists, and social workers—are likewise naive. But the evidence indicates that expert judgments track the very same patterns as laypeople's. Even practiced clinicians seem to think that conditions that are explained biologically will face especially poor prognoses, and can really only be managed by biological interventions. And they seem to think the inverse, *mutatis mutandis*, for psychological or environmental explanations.

What is even odder is that clinicians seem to make these and related judgments in a manner that is totally inconsistent with some very basic tenets of clinical reasoning. For example, experts in these studies often seem to assume that, if a biological factor can explain some set of symptoms, psychological or environmental factors couldn't very successfully explain it, and couldn't be leveraged in order to treat it. In other words, they appear to reason as though explanations and interventions are implicitly "competitive" or "exclusionary": the availability of a good biological explanation or intervention leads them to think that all other kinds of explanations or interventions are *less plausible* (see, e.g., Ahn et al., 2009; Miresco and Kirmeyer, 2006; Lebowitz & Ahn, 2014).

But this should be shocking. Mental health professionals know better than anyone that there are typically many causes of a mental health problem, that these causes typically complement one another in complicated ways, and that effective interventions can target any, many, or even none of them. These are all foundational principles of the biopsychosocial model of mental illness, which is commonly regarded as the presiding "psychiatric orthodoxy" (Pilgrim, 2002; Ghaemi, 2010/2011; see also Bolton & Gillett, 2019). And we know that practiced clinicians tend to endorse these principles. For example, they clearly understand that mental illnesses don't usually have a single cause (Ahn et al., 2009), that different causal explanations are often complementary (Harland et al., 2009; Proctor, 2008; Brog & Guskin, 1998), and that effective treatments need not target any particular causal pathway (Ahn et al., 2006). But this rich causal understanding appears to be completely belied by the actual judgments they make.

In light of these striking contrasts between what clinicians *know* (in principle) and what they seem to *do* (in practice), an analysis of the data that hinges on the assumption that they are simply ignorant, or that they are inveterately sloppy causal reasoners, should start to seem much less convincing. But, of course, we might reach for a more complicated story. For example, many researchers who have contributed to the empirical literature on explanatory framing effects have at some point suggested that, if even experts are susceptible to such obvious errors, there must be powerful covert intuitions, heuristics, or cognitive biases that distort their reasoning. One

especially common proposal is that people's judgments in these cases are guided by implicit dualist or essentialist intuitions, rather than by any explicit beliefs they might actively endorse.[8]

Maybe this proposal is onto something: maybe deep-seated intuitions do often quietly guide people's thinking and jam up their judgments, irrespective of their claims to expertise. But even if this is true, I do not think it yields a very satisfying general-purpose account of the data we observe. For one thing, there is just not much evidence for the impact of these intuitions—and recent experimental studies that have sought to capture the influence of essentialist intuitions, in particular, did not find the same kinds of effects we observe in the wider literature on framing effects in psychiatry (Peters et al., 2020). We also know that mental health professionals often actively and conscientiously disavow dualism and essentialism. For example, Ahn et al. (2006) found that expert clinicians generally resisted the suggestion that mental illnesses have causal essences, and tended to believe that, even if there *were* one basic kind of cause for a mental illness, effective psychiatric interventions would not need to target it.[9]

This points to an even more general problem with the "big, bad biases" hypothesis. This is that, if implicit biases are to blame for the errors in experts' reasoning, we should expect that explicitly correcting for these biases would lead to significant improvements in outcomes. But in some of the most striking studies to date, participants were actively reminded of the importance of non-reductionism and causal pluralism while reasoning about explanations of mental illness. For example, Ahn et al. (2009) stressed to participating clinicians that "biological, psychological, and environmental causes [are] *non-mutually exclusive domains that could be overlapping*". And, indeed, on a free recall task, participants volunteered an average of 5.4 different causes for various mental disorders, and judged a full third of these causes to be "both biological *and* psychological" in nature. But these same clinicians later seemed to think that conditions with a significant biological basis would *not* have a very significant psychological and environmental basis, and vice versa. If implicit bias were really to blame for these effects, it is difficult to see why explicit correction did not mitigate them.[10]

---

[8] We have already seen some examples of this interpretative line: recall, for example, the suggestion by Ahn and colleagues that clinicians' apparently competitive judgments about the different causal bases of mental illness might be driven by illicit inferences from an intuition of explanatory exclusion (e.g., "if a genetic explanation is relevant, other explanations must be irrelevant").

[9] In their discussion of this result, Ahn and colleagues characterized it as an "effect of expertise", and emphasized that novices were more likely to endorse essentialist views about mental illness. Notably, however, *even novices* endorsed such views at a much lower rate than they did for medical disorders, and at comparable rate to their essentialist judgments about nominal kinds like "trees planted in the year 2002" and "dogs whose names begin with 'F'" (2006: 766).).

[10] We find further evidence for this general pattern in Miresco & Kirmeyer (2006), which studied psychologists' and psychiatrists' clinical reasoning vis-a-vis causal attributions. "Instead of treating [biological and psychological explanations] as different levels of explanation," the authors report, "[their] endorsement of biologically construed bases of behaviors (e.g., genes, brain structures, neurotransmitters) appears to be *inversely related* to their endorsement of psychologically construed bases of behaviors (e.g., intentionality, desire, motivations)." In other words, experts seem to consistently reason as though biological and non-biological explanations of mental illness are fundamentally in tension with one another. Note also that ~30% of clinicians in Miresco and Kirmeyer's study correctly guessed the experimenters' hypothesis — but their responses were not statistically different from other respondents'. This suggests, as the authors note, that *even explicit awareness of the research question* did not impact experts' tendency to reason about different kinds of causal attributions in a dualistic, competitive fashion.

So there are deep problems with familiar attempts to explain explanatory framing effects. Appeals to deep-seated biases, or indeed to judgments about causal relevance, do not explain some of the strangest features of the empirical data, especially the "exclusionary" character of people's explanatory judgments. More generally, and more damningly still, these analyses would have us accept that mental health professionals are nothing like the expert clinical reasoners we tend to think they are. After all, the apparent errors in their judgments are perfectly congruent with laypeople's, and perfectly incongruent with widely endorsed principles of clinical and causal reasoning. So we seem forced to say that clinicians are as likely as is the proverbial man off the street, or the typical undergraduate, to reason in extremely crude ways about mental illness—to think things like "every psychiatric condition has one kind of cause and one good intervention", or "the mind and the brain must be completely different systems", or "if something can be explained by a biological factor, that must be the only explanation we can give".

On reflection, however, this claim should seem deeply uncharitable. It should also seem downright implausible. Although the assumption of ignorance is often casually thrown about in the literature, it seems to me that such a damning conclusion about the poverty of expert reasoning should come as a last resort, not a first guess.[11] If clinicians are consistently displaying a distinctive pattern of judgment, it seems like good interpretive practice to at least wonder whether something deeper might be going on. Even if ignorance and cognitive sloppiness can illuminate some of its details, the broader picture painted by the evidence still calls out for explanation. Fortunately, I think there are deeper and more satisfying ways of making sense of it. I now want to consider a novel analysis of just this kind. I will call this the "normative model" of explanatory framing effects.

4. Causes, Norms, and Explanations

Before I start filling out the normative model, it will be helpful to think for a moment about explanations more generally. In particular, I want to take a quick step back to consider what it is we are really *doing* when we explain things. One extremely intuitive answer to this question—so intuitive, in fact, that you might not think there are viable alternatives to it—is that we explain things to one another in order to share causal information. In other words, we seek explanations primarily because we want to acquire true beliefs about the causes of some fact or event.[12]

As it turns out, however, there are good reasons for thinking that ordinary explanations do not track unvarnished facts about causal structure. If this is right, it suggests that the "causal

[11] For instance, in their discussion of their 2009 study results, Ahn and colleagues caution us that "new discoveries of genetic influences on a mental disorder could inspire possibly inaccurate inferences [by clinicians] that the disorder is no longer psychologically or environmentally influenced." But this should seem incredible. Even your average undergraduate would probably not make inferences quite this silly.

[12] This analysis, of course, doesn't capture the character of non-causal explanations. And even if we restrict ourselves to causal explanation, there are typically going to be further constraints—e.g., norms of relevance, nomological character, predictive force, counterfactual dependence—placed on the kinds of causal facts that can be properly explanatory. But these philosophical subtleties need not concern us here. The point is that this basic (and I think historically influential) picture should seem familiar.

communication" picture of explanation is *anemic*, even if otherwise correct. There are important and even essential features of ordinary explanations that it simply doesn't capture. One initial way of getting a feel for this is by reflecting on the fact that explanations are always selective and partial. They do not simply describe all the causes of an event we're interested in. Instead, various kinds of communicative, interpretive, and pragmatic norms systematically influence which kinds of facts they should include (see, e.g., Grice, 1989). For example, if you ask me "why did you start a PhD in philosophy?", I'm probably going to respond by highlighting only the factors that I expect you to find most relevant and insightful. If I instead start listing all the biological, environmental, and psychological links in the causal chain extending from my birth to my enrollment in a PhD program, my explanation would not just be strange; it would simply be bad.[13] In this way, explanations will always filter down facts about general causal structure to facts that are judged to be important in a particular context.

Research suggests, however, that our explanatory practices are influenced by implicit norms in even deeper ways than those indicated by considerations of mere relevance or contextual utility. For example, we now have a great deal of evidence that people's causal ascriptions—and therefore, it would seem, their causal-explanatory judgments—are influenced by considerations of moral valence, moral responsibility, statistical normality or abnormality, and norms of proper functioning (see, e.g., Kahneman & Tversky, 1982; Alicke 1992; Alicke et al., 2011; Hitchock and Knobe, 2009; Icard et al., 2017; Kirfel & Lagnado, 2018; Kirfel et al., 2024; Statham, 2020; Sytsma et al., 2012). When, for instance, the actions of two different people bring about some effect, but only one of them was not supposed to have acted as they did, people tend to say that it is the rule-breaker's actions (rather than the rule-follower's) that explain what happened (Hitchcock & Knobe, 2009). Similarly, when the functionality of a mechanism depends on the functions of many of its parts, but one part is functioning as designed and the other is functioning counter to design, people tend to say that the part that is *not* functioning as designed is the one that explains the mechanism's breaking down—even when an intervention into either of these parts would be sufficient to fix it (*ibid*.).

In this way, people in search of explanations seem to reason in light of "normalizing" counterfactuals. When trying to understand why something happened, they consider what would have happened if something more normal occurred instead. This suggests that people are sensitive to lots of surprisingly rich background norms about what's good, what's typical, or even what's purposeful when they reason about what caused what, and what explains what. An obvious question that arises in this context is why exactly this is. Various possible answers have already been carefully explored in the recent literature on norms in causal-explanatory judgment, so I won't say much on the matter here. One very convincing proposal, however, highlights the important role that explanations play in guiding our future action. The basic thought here is that,

[13] Or suppose I try to answer your question by thinking about possible counterfactuals: for example, (1) "if I hadn't read Nietzsche in high school, I wouldn't have decided to study philosophy," (2) "if I had been born five hundred years ago, I wouldn't have decided to study philosophy," or (3) "if I had been forced to go to law school, I wouldn't have decided to study philosophy." I will surely only consider the first of these possibilities: the others will simply seem irrelevant, although they might track equally plausible counterfactual claims.

if we want to change an outcome, we usually want to do so by making something go *better*—and this often means making sure something goes "less wrong", or becomes "less unusual", or functions "more optimally". In this way, focusing on abnormal events in our explanations helps us zero in on the most suitable possible interventions, by helping us see what *should* be made better in order for an outcome to change in the best possible way.[14]

I think this line of thinking captures something important. In fact, the "normative model" of explanatory framing effects, which I will now introduce, can be regarded as a variation on this theme. But it also stands to deepen our appreciation of the general theme, by providing a fuller picture of how and why we might arrive at judgments of normality, suitability, and relevance.

The Normative Model

The central idea underlying what I am calling the "normative model" is simple. It is just this: if explanations are sensitive to underlying judgments about the *normality* or *abnormality* of different nodes in a causal structure, and therefore judgments about the *suitability* or *unsuitability* of different ways of intervening in it, this is probably because they are sensitive to underlying judgments about which kinds of *problems* an outcome represents or implicates. In other words, in light of all the evidence for the impact of normative judgments on people's causal-explanatory reasoning, it seems very plausible to suppose that presenting people with different explanations of an outcome conveys broader normative information about—can reflect or further reinforce implicit judgments about—what exactly has "gone wrong" such that this outcome came about. In this way, different explanations would invite us to think not just in terms of different possible causal histories, but also in terms of different possible *kinds of wrongness*. And that is to say that they would encourage us to think in terms of different possible *problems.*

This proposal might sound suspiciously esoteric when considered in the abstract. But I think the basic idea it tracks is extremely intuitive. To see this, start by considering a very simple case. Suppose that my friend recently failed their qualifying exams, and I asked them *why* they failed. Here are two possible answers they might give me:

(1) I failed because the exam focused on Hegel's *Science of Logic*!

(2) I failed because I didn't focus on studying Hegel's *Science of Logic*!

I think it's clear that these explanations are not tracking different *causal* facts. If the exam was on a particular text that my friend did not know much about, both facts about the exam's contents, on the one hand, and facts about the state of my friend's knowledge, on the other, *jointly* led to their receiving a failing grade. In other words, these explanations are naturally interpreted as pointing to different features of the same causal structure. This structure licenses various counterfactuals: for example, if the exam had instead been on many different texts, *or* if my friend had instead mastered the Doctrine of the Concept, they would have passed rather than

[14] See Hitchcock & Knobe (2009) and Phillips et al. (2019) for more detailed developments of this view.

failed. But it's precisely because my friend didn't know much about Hegel that the first counterfactual is true, and because the exam only tested knowledge of Hegel that the second is.

On reflection, however, it should seem equally clear that each of these explanations communicates something very different about what *went wrong* with the exam. And, by the same measure, each communicates a very different picture of how the exam could have gone *right*. Consider *(1)*. This explanation suggests, especially when pronounced with a certain level of righteous indignation, that the problem with the exam was that it examined the *wrong things*. The question my friend is implicitly inviting me to consider here is something like this: "why did a qualifying exam, which really ought to assess general philosophical competence, focus entirely on one marginal and arcane text?" To the extent that I accept my friend's explanation, I will probably think this question is a fair one. And so I will probably start considering counterfactuals that involve ways in which the *exam* could have been better (more fairly, more aptly) designed. The relevant interventions suggested by these counterfactuals would then involve protesting or endeavoring to change this design—say, petitioning the department chair to declare the exam invalid, or pressuring the faculty examiners to rethink their standards of professional assessment.

Explanation *(2),* however, does something very different. It suggests that the problem with the exam was not the nature of its design, but rather my friend's *lack of preparation* for it. In light of this, it immediately invites the consideration of different questions (e.g., "why didn't you study more Hegel?"), different counterfactuals (e.g., "what if you had studied more Hegel?"), and different interventions ("master the method of determinate negation", "acquire a better understanding of the German Idealists", etc). But this is not because the second explanation explicitly or implicitly disputes any of the counterfactuals suggested by the first explanation, or indeed the efficacy of the interventions implied by them. It's still true that if the exam would have been designed differently, or if the exam results had been declared invalid, my friend would not have failed. Invoking *(2)* has the distinctive effect it does not by denying any of these causal or counterfactual features of the exam's outcome, but rather by communicating that the *real issue* with this outcome—the thing that really went wrong, and therefore the thing that really should be made right—is that my friend did not do a good enough job of preparing for it. This naturally suggests that the right kind of solution to the issue, the *real* solution, will involve changes to my friend's study habits and philosophical literacy, not changes to the nature of the exam.

So here we have two different explanations which, when considered in terms of their descriptive content, are not just consistent but fully complementary. Each is true precisely because the other is. But they seem to license very different ways of thinking about the outcome they jointly explain, by encouraging us to think in terms of different kinds of problems. In other words, they give us different senses of what kind of *wrongness* my friend's failure represents.[15]

---

[15] You might think that this gloss only works because "failing an exam" is *intrinsically problematic*. Would the same kind of analysis work if we were dealing with an ordinary explanation of some humdrum event? I think it often will. I use the language of "problems" here loosely, to indicate there is a particular way in which things veered off their normal course. Some such "veering off-course" is typically what makes us seek explanation in the first place. If everything is going precisely as expected, or precisely as I think it should, I probably won't ask searching questions about why this is the case. If I find myself confused or puzzled, it's usually because something isn't going the way I thought it would (the clear skies have suddenly turned stormy; my computer isn't booting; my friend is late to

There are three things that it is extremely important to notice here. The first is that there is no *empirical* fact of the matter about which of these explanations is the better one. We couldn't simply inspect the world, or our best causal models, in order to determine that one of these explanations gets things right and the other gets things wrong, or that one is more and the other less adequate. This is because to say "the real problem is X" is not to make an empirical claim about the way a situation has actually shaken out. It is to make an intrinsically normative claim about how we *ought to think about* its stakes and significance. One way to see this is to consider whether someone with comprehensive knowledge of the causal history of the exam would be able to authoritatively pronounce upon what the problem with it was. I want to suggest that they would not be. The question "what was the problem that led to this outcome?" can only be answered by appealing to a normative picture of how a good outcome should have been brought about. And even a full set of causal and historical facts are consistent with many such pictures.

The second thing to notice is that presenting a situation in light of a particular kind of problem involves the transmission of *complex* normative information. In other words, conveying "the problem with outcome *O* is feature *F*" isn't just a matter of communicating a single claim about the kind of badness, wrongness, or strangeness represented by a particular event. When my friend tells me that they failed the exam because it was on the *Science of Logic*, they're not *just* telling me that this happens to be a bad text to examine people on. They're suggesting a more general way of thinking about a number of deeper and related issues: for example, what the goals of academic assessments are or should be, which features of these assessments are worth significant attention, what kinds of issues might be associated with them, what criteria should inform our judgments of whether these issues have successfully been dealt with, what kind of activities or practical responses are really valuable with respect to dealing with them, and so on.

This is important, because it makes sense of why explanations that foreground different problems might reasonably recruit many different kinds of downstream inferences. They don't just lead us to think "*F* is bad with respect to *O*" or "let's focus on changing *F*". Instead, they tell us something like "think about *O* in light of the norms relevant to *F*-ness." (For example, being told "the problem with the exam was that it was on Hegel" does not communicate something like "the exam should have been on Schelling instead", but rather: "think about philosophy exams in light of changing norms of philosophical importance, academic competence, or fairness to students of different backgrounds or interests". Or indeed: "do not simply think about philosophy exams in light of particular students' preparedness.") In this way, presenting a problem to people naturally motivates them to think about the problematic situation in terms of a very specific, but

---

dinner; my father has taken up smoking; my brother can't get a job; my partner is ill; bad things happen to good people; etc.). Of course, sometimes these unexpected happenings are *positive* (my partner makes an abrupt recovery; my brother wins a Nobel). But I think we can understand these happenings as representing a *solution* to a particular kind of problem (the problem of illness, of not winning prestigious prizes, and so on). Or you might think of "problems" as generic "violations of norms", rather than as a concept associated with a specifically negative valence. In this way, any deviation from a normal state suggests a "problematic" interruption of the ordinary course of things. I want to maintain the language of "problems" here, rather than the more generic language of "norm-violations", because I think we have a very intuitive grasp of the general importance of problems, but not of the general importance of "norms."

potentially very rich, normative vocabulary—one that seems uniquely appropriate to capturing and exploring that specific kind of problem. And this can change a lot more than just a single belief about what went wrong in a single situation. It can lead people to shift their focus to particular kinds of default states, variables, relationships, and standards of assessment in their thinking about this sort of problem situation in general.

This narrowing of normative focus is, I think, an essential feature of problem-based reasoning. In order to identify and solve a problem, we need a firm general grasp of the things we should care about and deal with: this is what ensures that our reasoning about and reactions to the problem are guided and systematic, rather than chancy and haphazard. But notice that this kind of focus is plausibly just as essential to explanatory reasoning. If explanations are going to help us organize our interactions with the world, as most philosophers think they fundamentally should and do, they had better tell us what to care about, and in what way. The world is simply too complex to make the construction and interpretation of comprehensive models of causal structure a generally viable or desirable aim. So we need to be strategically blinkered. If we weren't, we would not know which features of an outcome, and which mechanisms of intervening in it, and which kinds of norms governing those interventions, are those that we should further explore and engage.

This brings me to one last and especially critical point. This is that the "narrowing of normative focus" that I am claiming is a natural function of problem-based thinking is also reflected in our *ordinary* ways of thinking and talking about problems. Although I can't defend this point in fullness here, it seems to me an extremely important fact about our ordinary grammar that we tend to talk about "the root problem" or "the real problem" with a situation. Considering that there are typically multiple things that could be said to be going wrong when something is troublesome, this might seem slightly strange. But I think it is true: we do not usually point to any number of problems that a situation might happen to reflect. Instead, we tend to focus our attention on what we take to be the "fundamental" problem, in a way that often shapes our understanding of why that situation is the kind of situation it is (say, a failure; an illness; a breakdown; a surprise).[16]

This is where talk of "problems" can start to seem importantly different from talk of "causes".[17] It is fairly easy to think about an event as having various different causes. In fact, we

[16] Return, for example, to the scenario of failing an exam. If I failed an exam, it's possible, and even probable, that many things went wrong such that this happened. I might simply not have studied the right things. But I might also have been sleep-deprived, or in the throes of a panic attack; or I might have lost my sense of time, or had trouble literally understanding the exam questions. Notice that many of these problems might be closely connected, and even mutually implicated. (Perhaps it's because I didn't study the right things that I had trouble understanding the questions, or perhaps I had trouble understanding the questions because I was sleep deprived, or perhaps I lost my sense of time because I was having a panic attack after recognizing that I had studied the wrong things, and so on.) But if you ask me why I failed, I will probably feel pressure to pinpoint one *deeper issue* that this whole complicated situation represents, especially if I am simultaneously trying to help you understand what kind of issue it is in the first place. In this case, I will take "why did you fail the exam?" to be a question about what fundamentally went wrong with the exam-taking such that it resulted in a failure. And so I will explain this exam in such a way that tells you exactly what I think the problem is (for example: "I was extremely anxious").

[17] Or, rather, where it will start to seem a lot like talk of "actual causes."

all simply *know* that every event has a complex causal history. It is much more difficult, however, to think of an event as representing different kinds of "root problems." For example, I can pretty easily accept that my friend's failing their qualifying exam is caused by facts about both their preparation and the nature of the examination. But once I think of their exam outcome as reflecting a basic *design* problem, it becomes difficult for me to think that it also, simultaneously, reflects a basic problem with my friend's preparedness. If the exam shouldn't have been on Hegel, the fact that my friend didn't study Hegel is, in a way, *besides the point*. Although it is true, it is not what is really concerning. The real problem is that the exam did not serve to assess graduate students in a fair and methodologically well-grounded way.

If this intuition tracks, it might explain why many people, including many philosophers, often seem to regard different explanations of at least some outcomes as competing or mutually incompatible, despite their appreciation for the complexity of causal history. When they say things like "these explanations can't both be right!", or "if this explanation is good, then this other one can't be very good!", they might not be calling out to us from the depths of explanatory chauvinism. They might simply be appropriately responsive to the claim to exclusivity implicit in our judgments about "real problems."

5. Problems and prescriptions in psychiatric explanation

Let's return to explanations of mental illness. My angle on this should now seem fairly obvious: I think that many of the puzzling effects that researchers have observed when giving people different explanations of psychiatric symptoms will begin to make a great deal of sense once we think of these explanations as pointing towards different kinds of "real problems". In fact, I think that this normative model can help us make sense of these effects not only from a diagnostic perspective, but also from a *rationalizing* perspective. In other words, it can help us see why it might actually be reasonable to make at least some of the inferences that people do.

Let's begin by spelling out a bit more carefully the general mechanism that, according to the normative model, would drive the effects we're concerned with. This model invites us to think about explanations as encoding information about a target event's representing a particular kind of *problem*, or a divergence from a particular kind of "normal", non-problematic case. Even at a very abstract level, this idea should seem to translate quite naturally into the explanatory context of psychiatry. Explanations of mental illness center on *illnesses*, which are essentially and even paradigmatically "problems."[18] And it's obvious that there are many ways of understanding what kinds of problems these are, and what kind of "unproblematic" states they ought to be contrasted with. But thorny ontological questions about the deeper nature of psychopathology are extremely difficult to answer. And so you might reasonably think that we should leave these questions to the research scientists, psychiatrists, and philosophers to sort

[18] I think the point generalizes, as I've suggested many times already, but that's not crucial for my argument here. Note, though, that the deep relevance of "norms and problems" talk is not restricted to psychiatry: all medical explanation, and very plausibly all biological explanation, will at some point make central reference to normatively thick notions of "function", "dysfunction", "pathology", and so on.

out—and that, in the meanwhile, we can simply explain (that is, describe the causes of) why mental illness occurs.

But if the normative model is right, it will be extremely difficult to divide up the explanatory labor in this way. This is because the model suggests that we are often implicitly coming down on these thorny questions, even in the apparently innocuous activity of giving and receiving explanations of psychiatric symptoms. In other words, the normative model predicts that our choices about precisely which factors to foreground in explanations of mental illness are influenced by, and will themselves influence, our general sense of what the problem represented by that illness really is. If this is right, it means that when I say, for instance, "Sally is depressed because of a neurotransmitter imbalance," what I am saying is not just "abnormal neurotransmitter levels are causally related to her depression". I am also saying that "Sally's problem—call it 'depression'—is really a neurotransmitter problem." If, on the other hand, I explain Sally's depression by reference to cognitive traits and attitudes (say, "Sally is depressed because she ruminates too much"), the normative model predicts that my audience will infer not only that Sally's depression is caused by her cognitive habits, but also that her problem is really one of cognitive style. Similarly, if I say "Sally is depressed because she's been out of a job all year", it predicts that my audience will take me to be saying that her depression is a problem of economic precarity—in other words, that it is basically a social or environmental problem.

This way of analyzing the impact of different explanatory claims should seem very intuitive. In fact, this is one of its major virtues. The strongest reason for taking the normative model seriously, I think, is that its predictions should seem not just obvious, but almost ineluctable. If I explain an episode of psychosis or anxiety to you by pointing to facts about biological mechanisms, it should seem extremely natural for you to infer that psychosis or anxiety are "biological mechanism problems." One way of seeing this is to try *not* to make this assumption. If you are anything like me, you will find this hard to do, at least without engaging in some fairly explicit philosophical reasoning. If psychosis or anxiety weren't biological mechanism problems, why would you be pointing me to precisely these factors in explaining them? Why wouldn't you point instead to facts about what you think the essential problem really is—say, facts about cognitive processing, or interpretive styles, or inferential patterns, or unusual or maladaptive forms of coping with existential, social, and environmental problems?

The analyses suggested by the normative model are, however, not only intuitive in the abstract. They also supply very compelling explanations of precisely the kinds of effects that researchers have observed when studying the impact of different explanatory framings of mental illness on people's judgments. To see this, let's reflect on how these explanations might run.

Consider first **changes to interventional inferences.** We've seen that when people—including expert psychiatrists and other mental health clinicians—are presented with biological explanations of psychiatric symptoms, they tend to think that medication, but not psychotherapy, will be an effective treatment. But when they are presented with psychological explanations of these same symptoms, they infer exactly the reverse. This should seem very strange. If, however, accepting a biological or psychological explanation of mental illness encourages us to think of

mental illness as a specifically biological or psychological *problem* (respectively), this preference for congruent kinds of interventions starts to look much more comprehensible.

This is because it is quite reasonable to think, when encountering a particular kind of problem, that a genuine solution to it—one that *really resolves* the root issue—will address the problem on its own terms. In other words, it's natural to assume that a good solution to a problem will directly address the bad-making or wrong-making features that make it the particular problem it is. For example, if I tell you that I keep missing my afternoon appointments because I have a problem with waking up before 2pm, it should seem obvious that the way for me to *really* solve this problem is to deal with my habit of oversleeping. Of course, I could also address the fact that I keep missing my afternoon appointments by making sure all my meetings are scheduled in the evenings instead. Rescheduling in this way would clearly be a helpful intervention. But I think it would, just as clearly, not be a "real solution" to the central issue.

This kind of sensitivity to the importance of "real solutions" is reflected in a lot of ordinary rhetoric about mental illness treatments—and indeed, as I have just suggested, in our problem-solving talk more generally. Think here of the familiar charge that something is "just a quick fix" or a "band-aid", and the deep skepticism and even disdain that these epithets convey about proposed solutions to a problem (e.g., depression, income inequality, or racial prejudice). The reason these charges are so effective, I think, is that everyone understands that there are often lots of different ways of temporarily coping with or handling a problem, but much fewer ways of addressing the real heart of the issue. This is why criticisms of "quick fixes" don't usually dispute the *possibility* of effectively implementing them. What they dispute is the normative *appropriateness* of pursuing them in lieu of a real solution.

"Real solutions", in contrast to quick fixes, are interventions that change an outcome in precisely the right way, rather than in any old way. They solve the problem by bringing a situation back into alignment with the right kinds of norms—the norms in light of which that situation is found to be lacking, and so is defined as a problem, to begin with. And we reasonably have strong preferences for these kinds of solutions: we either flatly judge them to be more "apt" responses to the problem, or we think they are the only way of genuinely, non-incidentally correcting it. If, for example, I think that anxiety is fundamentally a biological problem, it seems reasonable to think that I can only get a "real fix" by addressing the specifically biological wrongness at issue, such as might bring the anxious person back in line with biological norms. Cognitive behavioral therapy would then probably seem to be only a "coping mechanism", or perhaps even a temporary "band-aid" to wear until the real treatments kick in. The inverse would be true if anxiety were understood in terms of psychological, social, or environmental problems. Once I see anxiety symptoms under these lights, it would be hard not to conclude that medications would be, even if effective, only a stopgap measure. The kinds of treatment that would seem genuinely appropriate and potentially helpful would likely involve, say, psychodynamic explorations of emotional history, or support in challenging social norms.

Now consider **changes to prognostic beliefs.** As we've seen, studies have repeatedly found that broadly biological explanations of psychiatric symptoms lead to significantly greater

pessimism about the course of mental illness than do psychological or environmental explanations of those very same symptoms. But this is, again, very odd. Why should simply learning that a mental illness is caused by biological factors lead us to think that its symptoms will be more severe, more chronic, and less responsive to treatment?

We can start to make sense of these pessimistic inferences if biological explanations of mental illness communicate that mental illness is a fundamentally *biological problem*. After all, pessimism is a perfectly natural response to encountering problems that we do not really understand, and do not really have a sense of how to solve. I can only reasonably be optimistic about the resolution of an issue if I have a pretty clear idea of what it would mean to solve it, and a pretty firm basis for thinking that I *could* solve it. But biological problems, at least as they arise in the context of mental illness, are precisely not issues of this kind. We don't usually understand what they involve, and so they tend to leave us stumped and confused.

Suppose, for example, I tell you Pedro is depressed because he has a "neurotransmitter problem". What can you now realistically infer about the deeper nature of Pedro's problem? What can you reasonably assume about how to rightly respond to his distress, how to tell whether such a response has been effective, how long his distress will take to fully resolve, or even whether it can be effectively resolved at all? Unless you happen to be a neurophysiologist (and probably even if you are), these questions are likely to leave you somewhat bewildered. And this difficulty becomes even more glaring when we turn to genetic problems. What would it even look like to solve a genetic problem? Most of us don't really know what genetic facts involve, much less what genetic "problems" involve, much less what might be done to meaningfully correct such problems (we certainly can't intervene directly on the genome!). So if we learn that Pedro's suffering is a specifically genetic problem, we probably won't feel very clear-eyed or hopeful about his prospects. We will likely find it difficult to envision a way in which these problems could be truly resolved.

We will, however, have a much easier time getting a firm and reassuring grip on problems of other kinds. The category of psychological problems, for example, is extremely familiar to us. We all deal with various issues related to our thoughts, feelings, and cognitive habits, at least in some way or to some degree, and so we are well-practiced at making sense of the general class to which they belong. We intuitively know how to assess—at least in a sketchy and preliminary way—when and in what respect someone has a psychological problem, and what it would mean to see it genuinely resolved. We also know first-hand, and in light of a lifetime's worth of evidence, that people can and often do meaningfully respond to these problems.

Much the same is true, I think, for most social, cultural, and environmental problems, especially because they are often quite structurally similar to psychological problems. These are issues on which we already have a fairly strong "normative grip": we understand why they're bad, and what might genuinely make them better. So we are less likely to feel pessimistic on principle when we learn that someone must grapple with them. This would explain the finding that explanations that foreground these kinds of issues are also much less likely to induce the kinds of negative expectancies than explanations invoking biological problems.

Finally, consider the observed **changes to interpersonal attitudes**, such as effects on attributions of agency, stereotyping, and even baseline levels of empathy. I think these results, too, will become much more comprehensible once we recognize the impact of thinking about mental illness in terms of different basic problems. Consider, for example, how people might respond to being told that someone has "neurotransmitter problems." Such problems are conceptually so far removed from the ordinary vocabulary of personal and interpersonal life that it will probably take real effort to think about them as problems of more or less ordinary human agents. (Even professional moral philosophers sometimes struggle to do so.) In general, we all know that people are usually responsible in *some* way for their beliefs, desires, intentions, and behaviors; but it's much less clear that they can be responsible for their genetics, hormone levels, or neurological structures. So it makes sense that, when psychiatric problems are framed in terms of the latter sorts of things, we will have a much more difficult time relating to the persons whose problems they are as ordinary kinds of agents. We simply won't know how to understand their problems in terms of a commonsense psychological and moral framework, much less help resolve them. For this reason, we might have a hard time figuring out how to engage with them, or how to deal with the struggles they characteristically deal with. In the face of invitations to prediction and imaginative projection, we would find ourselves rudderless. A general reluctance to morally or rationally evaluate such people, and to sincerely and intimately engage with them, would then naturally follow.

If, however, we think of someone's suffering in terms of broadly familiar psychological problems—say, problems with rage, rumination, social anxiety, or self-control—we can much more easily employ our ordinary conceptual and interpersonal tools to try to understand and help them. For example, we might try to reason with them, criticize or defend their behavior, attempt to convince them to change their minds or habits, advise them to talk to their friends or loved ones, encourage them to find a psychotherapist, and so on. We will, in short, regard them as agents with whom we ought to rationally engage in a familiar, interpersonal sort of way. And this is exactly what researchers report in the empirical literature. As we've seen, when people are given psychological explanations of psychiatric symptoms, they tend to think that psychotherapy is the best kind of treatment; they are less likely to think of people with these symptoms as deeply abnormal; they are more inclined to hold them accountable for their behavior, and to assume that they can control or change their thoughts and feelings; and they are not disinclined from pursuing extended or particularly intimate forms of social contact with them. All of these effects are perfectly consistent with, and indeed predicted by, the normative model. They follow from an invitation to attend to particular, familiar problems.

6. Pick up your prescriptions

I began this paper by introducing some puzzling results from recent research on explanatory framings of mental illness. This literature suggests that people consistently respond in surprising ways to different explanations of even the very same psychiatric symptoms. I have

argued that these systematic impacts on the reasoning of so many people, including experts, should make us curious about what might be driving them. And I've suggested that explaining the data away as mere evidence of ignorance is uncharitable to the point of implausibility.

But there is a way of making sense of explanatory framing effects that does not require such uncharitable and implausible assumptions. If, as I have been suggesting, explanations of psychiatric symptoms don't simply communicate facts about their causal history, but also motivate judgments about the "real problem" represented by them, it's not surprising that people tend to think differently across different explanatory contexts. In fact, we should simply expect different explanations of mental illness to encourage distinctive judgments about (among other things) what kinds of norms to center in our reasoning about psychiatric conditions, and how we should bring these conditions, or the people whose conditions they are, back in line with them.

Of course, this needn't lead us to endorse all of the specific inferences people make in response to presentations of a problem. At least some of the inferences considered above should continue to seem pretty problematic.[19] I think the normative model does, however, illuminate the overall reasonability of the general *mechanism* by which people arrive at them. Although a detailed defense of this claim will have to be left to another occasion, it should be easy enough to see why thinking in terms of problems would serve as a crucial cognitive strategy. Even in very simple cases (as, for example, when moping about a failed exam or a missed appointment), we can often understand what is wrong, strange, or unusual about an outcome in a number of different ways. But learning about real problems attunes us to those of its features we really *should* care about, so that we have a sense of what norms to let guide our further reasoning about and responses to its occurrence. This guidance is crucial, especially over the longer run of inquiry and action, in focusing our thinking about particular kinds of outcome in ways we deem normatively apt—which, importantly, need not coincide with those that would be strategic for the purposes of locally optimized predictions or interventions.[20]

If this is right, the issue with explanatory framing effects is not simply that people are responsive to different presentations of a problem. It is rather that they do not realize, as they are so responding, that normative judgments are pulling their strings. But this is one thing it is absolutely essential that they do realize, given that questions about the nature of mental health problems are not just inevitably consequential, but simply inevitable. Satisfying answers to these questions are not decided merely by matters of fact; they are shaped by our sense of what the facts should have been and should be. And that is to say that we will often have to make real choices about which of these answers to endorse. Is the basic problem reflected by my

---

[19] For example, if a psychiatrist assumes that a particular mental illness represents a biological problem when they receive a biological explanation of symptoms, but would take it to be a psychological problem were they given information about psychological traits, they would seem to be making a completely unwarranted leap in either case. And this is one kind of inference we do observe in the data.

[20] This marks an important difference between the normative model I've considered here and more familiar accounts of norms in explanation. These familiar accounts tend to emphasize the role of norms in guiding us to optimal interventions, often with the implication that explanations are directed at maximizing their efficacy or reliability. The normative model, however, shows how we would go about determining which interventions are optimal *in the right kind of way*. In so doing, it deepens the role that norms play in guiding explanatory reasoning.

depression the fact that my neurotransmitters are out of whack, or is that I am succumbing to obsessive self-criticism? Or is it, perhaps, that I am profoundly isolated, or existentially adrift, or living in conditions of extreme economic precarity? As far as I can tell, there is no non-normative information that could dictate a single, uncontroversially correct answer to this question (although there could certainly be better or worse reasons for defending one answer rather than others).[21] This means, however, that the responsibility for supplying a convincing answer cannot be cleanly offloaded onto the empirical facts.

How best to respond to this responsibility is another issue. But, as with all such matters, the first step must be to recognize that we can, in fact, be called to account. And here the normative model of explanatory framing effects gives us just the resources we need. In pointing us towards the importance of representations of "real problems," it helps us to pick up on important kinds of prescriptions that we regularly make and receive, so that we do not find ourselves immediately and unreflectively rushing off to fill them. It also helps us predict and explain a great deal of otherwise puzzling empirical data, by supplying a convincing mechanism for the framing effects that we observe in the context of psychiatric explanations, and perhaps even in explanations more generally.[22] For reasons to which I have alluded above, I suspect that the influence of this kind of problem-based thinking on our reasoning is likely to be extremely cognitively and ethically significant. Its importance has, however, so far gone largely unappreciated. Fortunately, this is the rare kind of problem to which there is a clear enough solution.

References

Ahn, W. K., Flanagan, E. H., Marsh, J. K., & Sanislow, C. A. (2006). Beliefs about essences and the reality of mental disorders. *Psychological Science*, *17*(9), 759-766

Ahn, W. K., Proctor, C. C., & Flanagan, E. H. (2009). Mental health clinicians' beliefs about the biological, psychological, and environmental bases of mental disorders. *Cognitive science*, *33*(2), 147-182.

[21] At this juncture, you might want to resist the need for nominating a "real problem" at all. Shouldn't we insist that there are often many equally important problems with a situation of interest, rather than a single "root" problem? My answer is that we *can* say this, but it seems to me that we shouldn't. Although I cannot defend this claim here, I believe that judgments about "real problems" serve an ineliminable function in our cognitive economy. But we can rest content with a more modest claim: if there are even some cognitive benefits to reasoning in terms of basic problems, and if the evidence suggests we do reason in this way, there are strong (though defeasible) grounds for taking this seriously. There are also clear pragmatic grounds for assuming that we cannot really dispense with it.

[22] Nettle et al. (2023), for example, recently conducted a series of experiments that aimed to assess people's responses to different kinds of explanations of human behavior *in general*. Their results indicate effects very similar to those we observe in explanations of mental illness, which neatly complements the hypothesis I've been considering here. If problem-based reasoning is simply a feature of explanatory reasoning, then its influence should not be restricted to specifically psychiatric explanations.

Ahn, W. K., Kim, N. S., & Lebowitz, M. S. (2017). The role of causal knowledge in reasoning about mental disorders. In *The Oxford handbook of causal reasoning* (2017), ed. Waldmann, 603-618.

Ahn, W. K., Novick, L. R., & Kim, N. S. (2003). Understanding behavior makes it more normal. *Psychonomic Bulletin & Review*, *10*(3), 746-752.

Alicke, M. D. (1992). Culpable causation. *Journal of personality and social psychology*, 63(3), 368.

Alicke, M. D., Rose, D., & Bloom, D. (2011). Causation, norm violation, and culpable control. *The Journal of Philosophy*, *108*(12), 670-696.

Baek, C. H., Kim, H. J., Park, H. Y., Seo, H. Y., Yoo, H., & Park, J. E. (2023). Influence of biogenetic explanations of mental disorders on stigma and help-seeking behavior: a systematic review and meta-analysis. *Journal of Korean Medical Science*, *38*(3).

Bennett, L., Thirlaway, K., & Murray, A. J. (2008). The stigmatising implications of presenting schizophrenia as a genetic disease. *Journal of genetic counseling*, *17*, 550-559.

Bolton, D., & Gillett, G. (2019). The Biopsychosocial Model of Health and Disease: New Philosophical and Scientific Developments. Palgrave Pivot.

Brog, M. A., & Guskin, K. A. (1998). Medical students' judgments of mind and brain in the etiology and treatment of psychiatric disorders: a pilot study. *Academic Psychiatry*, *22*(4), 229-235.

Dar-Nimrod, I., Zuckerman, M., & Duberstein, P. R. (2013). The effects of learning about one's own genetic susceptibility to alcoholism: a randomized experiment. *Genetics in medicine*, *15*(2), 132-138.

Deacon, B. J., & Baird, G. L. (2009). The chemical imbalance explanation of depression: Reducing blame at what cost?. *Journal of Social and Clinical Psychology*, *28*(4), 415-435.

Gershkovich, M., Deacon, B. J., & Wheaton, M. G. (2018). Biomedical causal attributions for obsessive-compulsive disorder: Associations with patient perceptions of prognosis and treatment expectancy. *Journal of Obsessive-Compulsive and Related Disorders*, *18*, 81-85.

Ghaemi, S. N. (2010). *The rise and fall of the biopsychosocial model: Reconciling art and science in psychiatry.* Johns Hopkins University Press.

Ghaemi, S. N. (2011). The biopsychosocial model in psychiatry: A critique. *American Journal of Psychiatry*, *121*(1), 451-457.

Harland, R., Antonova, E., Owen, G. S., Broome, M., Landau, S., Deeley, Q., & Murray, R. (2009). A study of psychiatrists' concepts of mental illness. *Psychological medicine*, *39*(6), 967-976.

Haslam, N., & Kvaale, E. P. (2015). Biogenetic explanations of mental disorder: The mixed-blessings model. *Current Directions in Psychological Science*, *24*(5), 399-404.

Hitchcock, C., & Knobe, J. (2009). Cause and norm. *The Journal of Philosophy*, *106*(11), 587-612.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength. *Cognition*, *161*, 80-93.

Iselin, M. G., & Addis, M. E. (2003). Effects of etiology on perceived helpfulness of treatments for depression. *Cognitive therapy and research*, *27*, 205-222.

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, *11*(2), 123-141

Kemp, J. J., Lickel, J. J., & Deacon, B. J. (2014). Effects of a chemical imbalance causal explanation on individuals' perceptions of their depressive symptoms. *Behaviour research and therapy*, *56*, 47-52.

Kendler, K. S. (2005). Toward a philosophical structure for psychiatry. *American journal of Psychiatry*, *162*(3), 433-440.

Kim, N. S., & LoSavio, S. T. (2009). Causal explanations affect judgments of the need for psychological treatment. *Judgment and Decision Making*, *4*(1), 82-91.

Kim, N. S., Paulus, D. J., Gonzalez, J. S., & Khalife, D. (2012). Proportionate responses to life events influence clinicians' judgments of psychological abnormality. *Psychological assessment*, *24*(3), 581.

Kirfel, L., Harding, J., Shin, J., Xin, C., Icard, T., & Gerstenberg, T. (2024). Do as I explain: Explanations communicate optimal interventions.

Kirfel, L., & Lagnado, D. (2018). Statistical norm effects in causal cognition. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.). Proceedings of the 40th Annual Conference of the Cognitive Science Society. Madison, WI: Cognitive Science Society.

Kvaale, E. P., Haslam, N., & Gottdiener, W. H. (2013). The 'side effects' of medicalization: A meta-analytic review of how biogenetic explanations affect stigma. *Clinical psychology review*, *33*(6), 782-794.

Kvaale, E. P., Gottdiener, W. H., & Haslam, N. (2013). Biogenetic explanations and stigma: A meta-analytic review of associations among laypeople. *Social science & medicine*, *96*, 95-103.

Lam, D. C., & Salkovskis, P. M. (2007). An experimental investigation of the impact of biological and psychological causal explanations on anxious and depressed patients' perception of a person with panic disorder. *Behaviour research and therapy*, *45*(2), 405-411.

Lebowitz, M. S., Ahn, W. K., & Nolen-Hoeksema, S. (2013). Fixable or fate? Perceptions of the biology of depression. *Journal of consulting and clinical psychology*, *81*(3), 518.

Lebowitz, M. S., & Ahn, W. K. (2014). Effects of biological explanations for mental disorders on clinicians' empathy. *Proceedings of the National Academy of Sciences*, *111*(50), 17786-17790.

Lebowitz, M. S., & Appelbaum, P. S. (2017). Beneficial and detrimental effects of genetic explanations for addiction. *International Journal of Social Psychiatry*, *63*(8), 717-723.

Lebowitz, M. S., & Ahn, W. K. (2018). Blue genes? Understanding and mitigating negative consequences of personalized information about genetic risk for depression. *Journal of genetic counseling*, *27*, 204-216.

Lebowitz, M. S., & Appelbaum, P. S. (2019). Biomedical explanations of psychopathology and their implications for attitudes and beliefs about mental disorders. *Annual Review of Clinical Psychology*, *15*, 555-577.

Lebowitz, M. S., Dolev-Amit, T., & Zilcha-Mano, S. (2021). Relationships of biomedical beliefs about depression to treatment-related expectancies in a treatment-seeking sample. *Psychotherapy*, *58*(3), 366.

Loughman, A., & Haslam, N. (2018). Neuroscientific explanations and the stigma of mental disorder: a meta-analytic study. *Cognitive Research: Principles and Implications*, *3*(1), 43.

Magliano, L., Ruggiero, G., Read, J., Mancuso, A., Schiavone, A., & Sepe, A. (2020). The views of non-psychiatric medical specialists about people with schizophrenia and depression. *Community mental health journal*, *56*, 1077-1084.

Marsh, J. K., & Romano, A. L. (2016). Lay judgments of mental health treatment options: The mind versus body problem. *MDM policy & practice*, *1*(1), 2381468316669361.

Miresco, M. J., & Kirmayer, L. J. (2006). The persistence of mind-brain dualism in psychiatric reasoning about clinical scenarios. *American Journal of Psychiatry*, *163*(5), 913-918.

Nettle, D., Frankenhuis, W. E., & Panchanathan, K. (2023). Biology, society, or choice: How do non-experts interpret explanations of behaviour?. *Open Mind*, *7*, 625-651.

Peters, D., Menendez, D., & Rosengren, K. (2020). Reframing mental illness: The role of essentialism on perceived treatment efficacy and stigmatization. *Memory & Cognition*, *48*(8), 1317-1333.

Phelan, J. C. (2005). Geneticization of deviant behavior and consequences for stigma: The case of mental illness. *Journal of health and social behavior*, *46*(4), 307-322.

Phillips, J., Morris, A., & Cushman, F. (2019). How we know what not to think. *Trends in Cognitive Sciences*, 23(12), 1026-1040.

Pilgrim, D. (2002). The biopsychosocial model in Anglo-American psychiatry: Past, present and future?. *Journal of Mental Health*, *11*(6), 585-594.

Proctor, C. C. T. (2008). *Clinicians' and laypeople's beliefs about the causal basis and treatment of mental disorders*. Yale University.

Schroder, H. S., Devendorf, A., & Zikmund-Fisher, B. J. (2023). Framing depression as a functional signal, not a disease: Rationale and initial randomized controlled trial. *Social Science & Medicine*, *328*, 115995.

Schroder, H. S., Duda, J. M., Christensen, K., Beard, C., & Björgvinsson, T. (2020). Stressors and chemical imbalances: Beliefs about the causes of depression in an acute psychiatric treatment sample. *Journal of Affective Disorders*, *276*, 537-545.

Statham, G. (2020). Normative commitments, causal structure, and policy disagreement. *Synthese*, *197*(5), 1983-2003.

Sytsma, J., Livengood, J., and Rose, D. (2012). "Two Types of Typicality: Rethinking the Role of Statistical Typicality in Ordinary Causal Attributions." *Studies in History and Philosophy of Science Part C*, 43: 814-820.

Weine, E. R., & Kim, N. S. (2019). Systematic distortions in clinicians' memories for client cases: Increasing causal coherence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(2), 196.

Zimmerman, H., Riordan, B. C., Winter, T., Bartonicek, A., & Scarf, D. (2020). Are New Zealand psychology students more susceptible to essentialist explanations for mental illness? Neuroessentialism and mental illness stigma in psychology and non-psychology students. *New Zealand Journal of Psychology*, *49*(3), 16-22.