Problems and Prescriptions in Psychiatric Explanation

Sera Schwarz
Yale University
Department of Philosophy
451 College St., New Haven, CT 06511
sera.schwarz@yale.edu

Abstract

A growing body of research suggests that different kinds of explanations of mental illness can have very different effects on their audiences' downstream 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 models that focus exclusively on changing causal or metaphysical judgments. On the normative 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, we can often combine different causal explanations, but it's not clear that we can as happily combine different characterisations of a problem. This might explain why philosophers and clinical experts sometimes seem to regard different explanations of mental illness as 'competitive', despite their appreciation for the causal complexity of psychiatric conditions.

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 licence or a philosophy degree to recognise 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. Although people don't always have a rich philosophical vocabulary for defending this, most of them know that the aetiology of a mental illness tends to be more complicated than 'your genes made you feel it'. Clinicians and researchers clearly recognise this, as evidenced by their increasingly impassioned calls for 'holistic' or 'biopsychosocial' approaches to mental illness (Gask [2018]; Davies et al. [2020]; Bolton [2023]), and their growing attentiveness to the varieties of causal complexity exhibited in psychiatry (Cooper [2001]; Kendler [2008]; Ross [2023]). But even non-experts often talk about the importance of different explanatory factors—say, childhood trauma, neurotransmitter levels, and personality traits—in a single breath.

A substantial body of empirical literature, however, seems to tell a strikingly different story. Across a range of correlational and experimental 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 exert a surprising degree of influence on their downstream reasoning about mental illness (for a review of the evidence, see Lebowitz and Appelbaum [2019]; Haslam and Kvaale [2015]). In other words, we now have strong evidence for the significance of 'explanatory framing effects' in psychiatry. By this I mean that the particular kinds of explanation of mental illness people focus on—the explanatory 'frame' in terms of which psychiatric symptoms are characterised—appears to bear *significant, systematic*, and seemingly *unwarranted* effects on many of their other beliefs, attitudes, and behaviours. What is stranger still is that these effects don't seem to affect only the proverbial man off the street. They also emerge in studies of people with first-hand experience of psychiatric symptoms, as well as expert psychiatrists and psychologists (see, e.g., Ahn et al. [2009]; Lebowitz et al. [2021]).

In this paper, I'll argue that these effects are even more puzzling and philosophically interesting than first meets the eye. I'll also suggest a new way of making sense of them. But my route to this conclusion will be somewhat unconventional. After reviewing some of the empirical research on explanations of mental illness, and raising some specific questions about its results, I'm going to step back and reflect on the nature of explanation more generally. I'll argue that explanations—in psychiatry, but also elsewhere—don't serve only to convey information about the causes of an outcome. At least sometimes, they have an intrinsically normative function: they serve to characterise an outcome as representing a particular kind of *problem* or *issue*.

This way of thinking about explanation is, I think, compelling in its own right, and inherits additional credibility from recent work on norms in causal reasoning. But it also neatly illuminates the explanatory framing effects we observe in psychiatry. If

explanations are in the business of defining problems, people's changing inferences about mental illness across explanatory contexts will not seem so strange: they can be understood as reasonable responses to changing information about the kind of 'wrongness' a mental illness represents. If I am right, however, these effects need not reflect something unique about the way we think about mental illness. They might instead reflect something much deeper about the nature of explanation, and the role it can play in our cognitive economy.

1. 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 effects to have emerged in this literature. These can, for present purposes, be grouped into three basic categories. There are, first, studies that supply evidence for changes to *prognostic* reasoning associated with receiving different explanations of mental illness; second, studies that investigate the different inferences that people make about appropriate *interventions*; and, third, studies that supply evidence for changes to the valence and strength of various *interpersonal attitudes*.

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 (Lebowitz and Appelbaum [2019]). One influential meta-analysis, which reviewed data from 28 experimental studies, yielded evidence for a significant link 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' (Kvaale et al. [2013]).[1] A number of studies conducted since have further corroborated these results (see, e.g., Lebowitz et al. [2013]; Haslam and Kvaale [2015]; Loughman and Haslam [2018]; Lebowitz and Ahn [2018]; Zimmerman et al. [2020]).

One particularly striking feature of this literature, and one to which I will return at

[1]A meta-analysis of correlational studies yielded similar results (Kvaale et al. [2012]). I borrow the term 'prognostic pessimism' from Lebowitz and Appelbaum's ([2019]) review of the data.

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 similar ways to information about the biological bases of mental illness. You would certainly not expect information about biological causes to consistently *dampen* their outlooks on the likely course of these conditions.[2] Current research suggests, however, that prognostic pessimism emerges as either an effect or correlate of biological explanations not just in samples of laypeople (Phelan [2005]; Bennett et al. [2008]), but also among people experiencing psychiatric symptoms (Lebowitz et al. [2014]; Gershkovich et al. [2018]; Lebowitz et al. [2021]), and even some clinicians (Magliano et al. [2019]).

When people with anxiety symptoms were given genetic or neurobiological explanations of panic disorder, for example, they were more likely to think that a person with this 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 (Lam and Salkovskis [2007]). Similar effects have been observed among people with a diagnosis of generalised anxiety disorder when given biological explanations of GAD (Lebowitz et al. [2014]). People with symptoms of depression were also less confident in their ability to recover when given sham 'evidence' for a genetic predisposition to MDD, which presumptively explained their symptoms (Lebowitz and Ahn [2018]; Kemp et al. [2014]).[3]

Experts probably aren't immune from the pessimism effect, either. Although there has not yet been systematic research directed specifically at assessing prognostic pessimism among mental health professionals, a recent study 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 (Magliano et al. [2019]). We also have evidence that psychiatrists, psychologists, and social workers who endorse biological explanations for a mental illness are more likely to believe that recovery will require medication, and are less optimistic about the potential efficacy of psychotherapy (Ahn et al. [2009]; Lebowitz and Ahn [2014]). This suggests that biological explanations are associated with pessimism about at least some routes to recovery from psychiatric conditions.

[2]You might even think that this information could help to *mitigate* pessimism, at least among participants that are inclined to trust in the promise of biomedicine. 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 can and often do develop targeted, evidence-based treatments. This is a line often echoed in calls to move beyond the DSM; see e.g. (Insel et al. [2010]). But of course much will hang here on people's background beliefs.

[3]See (Schroder et al. [2020]) for correlational evidence of this effect in a sample of inpatients.

B. Influence on reasoning about interventions

Another important dimension across which different kinds of explanations seem to have a significant differential effect concerns judgments about appropriate *interventions*. Converging lines of evidence suggest that members of the general public, people with psychiatric symptoms, and even expert clinicians tend to reason differently about treatment options for a mental health problem in response to being presented with 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 given broadly biological explanations of clinical symptoms, participants in various studies were more likely to prefer treatment by medication over psychotherapy; but they made the inverse judgment when the same or similar symptoms were explained psychosocially or environmentally (Proctor [2008]; Deacon and Baird [2009]; Marsh and 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 a set of symptoms in light of, say, biological explanations, they don't simply judge that pharmacological interventions 'make more intuitive sense' than psychotherapy: they explicitly predict that medication will be more *effective* than psychotherapy, and that psychotherapy will be *less effective in general*. For example, one study found that participants presented with genetic explanations for alcohol use disorder or gambling disorder believed 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 (Lebowitz and Appelbaum [2017]; see also Lebowitz et al. [2021]). A similar pattern emerges when people are provided psychological explanations of a clinical vignette: they tend to say, in such cases, that psychotherapy is a more credible and more effective intervention than medication (Iselin and Addis [2003]).

Crucially, this preference for 'explanation-congruent interventions' does not arise only in specific populations. It recurs in studies of lay audiences (Marsh and Romano [2016]; Deacon and Baird [2009]; Iselin and Addis [2003]), people with clinical symptoms (Lebowitz et al. [2021]; Iselin and Addis [2003]), and—strikingly—even expert clinicians (Ahn et al. [2009]; Lebowitz and Ahn [2014]). So the inferences it reflects 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 and attitudes towards persons who experience mental illness. 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. For example, in one influential study (Miresco & Kirmeyer [2006]), 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'.[4]

A similar effect has been reproduced in clinical samples. Two independent studies found that explaining depression to people with depressive symptoms by appeal to 'chemical imbalances' diminished their self-blame, but also weakened their perception of their own agency with respect to recovering from or even managing their symptoms (Deacon and Baird [2009]; Kemp et al. [2014]). The same pattern has emerged in samples from the general public: people provided genetic rather than non-genetic explanations of a person's psychological condition, for instance, tended to reduce both their ascriptions of blame and their general ascriptions of agency and self-control (Lebowitz and Appelbaum [2017]). In one striking experiment, healthy participants even rated *themselves* as less able to control their drinking when they were told—baselessly—that they had a genetic predisposition to alcoholism (Dar-Nimrod, Zuckerman, and Duberstein [2013]).

Many other broadly interpersonal judgments seem to be modulated by different explanations of mental illness. In one particularly unsettling study, mental health clinicians reported feeling less empathy for hypothetical patients when their symptoms were explained biologically rather than psychosocially (Lebowitz and Ahn [2014]). In fact, this effect persisted even when *both* biological and psychosocial explanations were provided, so long as the biological information was foregrounded. Other research suggests 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, and Kim [2003]; Kim, Paulus, Gonzalez, and Khalife [2012]; Weine and Kim [2018]). These effects fit neatly with a number of experimental and correlational studies that link biological explanations of mental illness with a greater endorsement of negative stereotypes, including heightened perceptions of people with psychiatric conditions as potentially unpredictable or dangerous (Haslam and Kvaale [2015]; Angermeyer et al. [2018]; Baek et al. [2022]).

2. What's so strange about framing effects?

[4]Intriguingly, judgments of psychological and neurobiological etiology were themselves inversely correlated. This seems to suggest that clinicians perceived these bases as 'competitive', at least to some degree. I discuss further evidence for this effect in section 3.

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 catastrophising—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 think that the empirical data, when taken together, present a picture that is much stranger than this simple analysis would suggest. One way to get a sense for this is to notice that many of the specific effects that have emerged in the empirical literature should seem *unwarranted*. Whether they are laymen or experts, participants in these studies tend to draw conclusions that go well beyond the information actually presented to them. And, at least on their face, many of these conclusions seem to involve rather odd inferential leaps. For example, we've seen that, when participants learn that a psychiatric condition has some broadly biological causes, they tend to think that it is likely to be especially severe, or that it can only be managed by medication. But these inferences are hardly justified by the mere discovery that there is a causal story to be told about these conditions at the level of biology.[5] After all, it's commonly (though not uncontroversially) assumed that *most* mental states have complex biological causes.[6] But it certainly doesn't follow that all undesirable mental states will face a poor prognosis, or call for a biological intervention.

In their discussions of these effects, many researchers have converged upon a common hypothesis about why they occur. What the evidence suggests, these researchers claim, is that people tend to reason about mental illness in light of instinctive and largely unreflective metaphysical intuitions. One especially common proposal is that people's

[5]You might think that participants aren't responding to information about biological causes *alone*, but are rather inferring something about relative causal relevance (say, on the basis of pragmatic norms) from the fact that these particular causes are the ones presented to them. In this way, they might infer that, e.g., 'a biological cause is the most important cause in a causal hierarchy'. Alternatively, people might be reasoning in light of the fact that biological explanations are often presented in contexts in which biological interventions are pursued. I'm happy to accept these possibilities, which I take to be generally compatible and even complementary with the account I pursue in this paper. My point at this stage is simply that people do seem to reason beyond the immediate evidence.

[6]Of course, this is not to say that they *only* have biological causes.

judgments in these studies are guided by covert dispositions towards mind-body dualism or causal essentialism (Haslam [2000]; Marsh and Romano [2016]; Ahn, Kim, and Lebowitz [2017]; Loughman and Haslam [2018]; Buckwalter [2020]; Berent and Platt [2021]; see also Bloom [2004]). This 'intuitive metaphysician' hypothesis could explain, for instance, why participants seem to attribute less agency to people with biologically explicable mental health conditions, and to assume that only medication would really help them. If one were disposed to think that mental illnesses have causal essences, and that 'mental' and 'physical' causal processes are deeply distinct, these inferences might naturally follow from learning that a given condition has biological causes.

Maybe this proposal is onto something: maybe deep-seated dualist or essentialist intuitions do often quietly guide people's thinking and jam up their judgments. As it turns out, however, there is surprisingly little independent evidence for the impact of such intuitions—and recent experimental studies that have sought to capture the influence of essentialist biases, in particular, did not find the effects we observe in the wider literature on framing effects in psychiatry (Peters et al. [2020]). But even if we assume that ordinary people are inclined to think in dualistic or essentialist terms, it seems to me that this does not yet yield a fully satisfying general-purpose account of the data. This is because the balance of the evidence suggests, as we've seen, that the patterns of judgments we find in laypeople are surprisingly similar—at least in their broad contours—to those exhibited by highly practised mental health professionals. And even if we are prepared to accept that ordinary people are unwitting dualists or causal essentialists, I think it should be much more difficult to accept that this would be true of clinical experts.

For instance, in the studies reviewed above, even expert psychiatrists and psychologists seemed to think that mental health conditions that could be explained biologically would face especially poor prognoses, and could only effectively be managed by biological interventions. And they made the inverse inference, *mutatis mutandis*, when they received psychological or environmental explanations of these conditions. They also seemed to think that, if a biological factor could explain some set of symptoms, psychological or environmental factors couldn't very successfully explain it, and couldn't be efficiently leveraged in order to treat it. In other words, they appeared to reason as though explanations and interventions were implicitly 'competitive', at least to some degree: the availability of a good biological explanation or intervention led them to assume that other kinds of explanations or interventions were *less plausible* or *less viable* (see, e.g., Ahn et al. [2009]; Miresco and Kirmeyer [2006]; Lebowitz and Ahn [2014]).[7]

[7]Notably, however, experts do not draw this inference in every case in which multiple explanations are available. For instance, in the (Ahn et al. [2009]) studies, clinicians were happy to say that a given psychiatric symptom or condition could simultaneously have strong psychological and strong social bases; but they did not seem to think that it could have both a strong psychosocial basis and a strong biological basis. I think

But this should all seem very puzzling. Surely 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 complex 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]; Bolton and Gillett [2019]). And we know that practised clinicians tend to endorse such principles. They clearly understand, for instance, that mental illnesses usually don't have a single cause (Ahn et al. [2009]), that different causal explanations are often complementary (Harland et al. [2009]; Proctor [2008]; Brog and Guskin [1998]), and that effective treatments need not target any particular causal pathway (Ahn et al. [2006]). And they often clearly disavow dualist and essentialist views. For example, (Ahn et al. [2006]) found that clinicians generally resisted the claim that mental illnesses have causal essences, and tended to believe that, even if a mental illness *did* have one basic kind of cause, effective interventions would not necessarily need to target it.[8]

Research on explanatory framing effects has, moreover, sometimes been designed to actively dampen possible essentialist or dualist intuitions. In the studies by (Ahn et al. [2009]), for instance, clinicians were repeatedly told to reason about psychiatric etiology in a non-reductive, pluralistic way.[9] But, in spite of this explicit guidance, many clinicians' judgments continued to bear striking similarities to those found in laypeople. Although their considered views on mental illness were surely far more subtle and complex, they still seemed inclined to think that conditions with a biological basis would not be very effectively explained by psychosocial factors, and could not be very effectively treated by psychosocial interventions.

Of course, if the 'intuitive metaphysician' hypothesis is correct, it's possible that experts are subject to biases that are simply disconnected from their conscious beliefs, in the way that other powerful kinds of implicit bias, like gender or racial biases, often seem to be recalcitrant to correction. Some researchers have doubled down on

---

this finding is consistent with the problem-based account I provide below: psychological and social problems are usually of a very similar type, and make reference to a very similar normative vocabulary. Note also that this result complicates a potential analysis that would hinge on participants simply selecting for the single *most relevant* cause. If psychological and social causes can both be judged relevant, it's not clear why this wouldn't hold for e.g. biological and social causes.

[8] Ahn and colleagues emphasise that even *novices* did not strongly endorse essentialist views about mental illness (Ahn et al. [2006], p. 766). See also (Kendler [2005]) and (Novick and Ross [2020]) for important theoretical defences of anti-essentialism in psychiatry.

[9] In fact, at almost every stage of the experiment, participants were reminded that 'biological, psychological, and environmental causes [are] *non-mutually exclusive domains that could be overlapping*'. (In a free recall task in a follow-up study, clinicians happily listed an average of 5.4 potential causes, many from very different domains, for various mental health conditions.) But participants still seemed to think conditions with a strong biological basis would *not* have a very significant psychological or environmental basis, and vice versa.

appeals to the power of covert metaphysics in exactly this way.[10] In their review of the literature for the *Oxford Handbook of Causal Reasoning,* for example, Ahn, Lebowitz, and Kim ([2017], p. 12) explicitly warn that dualist intuitions can generate a felt tension between biological and psychological explanations 'strong enough to lead to an irrational bias in both clinicians and laypeople'. And, if implicit biases are indeed responsible for the effects we find, perhaps they are right to sound this cautionary note.

On reflection, however, I think this analysis can come to seem prematurely pessimistic. If even experts consistently exhibit a distinctive pattern of judgment, it seems like good interpretive practice to wonder whether something deeper than 'irrational biases' might account for it. Although appealing to illicit biases can often explain otherwise confusing data, it can also, perhaps just as often, obscure richer and more charitable possible interpretations of the evidence. And I think there are such alternatives available to us. I now want to consider a novel analysis of just this kind. I will call this the 'normative model' of explanatory framing effects.

3. Causes, Norms, and Explanations

Before I start filling out the normative model, it will be helpful to think for a moment about causal explanations more generally.[11] In particular, I want to take a quick step back to consider what it is we are really *doing* when we explain why things happen. One intuitive answer to this question—so intuitive, in fact, that you might not think there are viable alternatives—is that we explain things simply in order to share causal information. In other words, we seek explanations simply 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—at least when it comes to non-scientific explanation—the 'causal communication' picture is *anemic*, even if otherwise correct. There are important and even essential features of ordinary explanations that it just doesn't capture. One way of getting a feel for this is by reflecting on a fact very familiar to philosophers of science, which is that explanations are always selective and partial. They never describe *all* the causes of an event: instead, they filter down facts about general causal structure in light of

[10] Miresco and Kirmeyer ([2006], p. 916), for example, note that one-third of participating clinicians correctly guessed the aim of the study—namely, to assess the influence of dualist intuitions—without this exerting any perceptible effects on their judgments. They conclude that 'psychiatrists continue to operate according to a mind-brain dichotomy in ways that are often covert and unacknowledged' (p. 913).

[11] I suspect that much of what I say in this section will also apply to many *non-causal* explanations. But because I'm principally concerned with ordinary causal explanations in this paper, I've elected to emphasise this here.

[12] Of course, there are usually further constraints placed on the particular kinds of causal facts that can be properly explanatory. But these complexities need not detain us here.

various communicative and pragmatic norms relevant to a context (Bromberger [1966]; van Fraassen [1980]; Woodward [2003], [2021]; Longino [2013]; Potochnik [2017]).

Research suggests, however, that our explanatory practices are shaped by implicit norms in even deeper ways than this. In particular, we now have a great deal of evidence that people's causal ascriptions—and so, presumably, their causal-explanatory judgments—tend to be influenced, often very surprisingly, by considerations of moral valence, moral responsibility, statistical normality or abnormality, and norms of proper functioning (see, e.g., Kahneman and Tversky [1982]; Alicke [1992]; Alicke et al. [2011]; Hitchcock and Knobe [2009]; Icard et al. [2017]; Kirfel and Lagnado [2018]; Kirfel et al. [2024]; Statham [2020]; Sytsma et al. [2012]). For instance, when 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 caused what happened. 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 (Hitchcock and Knobe [2009]).

In this way, people in search of explanations seem to reason in light of 'normalising' 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. Of course, 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. But one very convincing proposal highlights the important role that explanations play in guiding our future action. The basic thought here is that, 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'. Focusing on abnormal events in our explanations therefore 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.[13]

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 our judgments of normality, suitability, and relevance.

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

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* stories. If the exam was on a particular text that my friend did not know much about, both facts about the exam's design, 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 been constructed differently, *or* if my friend had instead mastered the Doctrine of the Concept, they would have passed rather than 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 her 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', and so on). 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' the exam result represents.[14]

[14]You might think that this gloss only works because 'failing an exam' (much like 'having a mental health issue') is *intrinsically problematic*. But what about events that are non-problematic, or indeed positively valenced? For the moment, I would note only that studies of linguistic corpora suggest that ordinary talk of causation tends to be correlated with negatively valenced words like 'problem' or 'damage' (e.g., Stubbs [1995]; Glass [2023]). This provides some reason to think that the problem-based analysis of causal explanation is at least very *widely* applicable. But see footnote 17, below, for more on the question of non-problematic explananda.

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 here 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. When we ask 'what was the problem that led to this outcome?', we are asking for a normative picture of how a good outcome should have been brought about. But even the full set of causal and historical facts can be 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 do not mean to convey that the exam should have been on the *Wissenschaftslehre* instead.) They're rather saying something like 'think about qualifying 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 these exams in light of particular students' preparedness.'

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.' In this way, the presentation of a problem can influence our thinking about a situation very generally, by shifting our focus to specific kinds of default states, variables, relationships, and standards of assessment in our further thinking about and acting upon it. When we point to a problem, as when we point to an explanation, we are pointing to those features of a situation that we think are most important for really understanding and exploring it. This pointed focus is obviously useful, because it organises our interactions with an outcome of interest: it ensures that we reason about and respond to it in a way that is guided and systematic, rather than chancy and haphazard.

This leads me to one last and especially critical point. This is that the 'narrowing of normative focus' characteristic of problem-based reasoning—that is, the pressure it puts on us to regard an outcome as representing a specific kind of wrongness or strangeness—naturally leads us to think in terms of there being some kind of 'root' problem at issue in any particular case. Although I can't defend this point in fullness

here, I think it's an important fact about our ordinary way of thinking about problems that we often talk about 'the real problem' with a situation (as opposed to any number of problems it might happen to exhibit), and the importance of finding 'real solutions' to it (as opposed to the mere 'quick fix'). What this suggests is that coming to understand a problem often involves coming to understand an outcome as having a kind of normative 'core'. This is why it is so natural to think, once we take an outcome to represent a particular kind of problem, that we can only really understand it by thinking about it in a particular way—namely, in light of the particular wrong-making or strange-making features that make it the problem it is. And it is why we think that we can only fully respond to it by changing these features *in the right way*—by bringing them back in line with the right norms. When we have a sense of the 'real problem', we will often regard features and interventions that seem unrelated to it, or inappropriately related to it, as much less relevant to address.[15]

This is where talk of problems can start to seem importantly different from talk of causes.[16] It is fairly easy to think about an event as having various different causes. In fact, we 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 exam result is caused by facts about both their preparation and the exam-writers' decisions about which questions to ask them. But once I think of the 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 students in a fair and methodologically well-grounded way; so the real solution would be to ensure that it does. This would be true even if we could reliably change the outcome in other ways (say, by giving advance warning of the examiners' Hegelomania, or providing a cheat sheet, or allowing for do-overs). Although these would, in some sense, all be effective interventions, they would only be 'quick fixes'. They would not address the real issue.

If this is right, it might explain why many people, including many philosophers, often seem to regard different explanations of at least some outcomes as difficult to happily combine, despite their appreciation for the complexity of causal history. When they say things like 'if this explanation is a powerful explanation for O, this other explanation can't be very powerful!', they might not be calling out to us from the depths of explanatory chauvinism. Similarly, when they say 'only this intervention

[15]This seems to fit quite neatly with discussions of 'discounting principles' in the psychological literature on lay causal explanation and attribution theory. For example, Kelley's ([1971], [1972], [1973]) covariation model likewise predicts that, when people come to regard one explanatory factor as relevant or plausible, they will tend to regard other potential factors as *less* relevant.

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

would be *really* effective', they might not be experiencing sudden amnesia with respect to all other possible interventions. They might, instead, simply be appropriately responding to the claim to normative priority implicit in our judgments about 'real problems', and the 'real solutions' that would rightly resolve them.

To the extent that this analysis does indeed capture an important dimension of explanatory reasoning, many further questions will immediately arise. For instance, you might wonder about the precise source of these effects: do judgments about core problems fundamentally guide judgments of *causal* relevance, and only derivatively judgments of explanatory relevance, or does the line of influence run the other way around? You might also wonder about their scope of application. Do these judgments only inform reasoning about outcomes that are explicitly *problematic*, or might they also be operative in other contexts?[17] Are they restricted to causal explanations, or indeed 'folk' causal explanations, or could they be relevant more generally? Would they have any bearing on explanatory projects in the sciences?

These questions are deep and difficult, and so I cannot hope to address them adequately here. But although they point to many different ways of further filling out the normative model, I think that, when it comes to the issues with which this paper is principally concerned, the basic picture upon which they draw can already illuminate a great deal.

---

[17] In other words, would a similar kind of analysis work if we were dealing with (a) an ordinary explanation of some humdrum event, or indeed (b) the stable, regular kinds of phenomena usually studied in scientific contexts? This question is important, not least because one would expect clinicians to be far more familiar with scientific forms of explanatory reasoning than laypeople. But I can imagine at least two different possible paths to answering this question in the affirmative. First, you might think that representations of problems are a special case of representations of 'strangeness' or 'unpredictability', such that, in non-problematic contexts, explanations will tend to center instead on a feature of the explanandum that is non-normative in the sense of 'not otherwise predictable'. Alternatively, you might think that unpredictability is itself a kind of problem—not, perhaps, with the explanatory objects themselves, but rather with our epistemic relationship to them. This line derives some plausibility from the fact that we typically seek explanations, at least (but perhaps not only) in the ordinary case, precisely because something seems to be veering off an expected course. It is usually only when expectations break down that we find ourselves puzzled, and ask searching questions about why something is the case (e.g., the skies have suddenly turned stormy, or I have begun to feel ill; or perhaps a winter day turns sunny, or I make a sudden recovery). Although these unexpected happenings often have a *positive* rather than *negative* valence, I think we can understand them as representing—at minimum—a generic kind of problem: namely, something's being odd or unexpected.

In this way, any deviation from a normal state can be 'problematic' qua interruption of the ordinary course of things, though we can and usually do reach for more specific and stylised varieties of problems in ordinary cases. This gloss on the relationship between explanatory reasoning and representations of problems seems to align neatly with recent work on linguistic corpora (Glass [2023]), which finds that ordinary talk of causation tends to be correlated with 'negative-sentiment complements', including references to 'problems'. It might also begin to clarify how the normative model could accommodate at least some scientific explanations. When I seek to explain (e.g.) why salt dissolves in water, I am at least pursuing a solution to the problem of inexplicability: 'why is this solid stuff, which I expect to endure through spatial relocation, suddenly disappearing?' This generic kind of 'problem' is effectively invisible, but I suspect this is only because it is (unusually) thin and uncontroversial.

4. Problems and prescriptions in psychiatric explanation

Let's return, then, to explanations of mental illness. My angle on this should now seem fairly clear: 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 the normative model can help us make sense of these effects not only from a diagnostic perspective, but also from a *rationalising* 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.

Recall that the normative model invites us to think about explanations as encoding information about problems, or divergences 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 psychiatric conditions are paradigmatically centered on problems, whether these are conceived as 'brain-circuit disorders' (Insel [2009]), 'problems of living' (Szasz [1960]), or something in between. And it's obvious that there are many ways of understanding what kinds of problems these are, and what kind of 'unproblematic' states they should be contrasted with. This is, in fact, a concern that has consistently driven criticism of the medical model, including recent challenges from defenders of the neurodiversity framework (Chapman and Carel [2021]; Kukla and Williams [2024]), the 'Mad Pride' movement (Rashed [2018]), madness-as-strategy paradigms (Garson [2022], [2023]), and approaches informed by the philosophy of disability (Hogan [2019a]).[18] Such critics have often argued that the medical model simply assumes—unduly—that the problems associated with mental health conditions are to be understood in terms of internal disorders or dysfunctions. But if there are viable alternatives to this view, deep questions quickly arise about the metaphysical and ethical stakes of conceiving of these conditions as personal medical problems, rather than as, say, 'the product[s] of an unaccommodating and oppressive society' (Hogan [2019b], p. 16).

You might think, in light of these ongoing controversies, that we have good reason to remain devoutly agnostic about the deeper nature of psychopathology. Amid such thorny philosophical thickets, it can seem prudent to leave the bigger questions for the experts to hash out. But if the normative model is right, it will be extremely difficult to

[18] Although philosophers and activists associated with these movements sometimes argue that (at least some) mental health conditions are not problems at all, I take most of them to be committed to the more moderate view that—to the extent that symptomatic or diagnosed persons do suffer or struggle—we ought to understand this not in terms of *personal* problems, but rather social, cultural, or material problems. This raises the interesting possibility that even framing a psychiatric condition as something to be explained in terms of a *disorder*, without the specification of biological facts, might lead one to infer that it represents a problem internal to an individual, and consequently to prefer local, personal interventions.

divide up the intellectual labor in this way. This is because the model suggests that we are often implicitly coming down on these questions, even in the apparently innocuous activity of giving and receiving explanations of psychiatric symptoms. In other words, it predicts that our choices about 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.

This suggests 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 potentially conveying that Sally's problem—call it 'depression'—is really a neurotransmitter problem. If, however, I explain Sally's depression by reference to cognitive traits (say, 'Sally is depressed because she ruminates'), the 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 fundamentally a cognitive one. Similarly, if I say 'Sally is depressed because she's been out of a job all year', I will be suggesting that her depression is a problem of economic precarity—in other words, that it is basically a social or environmental problem.

This way of analysing the impact of different explanatory claims should seem very intuitive. It also illuminates the intuitive basis for an assumption which should seem otherwise odd, especially for psychiatrists: namely, the assumption that explanations of mental illness can *compete* for explanatory power. We've seen that clinicians in experimental studies sometimes seem to think that a condition that has a good biological explanation cannot have a very good psychological explanation, even though they accept that other kinds of explanations (e.g., psychological and social explanations) can actively complement one another. But these same clinicians also think that biological and psychological causes of mental illness often work together. So why would they conclude that conditions that can be explained biologically are not well-explained otherwise? The normative model produces a simple answer: once a condition is conceived as a basically biological problem, explanations that do not make reference to biological norms, or invoke features relevant to these norms, will tend to seem much less apt.

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 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. Why would even experts think that interventions can only be effective if they target particular kinds of causes?

Broad clinical consensus explicitly militates against this.

The normative model, however, makes this preference for congruent kinds of interventions much more comprehensible. This is because it is quite reasonable, as we have seen, to think that a real solution to a particular kind of problem—that is, a solution that *genuinely resolves* it—will address the very features that made it problematic to begin with. If I tell you, for example, 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 simply start to schedule my meetings in the evenings. This would seem to be a neat and effective intervention upon an undesirable outcome. But I think it would, just as clearly, not represent a 'real solution' to the central issue.

Real solutions are interventions that change an outcome in precisely the right way, rather than in any old way. And this is something we care a lot about. Think here of the familiar language of 'first-line' vs 'adjunctive' treatments, and the common assumption that adjunctive treatments will only facilitate or mitigate the side-effects of a 'preferred' solution.[19] Or think of the familiar charge that an intervention is just a 'band-aid' or a 'stopgap', and the deep skepticism and disdain that these epithets convey about proposed responses to a problem (e.g., depression, income inequality, or racial prejudice). One reason these charges are so powerful, I think, is that they often don't dispute the *possibility* of implementing many different kinds of interventions upon an outcome. What they dispute is the normative *appropriateness* of pursuing them in lieu of a real solution to the issue, which would aptly and non-incidentally correct for it.

The importance of this distinction is very clear in the case of mental illness. If, for example, I think that anxiety is fundamentally a biological problem, it seems very reasonable to think that I can only get a 'real fix' by addressing the specifically biological wrongness at issue: the correct, 'first-line' course of action is to bring the anxious person back in line with biological norms. Cognitive behavioural therapy might then seem to be only a 'coping mechanism', or perhaps even a temporary 'band-aid' to wear until the real treatments kick in. But 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

[19] See, for example, Miklowitz et al.'s [2020] case for the 'adjunctive' status of psychotherapy in the treatment of bipolar disorder—that is, its serving as a form of therapy that is potentially augmentative, but not itself necessary. Similar such examples are easily produced.

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, quite odd. Why should simply learning that a mental illness has biological causes 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, however, if biological explanations of mental illness communicate that mental illness is a fundamentally *biological problem*. After all, problems—unlike mere causal factors—are the kinds of things that we feel compelled to resolve, rather than simply identify or manipulate. And pessimism is a perfectly natural response to encountering problems that we cannot get a meaningful normative grip on. 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 psychiatry, 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 Dave is depressed because he has a 'neurotransmitter problem', or perhaps a 'genetic problem'. What can you now realistically infer about his prospects? 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 indeed whether it can be effectively resolved at all? Unless you happen to be a neurophysiologist or geneticist (and even if you are), these questions are likely to leave you somewhat bewildered. Most of us don't really know what 'neurotransmission problems' or 'genetic problems' involve, much less what might be done to meaningfully solve them (we certainly can't intervene directly on the genome!). So if we are told that Dave's suffering is a specifically neurological or genetic kind of problem, we probably won't feel very clear-eyed or hopeful about his prospects. We will likely find it difficult to envision how his problem could be meaningfully addressed.

We will, however, have a much easier time getting a firm and reassuring grip on problems of other kinds. For example, the category of psychological problems 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-practised 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 find such solutions. Much the same is true, I think, for most social, cultural, and environmental problems. Like psychological problems, these are the kinds of issues on which we already have a strong normative grip: we are well-equipped to understand why they're bad, and what might really

make them better. So we are less likely to feel pessimistic on principle when we learn that someone must grapple with them.

Finally, consider the **changes to interpersonal attitudes** associated with explanatory framing effects, such as changes to attributions of agency, stereotyping, and even baseline levels of empathy. It would be very odd, I think, if these effects were down to our simply learning that clinical conditions have biological causes. All human states have at least some biological causes; people with mental illness are no different in *this* respect. These results would make a great deal of sense, however, if they were tracking changing views of real problems. This is because some problems—for example, 'neurotransmitter problems'—are extremely difficult to think about as problems of more or less ordinary agents. They are conceptually so far removed from the ordinary vocabulary of interpersonal life that it can take real effort to integrate them into a familiar psychological or moral framework. (Even professional moral philosophers sometimes struggle to do so.) For this reason, we might have a hard time figuring out how to morally and rationally engage with people who wrestle with such problems. In the face of invitations to prediction and imaginative projection, we might find ourselves rudderless. Reaching for stereotypes would then be one strategy for compensating for this felt sense of uncertainty.

If, however, we think of someone's suffering in terms of broadly familiar psychological problems—say, issues 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, criticise or defend their behaviour, 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 precisely what we find in the literature. As we've seen, when people are given psychological explanations of psychiatric symptoms, they tend to prefer treatment by psychotherapy; they are less likely to think of people with these symptoms as deeply abnormal; they are more inclined to hold them accountable for their behaviour, and to assume that they can influence 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 would seem to follow from an invitation to attend to familiar, distinctively interpersonal problems, rather than those that are distinctly foreign.

5. Conclusion

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 interpreting the data as evidence only of recalcitrant metaphysical biases, or indeed irrational kinds of bias, is uncharitable to the point of implausibility—especially if alternative interpretations are available.

But there is a way of making sense of these data that does not require such uncharitable and implausible assumptions. If, as I have been arguing, explanations of psychiatric symptoms don't simply communicate facts about their causal history, but also motivate judgments about the 'real problem' they represent, the fact that people reason differently across explanatory contexts should no longer seem surprising. Different explanations of mental illness center different kinds of norms in our reasoning. This naturally informs a broad range of further judgments about psychiatric conditions, and what it would take to meaningfully address them.[20]

This analysis does not, however, only neatly account for some otherwise puzzling effects. It also captures the overall reasonability of the basic *mechanism* by which these effects are generated. Although a detailed defence of this claim will have to be left to another occasion, I think it should already be fairly easy to see how thinking in terms of problems can serve as a powerful 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 a sensitivity to 'real problems' attunes us to those of its features we really *should* care about, by attuning us to norms that can guide our general 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 optimised predictions or interventions.[21]

If this is right, representations of problems are *functional*. They are the means by which we get a handle on how generally to think about an outcome, especially in the face of plausible alternatives. But this, in turn, suggests that the problem with explanatory framing effects is not that people are unduly responsive to such

[20]Importantly, this analysis need not be restricted to psychiatric explanation. In fact, a recent series of experiments on the uptake of explanations of human behaviour *in general* produced effects very similar to those we observe in the context of mental illness (Nettle et al. [2023]). This neatly complements the broader 'problem-theoretic' hypothesis I've been exploring here.

[21]This marks an important difference between the normative model and more familiar accounts of norms in explanation. These accounts tend to emphasise the role of norms in guiding us to optimal interventions, often with the implication that explanations are directed at maximising their efficacy or reliability (Kirfel et al. [2024]). 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.

representations. It is rather that they do not usually recognise that this is what they are doing. In the context of psychiatry, however, this is an oversight we can ill afford. Whenever we are confronted with psychiatric conditions, we are confronted with difficult choices about how to think about them. Is the basic problem represented by my depression the fact that my neurotransmitters are out of whack, or is it rather that I have succumbed 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 these questions.[22] This is because satisfying answers to them are not decided by the causal facts; they are decided by our sense of what the facts should have been and should be. To endorse one answer over others is, in other words, to assume a kind of responsibility for adopting a particular normative view—and this is a responsibility that cannot be cleanly offloaded onto the scientist, the metaphysician, or the empirical data. How best to adjudicate this responsibility is an issue for another day. But, as with all such matters, the first point of business is to recognise that we are regularly summoned to judgment in this way. And here, I think, the normative model gives us just the resources we need.

Acknowledgements

I am deeply grateful to Josh Knobe for our many lengthy discussions about the issues addressed in this paper: the ideas presented here could not have been so carefully or so fruitfully developed without them. I am also grateful to Peli Grietzer, Daniel Greco, and Muhammad Ali Khalidi for their many insightful comments on earlier drafts; to members of the Knobe Lab at Yale University and participants in the *29th Annual Philosophy of Science Association (PSA) Conference* in New Orleans, USA for their encouragement—and their helpful provocations—when I first began workshopping these ideas; and, finally, to two anonymous reviewers, whose unerringly constructive engagement with this paper went well beyond the call of ordinary professional duties. Their remarks have significantly sharpened my thinking about the ideas taken up here; all remaining ambiguities, errors, and oversights are mine alone.

[22] At this juncture, you might want to deny 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* take such a line, but there are at least some reasons to resist it. Although I can't defend this claim in fullness here, I've already indicated that judgments about 'real problems' might serve an important functional role in our cognitive economy. In a slogan: we must focus on *something* in order to understand *anything*. In particular, we must know what exactly has gone wrong in order to know how it might generally go right. These claims can, at minimum, be defended on pragmatic grounds, although I suspect there is more to say on their behalf.

References

Ahn, W. K., Flanagan, E. H., Marsh, J. K. and Sanislow, C. A. [2006]: 'Beliefs About Essences and the Reality of Mental Disorders', *Psychological Science*, **17**, pp. 759–766.

Ahn, W. K., Proctor, C. C. and Flanagan, E. H. [2009]: 'Mental Health Clinicians' Beliefs About the Biological, Psychological, and Environmental Bases of Mental Disorders', *Cognitive Science*, **33**, pp. 147–182.

Ahn, W. K., Kim, N. S. and Lebowitz, M. S. [2017]: 'The Role of Causal Knowledge in Reasoning About Mental Disorders', in Waldmann, M. R. (ed.), *The Oxford Handbook of Causal Reasoning*, Oxford University Press, pp. 603–618.

Ahn, W. K., Novick, L. R. and Kim, N. S. [2003]: 'Understanding Behavior Makes It More Normal', *Psychonomic Bulletin & Review*, **10**, pp. 746–752.

Alicke, M. D. [1992]: 'Culpable Causation', *Journal of Personality and Social Psychology*, **63**, pp. 368–378.

Alicke, M. D., Rose, D. and Bloom, D. [2011]: 'Causation, Norm Violation, and Culpable Control', *The Journal of Philosophy*, **108**, pp. 670–696.

Baek, C. H., Kim, H. J., Park, H. Y., Seo, H. Y., Yoo, H. and 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,** e25, available at < https://doi.org/10.3346/jkms.2023.38.e25>.

Bennett, L., Thirlaway, K. and Murray, A. J. [2008]: 'The Stigmatising Implications of Presenting Schizophrenia as a Genetic Disease', *Journal of Genetic Counseling*, **17**, pp. 550–559.

Berent, I. and Platt, M. [2021]: 'Essentialist Biases Toward Psychiatric Disorders: Brain Disorders Are Presumed Innate', *Cognitive Science*, **45**, e12970, available at <https://doi.org/10.1111/cogs.12970>.

Bolton, D. [2023]: 'A Revitalized Biopsychosocial Model: Core Theory, Research Paradigms, and Clinical Implications', *Psychological Medicine*, **53**, pp. 7504–7511.

Bolton, D. and Gillett, G. [2019]: *The Biopsychosocial Model of Health and Disease: New Philosophical and Scientific Developments*, Palgrave Pivot.

Brog, M. A. and 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**, pp. 229–235.

Bromberger, S. [1966]: 'Why-Questions', in Colodny, R. G. (ed.), *Mind and Cosmos: Essays in Contemporary Science and Philosophy*, University of Pittsburgh Press, pp. 86–111.

Buckwalter, W. [2020]: 'Mind–Brain Dichotomy, Mental Disorder, and Theory of Mind', *Erkenntnis*, **85**, pp. 511–526.

Cooper, B. [2001]: 'Nature, Nurture and Mental Disorder: Old Concepts in the New Millennium', *British Journal of Psychiatry*, **178**, pp. 91–101.

Dar-Nimrod, I., Zuckerman, M. and Duberstein, P. R. [2013]: 'The Effects of Learning About One's Own Genetic Susceptibility to Alcoholism: A Randomized Experiment', *Genetics in Medicine*, **15**, pp. 132–138.

Davies, W., Savulescu, J. and Roache, R. (eds) [2020]: *Psychiatry Reborn: Biopsychosocial Psychiatry in Modern Medicine*, Oxford University Press.

Deacon, B. J. and Baird, G. L. [2009]: 'The Chemical Imbalance Explanation of Depression: Reducing Blame at What Cost?', *Journal of Social and Clinical Psychology*, **28**, pp. 415–435.

Garson, J. [2022]: *Madness: A Philosophical Exploration*, Oxford University Press.

Garson, J. [2023]: 'Madness and Idiocy: Reframing a Basic Problem of Philosophy of Psychiatry', *Philosophy, Psychiatry, & Psychology*, **30**, pp. 285–295.

Gask, L. [2018]: 'In Defence of the Biopsychosocial Model', *The Lancet Psychiatry*, **5**, pp. 548–549.

Gershkovich, M., Deacon, B. J. and 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**, pp. 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**, pp. 451–457.

Glass, L. [2023]: 'Using the Anna Karenina Principle to Explain Why Cause Favors Negative-Sentiment Complements', *Semantics and Pragmatics*, **16**, pp. 1–49.

Harland, R., Antonova, E., Owen, G. S., Broome, M., Landau, S., Deeley, Q. and Murray, R. [2009]: 'A Study of Psychiatrists' Concepts of Mental Illness', *Psychological Medicine*, **39**, pp. 967–976.

Haslam, N. and Kvaale, E. P. [2015]: 'Biogenetic Explanations of Mental Disorder: The Mixed-Blessings Model', *Current Directions in Psychological Science*, **24**, pp. 399–404.

Hitchcock, C. and Knobe, J. [2009]: 'Cause and Norm', *The Journal of Philosophy*, **106**, pp. 587–612.

Hogan, A. J. [2019a]: 'Moving Away From the "Medical Model": The Development and Revision of the World Health Organization's Classification of Disability', *Bulletin of the History of Medicine*, **93**, pp. 241–269.

Hogan, A. J. [2019b]: 'Social and Medical Models of Disability and Mental Health: Evolution and Renewal', *CMAJ: Canadian Medical Association Journal*, **191**, pp. 16–18.

Icard, T. F., Kominsky, J. F. and Knobe, J. [2017]: 'Normality and Actual Causal Strength', *Cognition*, **161**, pp. 80–93.

Insel, T. R. [2009]: 'Disruptive Insights in Psychiatry: Transforming a Clinical Discipline', *The Journal of Clinical Investigation*, **119**, pp. 700–705.

Iselin, M. G. and Addis, M. E. [2003]: 'Effects of Etiology on Perceived Helpfulness of Treatments for Depression', *Cognitive Therapy and Research*, **27**, pp. 205–222.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C. and Wang, P. [2010]: 'Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders', *The American Journal of Psychiatry*, **167**, pp. 748–751.

Kahneman, D. and Tversky, A. [1982]: 'On the Study of Statistical Intuitions', *Cognition*, **11**, pp. 123–141.

Kemp, J. J., Lickel, J. J. and Deacon, B. J. [2014]: 'Effects of a Chemical Imbalance Causal Explanation on Individuals' Perceptions of Their Depressive Symptoms', *Behaviour Research and Therapy*, **56**, pp. 47–52.

Kendler, K. S. [2005]: 'Toward a Philosophical Structure for Psychiatry', *American Journal of Psychiatry*, **162**, pp. 433–440.

Kendler, K. S. [2008]: 'Explanatory Models for Psychiatric Illness', *The American Journal of Psychiatry*, **165**, pp. 695–702.

Kelley, H. H. [1972]: 'Causal Schemata and the Attribution Process', in Jones, E. E., Kanouse, D. E., Kelley, H. H., Nisbett, R. E., Valins, S. and Weiner, B. (eds), *Attribution: Perceiving the Causes of Behavior*, General Learning Press, pp. 151–174.

Kim, N. S., Paulus, D. J., Gonzalez, J. S. and Khalife, D. [2012]: 'Proportionate Responses to Life Events Influence Clinicians' Judgments of Psychological Abnormality', *Psychological Assessment*, **24**, pp. 581–590.

Kirfel, L., Harding, J., Shin, J., Xin, C., Icard, T. and Gerstenberg, T. [2024]: 'Do As I Explain: Explanations Communicate Optimal Interventions', *Proceedings of the 46th Annual Conference of the Cognitive Science Society*, available at <https://doi.org/10.31234/osf.io/4kyfn>.

Kirfel, L. and Lagnado, D. [2018]: 'Statistical Norm Effects in Causal Cognition', in Rogers, T. T., Rau, M., Zhu, X. and Kalish, C. W. (eds), *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, available at <https://escholarship.org/uc/item/7g14d2f6

Kvaale, E. P., Haslam, N. and Gottdiener, W. H. [2013]: 'The "Side Effects" of Medicalization: A Meta-Analytic Review of How Biogenetic Explanations Affect Stigma', *Clinical Psychology Review*, **33**, pp. 782–794.

Kvaale, E. P., Gottdiener, W. H. and Haslam, N. [2013]: 'Biogenetic Explanations and Stigma: A Meta-Analytic Review of Associations Among Laypeople', *Social Science & Medicine*, **96**, pp. 95–103.

Kukla, Q. R. and Williams, R. M. [2024]: 'Introduction to the Kennedy Institute of Ethics Special Issue, "Situating Neurodiversity and Madness" ', *Kennedy Institute of Ethics Journal*, **34**, pp. ix–xv.

Lam, D. C. and 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**, pp. 405–411.

Lebowitz, M. S., Ahn, W. K. and Nolen-Hoeksema, S. [2013]: 'Fixable or Fate? Perceptions of the Biology of Depression', *Journal of Consulting and Clinical Psychology*, **81**, pp. 518–527.

Lebowitz, M. S. and Ahn, W. K. [2014]: 'Effects of Biological Explanations for Mental Disorders on Clinicians' Empathy', *Proceedings of the National Academy of Sciences*, **111**, pp. 17786–17790.

Lebowitz, M. S. and Appelbaum, P. S. [2017]: 'Beneficial and Detrimental Effects of Genetic Explanations for Addiction', *International Journal of Social Psychiatry*, **63**, pp. 717–723.

Lebowitz, M. S. and Ahn, W. K. [2018]: 'Blue Genes? Understanding and Mitigating Negative Consequences of Personalized Information About Genetic Risk for Depression', *Journal of Genetic Counseling*, **27**, pp. 204–216.

Lebowitz, M. S. and Appelbaum, P. S. [2019]: 'Biomedical Explanations of Psychopathology and Their Implications for Attitudes and Beliefs About Mental Disorders', *Annual Review of Clinical Psychology*, **15**, pp. 555–577.

Lebowitz, M. S., Dolev-Amit, T. and Zilcha-Mano, S. [2021]: 'Relationships of Biomedical Beliefs About Depression to Treatment-Related Expectancies in a Treatment-Seeking Sample', *Psychotherapy*, **58**, pp. 366–374.

Longino, H. [2013]: *Studying Human Behavior: How Scientists Investigate Aggression and Sexuality*, University of Chicago Press.

Loughman, A. and Haslam, N. [2018]: 'Neuroscientific Explanations and the Stigma of Mental Disorder: A Meta-Analytic Study', *Cognitive Research: Principles and Implications*, **3**.

Magliano, L., Ruggiero, G., Read, J., Mancuso, A., Schiavone, A. and Sepe, A. [2020]: 'The Views of Non-Psychiatric Medical Specialists About People With Schizophrenia and Depression', *Community Mental Health Journal*, **56**, pp. 1077–1084.

Marsh, J. K. and Romano, A. L. [2016]: 'Lay Judgments of Mental Health Treatment Options: The Mind Versus Body Problem', *MDM Policy & Practice*, **1**.

Miresco, M. J. and Kirmayer, L. J. [2006]: 'The Persistence of Mind–Brain Dualism in Psychiatric Reasoning About Clinical Scenarios', *American Journal of Psychiatry*, **163**, pp. 913–918.

* Miklowitz, D. J., Efthimiou, O., Furukawa, T. A., Scott, J., McLaren, R., Geddes, J. R. and Cipriani, A. [2021]: 'Adjunctive Psychotherapy for Bipolar Disorder: A Systematic Review and Component Network Meta-Analysis', *JAMA Psychiatry*, **78**, pp. 141–150.

Nettle, D., Frankenhuis, W. E. and Panchanathan, K. [2023]: 'Biology, Society, or Choice: How Do Non-Experts Interpret Explanations of Behaviour?', *Open Mind*, **7**, pp. 625–651.

Peters, D., Menendez, D. and Rosengren, K. [2020]: 'Reframing Mental Illness: The Role of Essentialism on Perceived Treatment Efficacy and Stigmatization', *Memory & Cognition*, **48**, pp. 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**, pp. 307–322.

Phillips, J., Morris, A. and Cushman, F. [2019]: 'How We Know What Not to Think', *Trends in Cognitive Sciences*, **23**, pp. 1026–1040.

Pilgrim, D. [2002]: 'The Biopsychosocial Model in Anglo-American Psychiatry: Past, Present and Future?', *Journal of Mental Health*, **11**, pp. 585–594.

Potochnik, A. [2017]: *Idealization and the Aims of Science*, University of Chicago Press.

Proctor, C. C. T. [2008]: *Clinicians' and Laypeople's Beliefs About the Causal Basis and Treatment of Mental Disorders*, Yale University Press.

Rashed, M. A. [2019]: 'In Defense of Madness: The Problem of Disability', *The Journal of Medicine and Philosophy*, **44**, pp. 150–174.

Ross, L. N. [2023]: 'Explanation in Contexts of Causal Complexity: Lessons From Psychiatric Genetics', in Bausman, W. C., Baxter, J. K. and Lean, O. M. (eds), *From Biological Practice to Scientific Metaphysics*, University of Minnesota Press, pp. 209-235.

Schroder, H. S., Devendorf, A. and Zikmund-Fisher, B. J. [2023]: 'Framing Depression as a Functional Signal, Not a Disease: Rationale and Initial Randomized Controlled Trial', *Social Science & Medicine*, **328**.

Schroder, H. S., Duda, J. M., Christensen, K., Beard, C. and Björgvinsson, T. [2020]: 'Stressors and Chemical Imbalances: Beliefs About the Causes of Depression in an Acute Psychiatric Treatment Sample', *Journal of Affective Disorders*, **276**, pp. 537–545.

Statham, G. [2020]: 'Normative Commitments, Causal Structure, and Policy Disagreement', *Synthese*, **197**, pp. 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**, pp. 814–820.

Szasz, T. S. [1960]: 'The Myth of Mental Illness', *American Psychologist*, **15**, pp. 113–118.

Van Fraassen, B. [1980]: *The Scientific Image*, Oxford University Press.

Weine, E. R. and Kim, N. S. [2019]: 'Systematic Distortions in Clinicians' Memories for Client Cases: Increasing Causal Coherence', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **45**, pp. 196–212.

Woodward, J. [2003]: *Making Things Happen: A Theory of Causal Explanation*, Oxford University Press.

——. [2021]: *Causation With a Human Face: Normative Theory and Descriptive Psychology*, Oxford University Press.

Zimmerman, H., Riordan, B. C., Winter, T., Bartonicek, A. and 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**, pp. 16–22.