The Charity Trap

Trust, Attribution, Emotional Dysregulation, and the Hidden Costs of Explaining Other People's Behavior

Flyxion
Independent Researcher

October 2026

Abstract

Ordinary social life depends on a habit so familiar that it rarely receives a name: when another person does something surprising, we supply a reason for it. The habit is indispensable, because a world in which every unexplained act counted as evidence of hostility or incompetence would be unlivable. It also has a cost that grows when a relationship becomes unreliable. This essay argues that the same interpretive generosity that makes cooperation possible can prevent an observer from updating their expectations, because the observer may keep generating the doubt that licenses the next act of forbearance. The argument is developed in four steps. First, the psychology of self-explanation and attribution is reviewed, including the finding that a widely taught actor–observer asymmetry is much weaker than its reputation. Second, a small formal apparatus is introduced, with five propositions and two corollaries, each proved, that separates explanatory possibility from justification, locates the silent threshold at which accepting commitments stops minimizing expected cost relative to verification, shows how an unconstrained supply of excuses removes the evidential content of any observation, shows that decisions based on conduct do not depend on the hidden cause, and shows how precedent-setting accommodation can exhaust a finite capacity in bounded expected time. Third, the framework is applied to relationships with people who have borderline personality disorder or bipolar disorder, which are treated as distinct conditions with distinct time structures, distinct treatments, and, for BPD in the one long cohort reviewed, high rates of symptomatic remission, and which are not treated as synonyms for manipulation or danger. Fourth, a practice of bounded charity is proposed in which compassion for a person's experience is kept separate from acceptance of every consequence of their conduct. The mathematical results are conditional statements about models and are not empirical findings. Clinical statements rest on sources checked against their bibliographic records and abstracts, with two checked in full, and their evidential strength is stated.

1 Introduction: The Reasonable Explanation

A man is told by a friend of twenty years that a scheduling change has been taken as a betrayal. A woman receives a string of messages at three in the morning that accuses her of things she did not do, followed by an apology so complete that the accusations seem to belong to

someone else. A manager agrees a plan on Monday, finds on Thursday that the plan has been disowned, and is told that nothing was ever agreed. A regulator receives a fourth explanation from the same operator for the same recurring failure, and finds it as reasonable as the first three. In each case the person on the receiving end performs a small piece of mental work so quickly that it passes for perception. They conclude that there must be a reason. Something happened that they do not know about. The friend has been under strain, the sender has a history that makes certain words land hard, the colleague was overloaded. The explanation is not invented from nothing, and it is frequently true. That is exactly why it is dangerous when it is not.

This essay concerns the habit of supplying reasons for other people's conduct, and the point at which that habit stops serving the person who has it. It begins from an observation that is easy to state and surprisingly hard to hold in mind: understanding why a person behaved as they did is a different act from establishing what others are obliged to tolerate. The first is an exercise in explanation. The second is a judgment about conduct and its consequences. A great deal of difficulty in close relationships arises because the first slides into the second without anyone noticing the join. The slide is not a moral failing. It is the ordinary operation of a very useful social faculty, applied in a setting where its usual guarantees have lapsed.

The topic has an unavoidable clinical shadow, and the essay deals with it directly because it is where the stakes are most visible. Relationships with people who have borderline personality disorder (BPD) or bipolar disorder can present circumstances in which ordinary expectations of proportion, continuity, and reciprocity stop working, in which the observer comes to monitor their own speech, and in which the explanations on offer are all individually plausible, and in such circumstances the mechanism described here may be relevant. How often it operates has not been measured. There is a considerable literature on the burden carried by relatives of people with these conditions (Bailey and Grenyer, 2013), which documents costs and not the inferential procedure formalized here, and a small body of evidence on programs designed to help them (Hoffman et al., 2005; Miklowitz et al., 2003). There is also a literature on the stigma that attaches to these diagnoses, which holds, among other things, that people with personality disorders are often perceived as purposefully misbehaving and that clinician stigma is a particular problem for BPD (Sheehan et al., 2016). Any honest treatment has to hold both facts at once. The relationships are hard, and the people in them are neither villains nor children.

A commitment of method follows. The essay does not claim that the framework applies only to psychiatric illness, and it does not claim that any diagnosis implies any pattern of behavior. The failure it describes is a failure of inference on the part of the observer, and it occurs in workplaces, families, institutions, and friendships with no diagnosis anywhere in sight. Psychiatric conditions enter as important and carefully differentiated applications, because they are cases in which behavior can be surprising, sincere, mistaken, and costly all at once, and in which the observer often has a strong and admirable motive to keep giving the benefit of the doubt. The essay does not diagnose anyone, does not teach readers to diagnose others, and treats the armchair attribution of a psychiatric label to a person in one's life as a hazard of its own, discussed in Section 7.

A word on register. The subject invites two bad tones, the clinical and the vindictive, and the essay tries to avoid both. A certain amount of dry humor about the absurdities of ordinary

social expectation is permitted, because the phenomenon is partly absurd: most people carry around a quietly heroic belief that everyone they meet is secretly reasonable, and they keep it up well past the point at which the evidence has stopped cooperating. The mathematics is permitted for a different reason. The central claims are claims about what follows from what, and such claims are best stated where they can be checked. The formalism is not decoration. Its main job is to state one point, concerning explanations chosen after the evidence arrives, precisely enough that it can be seen to be a fact about a procedure and not a matter of temperament, which prose alone does not manage. Each proposition is accompanied by its proof, its assumptions, and a plain statement of what it does not show. Where the essay relies on empirical findings it says how strong they are, and Table 3 in the final sections collects those judgments.

2 Two Conditions, Not One

It is common in everyday talk to treat borderline personality disorder and bipolar disorder as neighboring entries on a list of conditions that make people unpredictable. The habit is understandable and mistaken, and the mistake matters for the argument, because the two conditions give the observer very different information about what to expect.

Borderline personality disorder is defined in the current diagnostic manual as a pervasive pattern of instability in interpersonal relationships, self-image, and affect, together with marked impulsivity, beginning by early adulthood and present across contexts (American Psychiatric Association, 2022). The diagnosis is made when at least five of nine listed criteria are met, which means that two people with the same diagnosis may share only a minority of their features, and heterogeneity is the rule. A prominent account holds that the emotional difficulty in BPD can be analyzed into four components: heightened emotional sensitivity, heightened and labile negative affect, a deficit of adaptive regulation strategies, and a surplus of maladaptive ones (Carpenter and Trull, 2013). That analysis is itself an application of the biosocial model associated with dialectical behavior therapy, and the review that organizes it concludes that research needs to specify which component is under study. The full text adds that the evidence for some components is mixed: average levels of negative affect in BPD are not always much higher than in comparison groups, and its instability over time, including its capacity to intensify rapidly and without much warning, is described as the more characteristic feature. The authors also state that there is not yet enough empirical evidence to detail how the four components interact. Related work on rejection sensitivity, defined as anxious expectation of, ready perception of, and overreaction to cues of rejection, found that patients with BPD scored highest on a rejection sensitivity measure and on borderline-specific cognitions when compared with several other clinical groups and with healthy controls, including patients with social anxiety disorder (Staebler et al., 2011). The full text adds qualifications. The BPD groups were small (20 outpatients and 26 inpatients, the inpatients all women in the first weeks of a dialectical behavior therapy program), the measures were self-report questionnaires, the design was cross-sectional, and the authors describe the results as preliminary and in need of replication in larger samples. These limits restrict what can be said about any individual, but the finding supports a cautious statement that interpersonal threat cues carry unusual weight for many people with the diagnosis. The practical implication, which the later sections

develop, is that a response which appears out of proportion to its apparent trigger may be proportionate to the trigger as the person perceives it.

Bipolar disorder is a different kind of object. The disorders are chronic and recurrent, affect more than one percent of the world population, are among the most heritable psychiatric conditions, and are marked by episodes of elevated mood (mania or hypomania) and of depression (Vieta et al., 2018). The diagnostic manual defines the elevated episodes by duration, with a hypomanic episode lasting at least four consecutive days and a manic episode at least one week unless hospitalization intervenes (American Psychiatric Association, 2022). The temporal structure is the important point for the observer. A person with bipolar disorder who is between episodes may be entirely reliable, and may deeply regret what was said or done during an episode. That is possible and not guaranteed, since the disorders can also bring cognitive and functional impairment (Vieta et al., 2018). Diagnosis can be hard, the same review notes, and patients and their families do not always understand the significance of their symptoms, especially hypomanic and manic ones (Vieta et al., 2018). Treatment typically combines mood-stabilizing or antipsychotic medication with psychosocial strategies.

The two conditions co-occur, and a literature review found that their relationship is neither consistent nor specific, that there are overlaps but important differences in phenomenology and medication response, that it is unusual for BPD to evolve into bipolar disorder, and that existing data do not support the claim that the two lie on a single spectrum (Paris et al., 2007). For the present purposes the review licenses a modest conclusion: the conditions should be handled as separate hypotheses about a person, each with its own implications for the timing and predictability of difficult conduct. To a first approximation, bipolar episodes organize difficulty in time, so that conduct during an episode may differ sharply from conduct before and after it. In BPD, instability in relationships is itself a defining feature in the diagnostic description (American Psychiatric Association, 2022), and the rejection-sensitivity findings above suggest that interpersonal cues are especially potent triggers for many patients, so that difficulty may recur across many ordinary situations. These are generalizations about populations and not predictions about individuals, and the next sections show how little of the practical argument depends on them.

Two further facts need to be stated early, because without them the discussion tilts toward fatalism. The first concerns outcomes in BPD. In a sixteen-year prospective study of 290 inpatients with the diagnosis, sustained symptomatic remission was reached at rates between 78 and 99 percent, depending on how long remission was required to last, whereas sustained recovery, which includes good social and vocational functioning, was reached by 40 to 60 percent. Recurrence and loss of recovery were more common than in comparison patients with other personality disorders (Zanarini et al., 2012). The authors' conclusion is that remission is far more common than recovery and that both are harder to attain and keep with BPD than with other personality disorders. The picture is therefore neither hopeless nor trivial. Two cautions apply. The distinction between remission and recovery should stay in view, since relief of symptoms does not guarantee good social and vocational functioning. And the cohort consisted of patients identified at an inpatient admission, so its findings cannot be assumed to hold for every clinical population. The second fact concerns treatment. A recent practice guideline of the American Psychiatric Association recommends a structured, evidence-supported psychotherapy that targets the core features of BPD and suggests that any psychotropic medication be used alongside psychotherapy, with a defined target and an end

point (American Psychiatric Association, 2024a). The guideline's chair, announcing it, said that no therapy emerged as a gold standard and that no evidence was found for the effectiveness of any pharmacotherapy in treating the core symptoms of the disorder (American Psychiatric Association, 2024b). A Cochrane review of 75 randomized trials with 4,507 participants, mostly of dialectical behavior therapy and mentalization-based treatment, found moderate-quality evidence that psychotherapy reduces borderline symptom severity compared with treatment as usual (standardized mean difference -0.52), low-quality evidence for reductions in self-harm and suicide-related outcomes, and no clear difference among therapy types (Storebø et al., 2020). Individual trials point the same way. Dialectical behavior therapy outperformed community treatment by experts on suicide attempts and hospitalization in a two-year trial of 101 women (Linehan et al., 2006), and mentalization-based treatment outperformed structured clinical management in a trial of 134 patients, although both groups improved (Bateman and Fonagy, 2009). For bipolar disorder, family-focused therapy combined with medication produced fewer relapses (35 percent against 54 percent) and better medication adherence than brief crisis management in a randomized trial of 101 patients (Miklowitz et al., 2003).

The stigma literature closes the section. Public knowledge of personality disorders is described as low, people with these diagnoses may be perceived as purposefully misbehaving rather than as ill, and health provider stigma appears especially strong for BPD (Sheehan et al., 2016). This matters for the argument in two directions. It means that a framework which tells observers to stop explaining away conduct risks being heard as permission to withdraw sympathy, and the essay must be read against that misreading. It also means that the people who live closest to these conditions often arrive at their conclusions without help from any clinician who understands the family's position, which is part of why the burden reviewed by Bailey and Grenyer (2013) is so heavy. In a systematic review of six studies covering 465 carers, five of them concerning BPD, carers showed elevated objective and subjective burden, grief, and mental health problems, with burden scores about half a standard deviation above those of carers of inpatients with other serious mental illnesses.

3 Why Ordinary Minds Explain Other People Generously

The human habit of supplying reasons for behavior begins at home. A classic review of the evidence argued that people have little direct access to the causes of their own behavior and often report on them by applying plausible theories, not by introspecting (Nisbett and Wilson, 1977). A more recent line of experiments produced a vivid illustration. In studies of choice blindness, participants chose which of two faces they found more attractive, and the experimenters covertly swapped the outcome so that participants were handed the face they had not chosen. Many failed to notice the swap, and many then offered reasons for the choice they had not in fact made (Johansson et al., 2005). The finding is not that people are liars. It is that the machinery of explanation runs smoothly even when its input is wrong, and that its output is experienced as a report and not as a construction.

If the machinery generates fluent reasons for one's own acts, it can generate reasons for another person's acts with very little adjustment, and a good deal of social psychology is about the form this takes. The correspondence bias, in the classic review, is the tendency to infer a person's enduring dispositions from behavior that can be entirely explained by the situation

in which it occurs (Gilbert and Malone, 1995), a tendency first demonstrated in studies in which observers inferred a writer's attitude from an essay even when told that the position had been assigned (Jones and Harris, 1967), and later dubbed the fundamental attribution error (Ross, 1977). The standard textbook companion is the actor–observer asymmetry, the claim that people explain their own behavior by circumstances and other people's by traits. Here the record needs correcting, since the asymmetry is much weaker than its reputation. A meta-analysis of 173 published studies found average effect sizes between $d = -0.016$ and $d = 0.095$, an asymmetry that held only under particular conditions (including when actor and observer were intimates, and for negative events, with a reverse pattern for positive ones), and an overall conclusion that, across valence, no actor–observer asymmetry exists (Malle, 2006). Cross-cultural work adds that the preference for dispositional explanations varies with culture while some patterns appear universal (Choi et al., 1999). The foundations on which popular accounts of attribution rest are therefore softer than they seem.

That has a consequence for the argument, and it is where the essay departs from the textbook attribution story. If the asymmetry is weak, the essay cannot rest on the claim that people systematically blame others while excusing themselves. The phenomenon of interest points in the opposite direction in one kind of case. It concerns the people we care about, whose surprising behavior we explain by hidden circumstances precisely because we have a model of them as reasonable. Research on close relationships has described this process directly. People in satisfying relationships have been shown to reinterpret a partner's faults in ways that preserve their image of the partner, seeing virtues in faults (Murray and Holmes, 1993), and to idealize partners in ways that predict relationship satisfaction (Murray et al., 1996). The relevant claim is not that such positive construction is foolish. In the research it is associated with satisfaction. The claim is that the same process, running in a relationship whose reality has changed, may keep the model of the partner from being updated.

Philosophy supplies the name for the underlying principle. In the tradition of radical interpretation, understanding another speaker requires assuming that they are largely rational and largely right, because without that background there is no way to attribute any particular belief at all (Davidson, 1973). The principle of charity is therefore not a courtesy. It is a constraint on interpretation, and it is the reason that a reading of a surprising remark in which the speaker means something sensible is almost always the first to occur. Charity in this sense governs what a remark means. It does not decide what an act obliges its audience to absorb. The step from the first to the second is the step this essay is about, and the history of psychology and philosophy both indicate why it feels so natural: the faculty that interprets and the faculty that excuses use largely the same inputs.

A last piece of background concerns the opposite pressure. Human beings are not credulous by default, and the literature on epistemic vigilance argues that communication would be unstable if listeners had no mechanisms for evaluating the reliability of sources and the plausibility of what they are told (Sperber et al., 2010). Vigilance is costly, though, and it is deployed selectively. Close relationships are precisely the setting in which the cost of appearing to doubt is highest, because to question a loved person's account is itself an act with consequences. The result is that the setting in which vigilance would be most informative is the setting in which it is most expensive to exercise, and the next section makes that cost explicit.

4 The Charitable Inference Problem

The argument can now be stated more exactly. An observer O interacts with another person P over periods $t = 1, 2, \ldots$. At each period P performs an act a_t drawn from a set $\mathcal{A}$ of possible acts. The act is influenced by a circumstance c_t drawn from a set $\mathcal{C}$, which may include P's stress, a misunderstanding, information O does not have, the state of an illness, or an ordinary reason of any kind. The observer does not see c_t. The observer sees the act, and has a history H_t of the relationship. When the observer asks why the act occurred, the rational answer is the posterior over circumstances,

$$\mathbb{P}(c \mid a_t, H_t) \;=\; \frac{\mathbb{P}(a_t \mid c, H_t)\,\mathbb{P}(c \mid H_t)}{\sum_{c' \in \mathcal{C}} \mathbb{P}(a_t \mid c', H_t)\,\mathbb{P}(c' \mid H_t)}. \tag{1}$$

Let $\mathcal{C}(a) = \{c \in \mathcal{C} : \mathbb{P}(a \mid c, H) > 0\}$ be the set of circumstances compatible with act a. Let $J(a, c) \in \{0, 1\}$ record whether, under circumstance c, the act is one the observer is obliged to absorb, so that $J(a, c) = 1$ means that the act is justified in that circumstance. The function J is a normative judgment supplied by the observer's values and obligations, and mathematics does not provide it. Equation (1) is the part that mathematics governs.

The central error occurs when plausibility is treated as justification. It can be stated as a proposition.

Proposition 1 (Possibility is not justification)**.** *The existence of a justifying circumstance compatible with an act does not imply that the act is justified. Formally, $\exists\, c \in \mathcal{C}(a) : J(a, c) = 1$ does not imply $\forall\, c \in \mathcal{C}(a) : J(a, c) = 1$, and the posterior probability that the act is justified,*

$$\mathbb{P}(J = 1 \mid a, H) \;=\; \sum_{c \in \mathcal{C}(a)} J(a, c)\,\mathbb{P}(c \mid a, H),$$

can be made arbitrarily small even though a justifying circumstance with positive prior probability exists.

Proof. Suppose $c_1, c_2 \in \mathcal{C}(a)$ with $J(a, c_1) = 1$ and $J(a, c_2) = 0$. The existential statement holds because of c_1 and the universal statement fails because of c_2, so the implication is false. For the quantitative claim, let $\mathbb{P}(c_1 \mid H) = \varepsilon > 0$ and $\mathbb{P}(c_2 \mid H) = 1 - \varepsilon$, with $\mathbb{P}(a \mid c_1, H) = \mathbb{P}(a \mid c_2, H)$. Then equation (1) gives $\mathbb{P}(c_1 \mid a, H) = \varepsilon$, so $\mathbb{P}(J = 1 \mid a, H) = \varepsilon$, which can be taken as small as desired while $c_1 \in \mathcal{C}(a)$ remains true. □

The proposition is modest. Its value is pedagogical, because it isolates the exact place at which a familiar interpersonal inference becomes invalid. The first three steps of the ordinary line of thought are reasonable: I do not understand why they acted this way, there may be causes I do not know, and those causes may make the behavior understandable. The fourth step, that the behavior is therefore justified, does not follow. What licenses it, when it is licensed, is evidence about which circumstance actually obtained, and a possible circumstance is not evidence of an actual one.

Four properties of an act under a circumstance are easily confused, and the confusion is the engine of the trap. Table 1 sets them out.

Property	Question it answers	What can be true while it is false
Intelligibility	Can the act be explained by some circumstance?	The act is explained, yet unjustified.
Sincerity	Does the stated feeling match the person's actual state?	The feeling is real, yet the belief behind it is mistaken.
Accuracy	Does the person's belief match the situation?	The belief is accurate, yet the response is disproportionate.
Justification and cost	Is the act one that others are obliged to absorb, and what does it cost them?	The cost is real whether or not anyone intended it.

Table 1: Four properties of an act that are often collapsed. They are conceptually distinguishable and do not in general entail one another, so any of them can hold while the others fail.

Consider a concrete case that does not depend on any diagnosis. A person sincerely believes they have been abandoned because a reply took a day to arrive. The feeling is sincere. The belief that they were abandoned is mistaken. The reaction, which consists of dozens of messages and an ultimatum, imposes a cost on the recipient. All three statements can be true together, and an observer who grasps this can respond to the feeling ("that sounds frightening") without endorsing the belief ("I was not abandoning you") and without accepting the ultimatum. The common failure runs the other way. Because the feeling is sincere, the observer treats the belief as accurate, and because the belief is understandable, the observer treats the conduct as warranted.

This is the point at which the distinction between intentional exploitation and destabilization regardless of intent does its work. A person may sincerely experience betrayal where an observer sees a changed schedule. If the resulting accusations, demands, or reversals recur, the observer will find that ordinary expectations of reciprocity and continuity no longer work, whatever anyone intended. Nothing in the analysis depends on settling intent, which is fortunate, because intent is the one thing the observer cannot observe, and attempts to infer it tend to produce either sentimentality or paranoia.

5 The Economy of Default Trust

Nobody can run a life on verification. A person who asked for proof of every promise, checked every account of where a friend had been, and prepared a defensive reply for every possible reversal would have no attention left for anything else, and would probably have no friends. Default trust is an adaptive economy of attention and not a lapse in intelligence. The question for the observer is not whether to trust, but when the economy of trust stops paying.

The simplest model that captures this treats each interaction in which P makes a claim or commitment as a decision. The observer may verify, at a cost $v > 0$ that includes the relational cost of appearing to doubt, in which case any misrepresentation is avoided. Alternatively the observer may accept the commitment at face value. If the commitment is honored with probability p, and a broken commitment costs the observer a loss L, accepting has expected loss $(1 - p)L$.

Proposition 2 (The silent threshold)**.** *Suppose $0 < v < L$. Accepting commitments without verification is weakly cost-minimizing if and only if $p \geq p^*$, where $p^* = 1 - v/L$, and the two options tie at $p = p^*$. If $v \geq L$, accepting is optimal for every $p \in [0, 1]$.*

Proof. The expected cost of accepting is $(1 - p)L$ and the cost of verifying is v. Accepting is weakly better when $(1 - p)L \leq v$, that is, when $p \geq 1 - v/L$. If $v \geq L$ then $(1 - p)L \leq L \leq v$ for every p, so accepting is weakly better throughout. □

The comparison is between accepting and perfect verification only, and says nothing about other policies such as partial checking or conditional commitments.

Two features of the result deserve comment. The first is that the threshold falls as verification becomes more costly relative to the loss, which is why close relationships tolerate so much. The relational cost of checking up on a partner or parent or old friend is high, so v is large, so p^* is low, and a relationship can survive a large number of broken commitments before the economy of trust turns against it. The second is that adult stakes raise L. A broken commitment about a dinner reservation is cheap. A broken commitment about rent, a loan, the care of children, or the terms of a shared business is not, and as L grows with the stakes, p^* rises toward one. A relationship in which the commitments carry high stakes therefore needs a higher reliability than one in which they do not, and the observer who applies the thresholds appropriate to a low-stakes friendship to a high-stakes arrangement is making an error that no amount of kindness corrects.

The threshold is silent because nothing in experience announces that p has crossed it. Each broken commitment is individually small and individually explicable, and the sum is invisible unless someone is keeping count. Suppose the observer holds a Beta(α, β) prior over p and has seen m honored and n broken commitments, so that the posterior mean is $\hat{p} = (\alpha + m)/(\alpha + \beta + m + n)$.

Corollary 1 (The exchange rate of goodwill)**.** *Under the stationarity assumption of the Beta model, $\hat{p} < p^*$ if and only if*

$$n \;>\; n^*(m) \;=\; \frac{v(\alpha + m) - (L - v)\beta}{L - v},$$

and $n^*(m)$ is increasing in m with slope $v/(L - v)$.

Proof. Write $p^* = (L - v)/L$ and $1 - p^* = v/L$. The condition $\hat{p} < p^*$ is $\alpha + m < p^*(\alpha + \beta + m + n)$, equivalent to $p^* n > (1 - p^*)(\alpha + m) - p^* \beta$. Dividing by p^* and substituting gives $n > \big[(v/L)(\alpha + m) - ((L - v)/L)\beta\big] \big/ \big((L - v)/L\big) = \big[v(\alpha + m) - (L - v)\beta\big]/(L - v)$. The slope in m is $v/(L - v)$. □

The corollary gives an informal bookkeeping principle, a model-implied exchange rate and not an empirical unit of goodwill. Each honored commitment buys $v/(L - v)$ broken ones before the economy of trust turns negative, so goodwill is a bank account with an exchange rate set by the cost of checking and the size of the loss. A long history of reliability is a large balance, and it is rational for a long history to delay the conclusion that something has changed. The assumption of stationarity is the weak point. If the process generating reliability has itself changed, as when an illness enters a new phase or a life circumstance shifts, the Beta model

keeps counting the old balance as current, and an observer who discounts old evidence at some rate would cross the threshold sooner. The mathematics does not tell the observer which model is right. It shows that a long, good history and a recent run of failures are both facts, and that the weight given to each is a modeling choice that deserves to be made deliberately.

The corollary also marks a degenerate regime. As v approaches L the exchange rate $v/(L-v)$ grows without bound, and when $v \geq L$ the threshold disappears altogether (Proposition 2). This is the situation in which checking is nearly as costly as being wrong, so that no accumulation of evidence can make verification the cheaper policy and, under this cost structure, nothing can change the observer's decision, and it is arguably the case of greatest practical concern: a relationship in which questioning is so costly that the observer can never afford to look.

There is also a connection to the study of reciprocity. In the repeated games that Axelrod and Hamilton analyzed, a strategy of conditional cooperation, which cooperates at first and thereafter does what the partner did last time, did well against a wide range of alternatives (Axelrod and Hamilton, 1981). The relevant feature here is that the strategy responds to the partner's conduct and not to the partner's motives. It never asks why. The simplicity is not a deficiency. It is what makes the strategy robust to exactly the explanatory uncertainty that this essay is about.

6 When Excuses Inflate

The economy of trust assumes that the observer counts evidence honestly. The next result shows how counting can fail without anyone intending it. Consider two hypotheses about P. Under M_R, P is reasonably reliable in the sense relevant to the relationship, and surprising acts have ordinary explanations. Under M_U, P's conduct is not reliably governed by considerations the relationship depends on. The observer has prior odds $\Omega_0 = \mathbb{P}(M_R)/\mathbb{P}(M_U)$ and watches a sequence of surprising acts $v_1, \ldots, v_n$, each of which violates the observer's expectations.

Under M_U the probability of the tth surprising act is $u_t = \mathbb{P}(v_t \mid M_U, H_t) > 0$. Under M_R the surprising act arises only through some circumstance. Let $\ell(v \mid c) \in [0,1]$ be the probability of act v under circumstance c given that P is reasonably reliable, and let π_t be the distribution over circumstances that the observer entertains at time t. Then

$$q_t \;=\; \mathbb{P}(v_t \mid M_R, H_t) \;=\; \sum_{c} \pi_t(c)\, \ell(v_t \mid c).$$

By sequential application of Bayes' theorem, the posterior odds after n surprising acts are

$$\Omega_n \;=\; \Omega_0 \prod_{t=1}^{n} \frac{q_t}{u_t}. \tag{2}$$

Proposition 3 (Excuse inflation)**.** *Let $\ell^*(v) = \max_c \ell(v \mid c)$, assumed attained.*

(a) Disciplined charity updates. *If the excuse distribution π_t is fixed independently of v_t and $q_t \leq \rho\, u_t$ for all t with $\rho < 1$, then $\Omega_n \leq \rho^n \Omega_0$. In particular $\Omega_n \leq \omega$ once $n \geq \max\{0, \lceil \ln(\Omega_0/\omega)/\ln(1/\rho) \rceil\}$.*

(b) Undisciplined charity does not update. *Suppose the observer chooses π_t sequentially, each after observing v_t, and that $\ell^*(v_t) \geq u_t$ for every t. Write $\tilde{q}_t = \sum_c \pi_t(c)\,\ell(v_t \mid c)$ for the adaptively fitted score and $\tilde{\Omega}_n = \Omega_0 \prod_{t=1}^{n} \tilde{q}_t / u_t$ for the resulting apparent odds. Then there is a sequence of choices for which $\tilde{\Omega}_n \geq \Omega_0$ for every n. If moreover $\min_c \ell(v_t \mid c) \leq u_t \leq \ell^*(v_t)$ for every t, the observer can choose π_t so that $\tilde{\Omega}_n = \Omega_0$ exactly.*

Proof. (a) By hypothesis each factor $q_t/u_t \leq \rho$, so equation (2) gives $\Omega_n \leq \rho^n \Omega_0$. The bound $\rho^n \Omega_0 \leq \omega$ holds exactly when $n \ln(1/\rho) \geq \ln(\Omega_0/\omega)$. Since n is an integer, the smallest such n is the ceiling of the right-hand quotient, and $n = 0$ suffices if $\Omega_0 \leq \omega$.

(b) Choose π_t to place all its mass on a circumstance c_t^* attaining $\ell^*(v_t)$. Then $\tilde{q}_t = \ell^*(v_t) \geq u_t$, each factor $\tilde{q}_t/u_t$ is at least one, and $\tilde{\Omega}_n \geq \Omega_0$. For the second claim, the map $\pi \mapsto \sum_c \pi(c)\ell(v_t \mid c)$ is continuous on the simplex of distributions, which is connected, so its image is an interval containing both its minimum $\min_c \ell(v_t \mid c)$ and its maximum $\ell^*(v_t)$. Any value u_t in that interval is attained by some π_t, giving $\tilde{q}_t = u_t$ and a unit factor at every step. □

The proposition formalizes a commonplace worry. Part (a) says that if the observer decides in advance, or on independent evidence, how much weight excuses deserve, then repeated surprising acts drive the odds on reliability down at a geometric rate, and a finite number of events changes the observer's mind. Part (b) says that if the observer is free to tailor the excuse to the act, then no finite or infinite sequence of surprising acts need move the odds at all. The condition $\ell^*(v_t) \geq u_t$ is exactly the thought, *there may be causes I do not know*, since it says that some imaginable circumstance makes the act at least as likely under M_R as under M_U. That thought is almost always true. A sufficiently inventive observer can find a circumstance for any act whatever. What is false is the unstated premise that the circumstance was actually present with the weight the argument requires. This is the sense in which the observer can be said to manufacture the doubt that licenses continued forbearance: the evidence is never allowed to count because the explanation is always chosen after it arrives.

Methodological status. Part (b) is not a theorem about Bayesian inference. Ordinary Bayesian updating fixes the likelihood model before the data arrive. Choosing π_t after seeing v_t and then inserting the resulting score $\tilde{q}_t$ in place of q_t in equation (2) is a form of post-selection, or adaptive model fitting, in which the data influence the model that is then used to assess them. The result shows what such a procedure does, which is why it models a human habit and not a defect in Bayes' rule. An observer whose predictive model is specified independently of the current observation avoids the particular post-selection defect described in part (b). Such an observer updates with the predictive likelihood obtained by integrating over circumstances before the event is observed. Part (a) applies only when those predictive likelihood ratios additionally satisfy the uniform bound $q_t \leq \rho u_t$. A coherent procedure does not guarantee that every surprising observation counts against reliability, and a properly specified model can assign a surprising event a ratio above one. What it guarantees is that the evidential weight is computed under a model that accounts for its own uncertainty, and not under one selected solely to accommodate the observation. For the same reason the quantities that the post-selected procedure produces are not posterior odds under any single coherent joint model fixed before the data. They are best read as apparent odds, or pseudo-posterior odds, and calling them posterior odds would grant the procedure a legitimacy it lacks.

A worked comparison. Hold the structure fixed and compare two procedures symbolically.

Under disciplined charity, suppose each surprising act has $q_t/u_t = \rho < 1$. Under undisciplined charity, suppose the observer picks the best available excuse each time, so that $\tilde{q}_t = \ell^*(v_t)$ and $\tilde{q}_t/u_t = r \geq 1$. Assume the same initial odds Ω_0 for both procedures and constant ratios across events. The odds on reliability, apparent odds in the undisciplined case, then follow Table 2. The disciplined column falls geometrically however mild ρ is, and the undisciplined column never falls. The ratio of the undisciplined to the disciplined odds after n events is $(r/\rho)^n$, so even a small per-event difference compounds, which is why the divergence is invisible in any single case and decisive across many. The third column shows that perfect tailoring is not needed for failure to update: any tailoring that keeps $\tilde{q}_t/u_t \geq 1$ suffices, and the disciplined column is the only one that moves. No numerical values are assumed, because the point is structural.

Events n	Disciplined, $\rho < 1$	Undisciplined, $r \geq 1$	Matched excuse, $q_t = u_t$
0	Ω_0	Ω_0	Ω_0
1	$\Omega_0\rho$	$\Omega_0 r$	Ω_0
2	$\Omega_0\rho^2$	$\Omega_0 r^2$	Ω_0
n	$\Omega_0\rho^n$	$\Omega_0 r^n$	Ω_0

Table 2: Odds on reliability after n surprising acts under three procedures for weighing excuses, with equal initial odds and constant ratios. The second and third columns are apparent, not posterior, odds.

Two further remarks keep the result honest. First, the proposition is symmetric. An observer who selects sinister hypotheses after the fact, supplying a hidden motive for each innocent act, achieves the mirror-image result and is equally unable to learn. Suspicion can inflate in the same way charity can, and the cure is the same: the weight on explanations must be set by something other than the act that they are meant to explain. Second, the proposition describes a procedure and says nothing about any person. It does not claim that anyone generates surprising acts at any rate, and it does not claim that the people who care about a person with a psychiatric diagnosis are doing this. It shows what the procedure of tailoring excuses to acts would do, so that the observer can notice when the procedure is being run.

What does discipline look like in practice? Three practices follow from the mathematics and need no clinical knowledge. The observer can record surprising acts when they happen, so that the count in equation (2) is not reconstructed from a memory biased by the latest explanation. The observer can distinguish excuses supported by independent information, such as a death in the family the observer actually knows about, from excuses that are merely conceivable, and give only the first any weight beyond the base rate. And the observer can fix in advance what number or pattern of events would change their conduct, which is the only way to prevent the threshold from moving as the evidence arrives. Each is a version of part (a), the restriction of π_t to something not chosen after the act.

The phrase independent support needs an operational meaning. What counts is information the observer held before the act, such as a bereavement or illness the observer already knew about, a documented medical event, a change in circumstances confirmed by a third party, or a recurring pattern that predicts the act, such as a known anniversary or a known period of strain. What does not count is an explanation first offered at the moment of accounting,

a story that could accompany any act whatever, or the observer's own surmise. The test is whether the circumstance would have been believed before the act occurred. Explanations offered by the person afterwards are not worthless. They are testimony about the person's experience and may be weighed as such, but testimony that arrives only after the act and is tailored to it is the least independent kind of support, and it is the kind the proposition shows to be uninformative about reliability when it is the only kind available.

7 Conduct Over Cause

The framework has so far allowed the observer to be uncertain about circumstances. It remains to ask what the observer should do, and here the central practical result is that the decision can often be made without resolving the uncertainty at all.

Let the observer's decision be a policy d chosen from a set $\mathcal{D}$ (continue as before, change the terms of contact, withdraw, seek help). Let $X = (X_{t+1}, \dots, X_T)$ be the future conduct of P as it affects O, and let $u(d, X)$ be the observer's utility from policy d when conduct is X. Let hypotheses $M_1, \dots, M_k$ be competing accounts of the hidden cause. Each hypothesis, given history, induces a predictive law F^i for X.

Proposition 4 (Cause-equivalence of conduct-based decisions)**.** *If two causal hypotheses M_i and M_j induce the same predictive law for future conduct, $F^i = F^j$, then every expected-utility-maximizing policy under M_i is expected-utility-maximizing under M_j. More generally, if the observer's beliefs are a mixture $\sum_i w_i M_i$ then the optimal policy depends on the hypotheses only through the mixture predictive law $\bar{F} = \sum_i w_i F^i$.*

Proof. The expected utility of policy d under M_i is $\mathbb{E}_{F^i}[u(d, X)]$. If $F^i = F^j$ the two expectations are equal for every d, so their maximizers coincide. Under the mixture, expected utility is $\sum_i w_i \mathbb{E}_{F^i}[u(d, X)] = \mathbb{E}_{\bar{F}}[u(d, X)]$ by linearity of expectation, which depends on the hypotheses only through $\bar{F}$. □

The proposition has a consequence that is easy to state and easy to forget. The question the observer cannot answer, which is why the other person did what they did, matters for decisions only to the extent that different answers imply different predictions about what the person will do next. When the competing explanations are *predictively equivalent* for the horizon that matters, the decision does not need to wait for the answer, and the relationship need not stall on the debate. This is the formal content of the advice to attend to conduct and not to motives.

The result is conditional on the form of the utility function. It holds when $u(d, X)$ depends on conduct and the observer's own costs. Real decisions also turn on considerations that can depend on the cause directly: judgments of moral responsibility, obligations arising from the relationship, assessments of intention, and the intrinsic value placed on maintaining the relationship. If utility is allowed to depend on the circumstance as well, $u(d, X, c)$, the equal-forecast argument no longer goes through, because two hypotheses with identical forecasts can still differ in expected utility. The proposition is therefore a conditional separation result. It identifies when the causal question can be set aside, namely when the decision turns on conduct and cost, and it does not claim that motives are irrelevant to every decision.

The proposition does not say that causes never matter. Cause matters through prediction. A hypothesis that a surprising act belongs to a bipolar episode predicts that the conduct will change when the episode ends, while a hypothesis that it expresses a settled pattern predicts that it will not, and the optimal timing of major commitments differs. A hypothesis that an outburst arose from an interpersonal threat the person perceived predicts that conduct will vary with the presence of such cues, which is information the observer can use. The practical reading is that diagnosis, when it is known and shared by the person and their clinicians, is valuable as information about F, and that it is not a permission slip. A diagnosis informs forecasts. It does not change what a given pattern of conduct costs the people around it.

For the same reason the observer should not attempt to assign a diagnosis. The informal assignment of a psychiatric label to a person in one's life has three defects. It is unreliable, since diagnosis is hard even for clinicians, as the review of bipolar disorders emphasizes (Vieta et al., 2018), and since the conditions overlap and co-occur without a simple relationship (Paris et al., 2007). It is stigmatizing, in a literature that documents the burden of the label (Sheehan et al., 2016). And it is usually unnecessary, because Proposition 4 shows that decisions grounded in conduct can often proceed without it. A relative who says, "I do not know why this keeps happening, and the cost to me has reached a level I cannot sustain," has said everything the decision requires.

8 Applying the Framework to Familiar Situations

The preceding results can be exercised on situations that relatives, partners, friends, and colleagues describe repeatedly. In each case the procedure is the same: identify the four properties of Table 1 separately, ask what is known as against merely conceivable, and decide on conduct.

The sudden accusation. A remark that seemed innocuous is met with a severe accusation. The charitable hypotheses are numerous: the remark touched something painful, the person is under strain, context is missing. By Proposition 1 none of them settles whether the accusation was warranted. For some people with BPD, one of the explanations has independent support, since rejection sensitivity has been found to be unusually high in BPD samples (Staebler et al., 2011). That is useful to know, and it supports responding to the feeling with calm. It does not follow that the accusation was accurate, and an observer who says "you are right, I did that," to end the conflict creates a false record that will return at the next dispute. A response consistent with the framework acknowledges the feeling, states plainly what actually happened, and declines to negotiate the facts in the middle of the crisis.

The repeated request for reassurance. The observer is asked, again and again, to confirm that they are not leaving, not angry, not lying. Reassurance is kind and often works for a time. If reassurance is the only response available it can become a condition of the relationship, which is the dynamic modeled in Section 9. An alternative is to offer reassurance on a predictable schedule and in a stable form, and to decline to re-litigate it on demand, so that the content of the message does not depend on the intensity of the request.

The agreement that does not survive. Plans agreed on Monday are disowned on Thursday. The cleanest response is to record agreements in a form both people can see, such as a short message

that restates what was decided, not as a trap but as a shared memory. Part of the burden of such relationships, as carers report (Bailey and Grenyer, 2013), is that the observer must carry the memory alone. Writing it down distributes it.

The threat or statement about self-harm. This case differs from the rest and must be stated without ambiguity. A statement that a person may end their life, or a report of self-harm, is to be taken seriously every time, even when similar statements have been made before and even when the observer suspects it is intended to prevent a departure. The reasoning of this essay about conduct and consequence does not govern the immediate response, and nothing here should be read as advice to treat such statements as manipulation. The response should be proportionate to immediacy and severity: emergency services when danger is imminent, a crisis line (9-8-8 in Canada and the United States) for support and guidance, any existing safety plan, and the person's clinicians where they exist. A crisis line and an emergency response are different resources, and which is appropriate depends on the situation. Questions of long-term boundaries and planning remain legitimate, but they belong to a later calm period and to a plan made with professional support, after immediate safety has been addressed.

The temptation to become the therapist. Relatives who understand a good deal about a condition are tempted to apply it. The treatment evidence reviewed here concerns structured psychotherapies delivered by trained clinicians (American Psychiatric Association, 2024a; Storebø et al., 2020), and none of it concerns informal interpretation by relatives, so there is no basis for treating the latter as a substitute. Family members can help most by supporting access to treatment, by learning the skills that programs for relatives teach, and by retaining their own role as a relative.

9 The Ratchet of Accommodation

The sections so far concern what the observer believes. This section concerns what happens to the observer's capacity when accommodation is the standing response. Much of the lived difficulty described by relatives is not a single event but a drift: the demands do not feel large on any given day, and then a year later the arrangement of the household or the friendship has organized itself around preventing the next crisis. One person requires increasing reassurance, the other devotes increasing attention to anticipating reactions, and eventually the second person monitors their own speech, withholds ordinary disagreement, or concedes unreasonable demands to preserve immediate calm. The pattern can be self-reinforcing without either participant having designed it.

A small model shows how little is needed to produce the drift. Let $E_t \geq 0$ be the precedent level at time t, meaning the level of accommodation that has come to be treated as normal. At each period P makes a demand D_t, with $E_t \leq D_t \leq E_t + m$ for some bound $m > 0$, so that demands do not fall below precedent and do not jump by more than m. Suppose that conditional on the past, the demand exceeds precedent by at least $\delta > 0$ with probability at least $p > 0$. The observer chooses an accommodation A_t. Precedent is set by what has been granted, $E_{t+1} = \max(E_t, A_t)$. The observer has a finite capacity $C > E_0$, the largest sustained level of accommodation they can bear. Let $T = \min\{t : E_t \geq C\}$ be the time at which precedent reaches capacity.

Proposition 5 (Ratchet). *If the observer always yields, $A_t = D_t$, then $\mathbb{E}[T] \le (C + m - E_0)/(p\delta)$. In particular $T < \infty$ almost surely.*

Proof. With $A_t = D_t$ we have $E_{t+1} - E_t = D_t - E_t \in [0, m]$, and this increment is at least δ with probability at least p, so $\mathbb{E}[E_{t+1} - E_t \mid \mathcal{F}_t] \ge p\delta$. Define $Y_t = E_{t\wedge T} - p\delta\,(t \wedge T)$. On $\{T > t\}$ the conditional expected change of Y_t is nonnegative, and on $\{T \le t\}$ it is zero, so Y_t is a submartingale and $\mathbb{E}[Y_n] \ge Y_0 = E_0$. Hence $p\delta\,\mathbb{E}[T \wedge n] \le \mathbb{E}[E_{T\wedge n}] - E_0$. Before T we have $E_t < C$, and at T the increment is at most m, so $E_{T\wedge n} < C + m$. Therefore $\mathbb{E}[T \wedge n] \le (C + m - E_0)/(p\delta)$ for every n, and letting $n \to \infty$ by monotone convergence gives the claim. □

Corollary 2 (A cap holds). *If the observer instead follows the policy $A_t = \min(D_t, a^*)$ with $E_0 \le a^* < C$, then $E_t \le a^* < C$ for all t, so $T = \infty$ with certainty. The price is that every demand above a^* is partly refused.*

Proof. By induction, $E_0 \le a^*$, and if $E_t \le a^*$ then $E_{t+1} = \max(E_t, \min(D_t, a^*)) \le a^*$. □

The result does not say that anyone makes escalating demands. It says that if demands vary, if some exceed precedent by a non-trivial margin with non-trivial frequency, and if granting a demand turns it into the new floor, then an observer who always yields reaches the limit of their capacity in bounded expected time. Nothing in the model requires ill will, a psychiatric diagnosis, or even a particularly demanding person. A coworker whose deadlines slip, a landlord whose requests accumulate, and a relative whose needs rise will all produce the same ratchet if each concession becomes a new norm. The corollary shows the remedy and its price. A limit set in advance and held is the only policy in this family that preserves capacity. Declining some demands is not a failure of compassion. In this model it is the condition under which the modeled capacity is never reached, which is what allows the observer to keep helping. The conclusion concerns the modeled accommodation level and not the observer's total psychological, financial, or relational costs.

The assumptions do most of the work here, and it is worth saying which. The model describes an irreversible accommodation process, in which precedent never decays and demands never fall below it, so it establishes the consequences of that process and not the inevitability of escalation in real relationships. If demands do not track precedent the conclusion fails. Suppose $D_t \le B$ almost surely, that the observer yields, and that precedent decays by a factor $\lambda \in (0, 1]$ when it is not renewed, $E_{t+1} = \max(\lambda E_t, A_t)$. By the same induction as in Corollary 2, $E_t \le \max(E_0, B)$ for all t, so a capacity $C > \max(E_0, B)$ is never reached. This variant shows only that the conclusion fails when demands have an exogenous bound below capacity, and it does not isolate the effect of decay, which would require accommodation sometimes to fall below existing precedent and demands to respond gradually to it. The present model deliberately states the irreversible case, and a model with decay would be a distinct piece of mathematical work. Whether real relationships resemble the irreversible case or the decaying one is an empirical question, and one reason to state limits in advance is that doing so makes the irreversible case less likely.

A limit that is invented in the middle of a crisis behaves differently, for it cannot be distinguished from a refusal made in anger, and it will be experienced by the other person as an

abrupt reversal. Limits are best stated in calm periods, framed as descriptions of the observer's own conduct ("I will end the call if the shouting continues, and call back tomorrow") and not as demands on the other person's, and applied consistently. The consistency matters, because a limit that is sometimes held and sometimes waived teaches that the demand is worth pursuing.

The two mechanisms of this essay should be kept apart, because they can occur separately. Excuse inflation concerns belief: it describes how an observer's assessment can resist correction. The ratchet concerns behavior: it describes how accommodation can become progressively more costly even when the observer's beliefs are accurate. A person may recognize perfectly well that a relationship is unreliable and keep accommodating it because the immediate cost of refusal is high, and another may hold an unjustifiably favorable interpretation without making costly concessions. The first needs a discipline of evidence, the second a limit set in advance, and a relationship may need one, the other, or both. The two may also reinforce one another, since a charitable reading increases accommodation and accommodation raises the cost of later refusal, which in turn gives a reason to preserve the charitable reading. A coupled model is a natural direction for later work and is not attempted here.

10 Bipolar Disorder: Episodes, Agreements, and the Well Interval

For bipolar disorder the temporal structure noted in Section 2 reorganizes the problem. Conduct during an episode of elevated or depressed mood may differ sharply from the person's conduct otherwise, and the person may regard what they did during an episode with regret. The observer's difficulty is to avoid two opposite mistakes. One is to treat conduct during an episode as the person's character, which is unfair to the person and wrong about the forecast. The other is to treat the existence of the illness as a reason to absorb every consequence of an episode, including consequences that fall on others, such as spending, commitments, and words said in anger.

The framework suggests how to separate them. Proposition 4 says that causes matter through prediction, and here the cause is highly informative about prediction: if conduct belongs to an episode, then it can be expected to change when the episode ends, and the person, once well, can take part in decisions about what should happen next time. This points to the most useful practical device in the whole discussion, which is the agreement made in a well interval. During a stable period, the person and the people close to them can decide together what the early signs of an episode are, who will raise the concern, what the person would want others to do, which decisions should be postponed (large purchases, resignations, the ending of relationships), and which clinicians should be contacted. An agreement of this kind has an unusual property: it is the person's own considered preference, expressed while well, applied when the person is not. It converts the observer's dilemma into a shared plan and relieves the observer of the need to decide, in the middle of the episode, whether conduct is illness or character.

The evidence for the underlying orientation is respectable, although it concerns family involvement in treatment and not the specific device. In a randomized trial of 101 patients, family-focused therapy, which consisted of 21 sessions of psychoeducation, communication training, and problem-solving skills training and was combined with medication, produced

fewer relapses and longer survival without relapse than a brief crisis management intervention, and it was associated with greater symptom reduction and better medication adherence (Miklowitz et al., 2003). The English guideline for bipolar disorder likewise contains recommendations for families and carers (National Institute for Health and Care Excellence, 2014). The agreements described above are an application of the same orientation and have not, to the knowledge of the essay, been tested as a separate intervention in the sources checked here, so they are presented as a reasonable practice and not as an established treatment effect. A partially analogous and better studied device is the psychiatric advance directive. In a randomized trial of 469 patients with severe mental illness, a facilitated session raised completion of an advance directive or proxy authorization to 61 percent, against 3 percent in a control group (Swanson et al., 2006). The trial tested completion and short-term effects on working alliance and treatment satisfaction, and not whether directives change what happens during an episode, so it shows that such plans can be made and not that they work. The collaborative plans described above are informal agreements, whereas a psychiatric advance directive is a formal document whose legal force depends on the jurisdiction.

Recognizing an episode is not always easy, and the review of the disorders notes that patients and families often do not appreciate the significance of hypomanic or manic symptoms (Vieta et al., 2018). That cuts in two directions. It argues for humility on the part of a relative who feels certain they can see an episode coming, and it argues against the opposite certainty that everything unusual is illness. The practical test is conduct and course: whether the change is out of character, persists for days, and tracks the signs the person and the clinicians have identified. When the answer is not obvious, the right step is usually to raise it with the person in a non-accusatory way and with their clinician, and the wrong one is to announce a diagnosis.

The framework also helps with a harder case, which is conduct that falls on others and that the person later regrets. Sincere regret is a fact about the person and does not undo the cost. Proposition 1 applies in its plain form: the act can be fully explained by the episode and still impose costs that the observer is not required to absorb. What the observer owes is a response in proportion to the situation, such as continuing the relationship while declining to be the guarantor of debts incurred during the episode, and this is compatible with warmth.

11 Borderline Personality Disorder: Validation Without Capitulation

The practical lesson of the clinical literature on BPD is that two things must be held together which ordinary intuition treats as rivals. One is that the person’s emotional experience, including the sense of threat and abandonment, is real and is not to be dismissed. The other is that the interpretation the person places on events may be mistaken and the conduct that follows may be costly. Dialectical behavior therapy is organized around the combination of acceptance and change, and as the term is used in that tradition, validation means communicating that a person’s responses make sense given their experience and context (Linehan, 1993). Validation is not agreement with the factual claim. It is possible to say, sincerely, that being afraid of losing someone makes sense, and also to say, without contradiction, that the person has not been abandoned and that a particular demand cannot be met.

The research on emotional functioning gives the practice a rationale. If the emotional difficulty in BPD includes heightened sensitivity and labile negative affect (Carpenter and Trull, 2013), and if interpersonal threat cues carry unusual weight (Staebler et al., 2011), then a reaction that appears out of proportion to its trigger may be proportionate to the trigger as perceived. The observer who understands this can respond to the perceived threat ("I can see that when the call did not come it felt like I had disappeared") without ratifying the interpretation. The review itself presents the expectation that events subtle or seemingly innocuous to others should produce rapid changes in mood in people high in emotion sensitivity as an inference from the model and not as a directly demonstrated result. The supporting evidence mixes self-report, laboratory, and daily-diary studies, some findings are inconsistent, and it describes groups, so the statement for any individual is a hypothesis to be tested against that person's own pattern.

The same literature argues for caution about the way the observer thinks about causes. The framing most damaging to relationships and to the person is a motive-based one in which every difficulty is read as manipulation. The stigma review notes that people with personality disorders are often perceived as purposefully misbehaving and that this attitude is common among providers (Sheehan et al., 2016). Proposition 4 offers an alternative that does not require a theory of motive in either direction. The conduct is what it is, the cost is what it is, and the limit can be set without an accusation. A relative can say, "I understand that this is distressing, and I cannot take calls after midnight," and both halves of the sentence can be true without any claim about intent.

Support for relatives exists and has some evidence behind it. Family Connections is a twelve-week manualized education program for relatives of people with BPD, led by trained family members and based on the strategies of dialectical behavior therapy; in an uncontrolled study of 44 participants from 34 families, grief and burden fell and mastery rose from before to after the program and the changes were maintained at six months (Hoffman et al., 2005). The design has no comparison group and relies on self-report, so the result shows that participants improved, not that the program caused it. The recovery data deserve emphasis for the opposite reason. In the one long prospective cohort reviewed here, most patients reached sustained symptomatic remission, a smaller share reached full recovery including good social and vocational functioning, and relapse was more common than in comparison patients (Zanarini et al., 2012). The cohort was identified at an inpatient admission and cannot be assumed to represent everyone with the diagnosis. It supports the more limited view that current conduct is not a permanent forecast, and that is a reason to keep the door open where it is safe to do so and to avoid the fatalism that treats present difficulty as fixed.

12 A Practice of Bounded Charity

The propositions can be turned into a short set of working principles. They are offered as practical judgment informed by the results above, and each is labeled with what supports it.

1. **Separate the four questions.** Ask separately whether the act is intelligible, whether the feeling is sincere, whether the belief is accurate, and what the act costs (Table 1). Answering one does not answer the others. *Support:* Proposition 1.

2. **Keep a count.** Record surprising or costly events as they occur, in a short note with the date. The purpose is not to build a case but to prevent the tally from being reconstructed from a memory bent by the latest explanation. The count is a private aid to memory and not a dossier. It should not be produced as ammunition in a dispute, and if the other person finds it, it is best described honestly as what it is, a way of keeping the observer's own account accurate. Used otherwise it becomes surveillance and defeats its purpose. A record also establishes what happened and is not a statistical sample: events need not be independent, equally consequential, or drawn from a stationary process, so a count is not a calibrated estimate of reliability, and the arithmetic of Corollary 1 is an illustration and not a measurement. *Support:* Proposition 3, parts (a) and (b); the practice itself has not been tested.

3. **Ration excuses by evidence.** Give weight to explanations supported by information the observer actually has, and treat merely conceivable explanations as conceivable. *Support:* Proposition 3.

4. **Decide in advance what would change your conduct.** State, privately and in calm, what pattern or number of events would lead to a change in contact, and hold to it. *Support:* Proposition 3(a); Corollary 1 for the exchange rate of goodwill.

5. **Set limits in calm and describe them as your own conduct.** Frame limits as what you will do, not as what the other person must do, and apply them consistently. *Support:* Proposition 5 and Corollary 2.

6. **Validate feelings, decline false facts.** Acknowledge that the person's reaction makes sense given their experience, and say plainly what you know happened. *Support:* the validation concept in dialectical behavior therapy (Linehan, 1993); Proposition 1.

7. **Support treatment and do not impersonate it.** Encourage access to evidence-supported care and, for relatives, to family programs. *Support:* guideline and trial evidence for structured psychotherapies in BPD (American Psychiatric Association, 2024a; Storebø et al., 2020) and for family involvement in bipolar disorder (Miklowitz et al., 2003).

8. **Make agreements visible, and make well-interval agreements where they apply.** A written summary of decisions distributes the memory. In bipolar disorder, plans made in well periods can govern behavior during episodes. *Support:* reasoned practice; family involvement is supported (Miklowitz et al., 2003), the specific devices are not tested in the sources checked.

9. **Protect your own capacity.** The burden on relatives is documented and heavy (Bailey and Grenyer, 2013). Sustaining help for years requires keeping something in reserve. *Support:* Proposition 5, Corollary 2.

10. **Do not let past investment decide the future.** The thought "I have already given so much" is a sunk-cost argument, and a long investment is a reason to be careful about changing course, not a reason to continue (Arkes and Blumer, 1985). Weigh what the arrangement will cost and offer from here on. *Support:* the sunk-cost literature, applied by analogy.

11. **Treat risk to life as a separate category.** Statements about suicide or self-harm are always

taken seriously, and the response is proportionate to immediacy and severity: emergency services when danger is imminent, a crisis line (9-8-8 in Canada and the United States) for support and guidance, any existing safety plan, and the person's clinicians. If a relationship involves threats or violence, safety planning with professionals takes priority over all the principles above.

None of these principles recommends ending a relationship, and none recommends keeping one. They recommend that the decision be made on the basis of conduct, cost, and capacity, with compassion intact. A relationship may well continue, at a different distance and on different terms, for decades. A relationship may also need to end or to be limited sharply, as when a person's wellbeing or the welfare of children is at stake, and the framework supports that choice without requiring a verdict on the other person's character. It is worth noticing that distance and abandonment are different things. A person who reduces contact, tells the other person clearly why and under what conditions it could change, and remains reachable for emergencies has not abandoned anyone, though it may be experienced that way, and that experience is itself a fact to be met with validation.

13 Beyond Psychiatry

The mechanism does not need a clinical setting. A workplace in which a manager's contradictory instructions are always excused by the pressure they are under, an institution whose repeated failures are always treated as exceptional because a good explanation is always available, and a committee that tolerates the same disruption because each case has a reason, all instantiate Proposition 3(b). A landlord or a vendor whose accommodations accumulate into a standing expectation instantiate Proposition 5. The reason the argument deserves wide application is that it explains a familiar social failure without needing a villain. Nobody has to be lying for a group of reasonable, kind, well-informed people to keep granting an exception for years.

Consider a regulator that oversees a licensed operator with a recurring compliance problem. After each incident the operator supplies an explanation: a supplier delay, a staffing gap, an unforeseeable weather event. Each explanation is plausible and some are verified. If the regulator reasons as in Proposition 3(b), fixing the weight of each explanation after receiving it, then repeated incidents never accumulate into a finding, because each is evaluated alone with its own reason attached. The discipline of part (a) corresponds to a rule fixed in advance: that a stated number of incidents of a given class within a stated period triggers a review regardless of the reasons offered, with reasons assessed at the review and not at the incident. The same structure appears in committees, where a standing disruption is excused case by case, and the corrective is identical, a threshold stated before the next case arrives. Such rules are sometimes experienced as uncharitable, and the framework says why they are not. They do not deny that reasons exist. They decide in advance how much weight reasons receive, so that charity is not recomputed under pressure.

There is a related point about environments. Recommendation systems, bureaucracies, and platforms that always supply a plausible rationale for their outputs exploit the same faculty, and the person on the receiving end faces the same inference problem: a rationale is always

available and says little about whether the output was warranted. The defense is also the same: ask what independent evidence supports the rationale, and decide on conduct and consequences.

14 Limits, Risks, and Misreadings

Several limits constrain the argument and the reader should hold them in mind.

First, the propositions are mathematical results about models. Proposition 2 assumes that verification is perfect and has a fixed cost; Corollary 1 assumes stationarity and exchangeability of commitments; Proposition 3 compares two stylized hypotheses and assumes that likelihoods can be compared; Proposition 4 is a statement about decision under equal forecasts; and Proposition 5 assumes that accommodation sets precedent and that demands vary within a bounded range. Real relationships violate every one of these assumptions. The results show what follows if the assumptions hold, and which assumption to examine if the conclusion looks wrong. They do not show that any particular relationship is in the situation described.

Second, the empirical base for the specific claim of the essay, that charitable explanation by observers prevents updating in relationships with people who have these conditions, is thin. The sources reviewed document burden on carers, treatment effects, outcome trajectories, and the psychology of attribution, and none of them tests the framework itself. The essay should be read as a theoretical proposal with a mathematical skeleton and as a pointer to a research program, and not as a finding. Table 3 states what each claim rests on.

Third, the essay's descriptions of clinical sources rest mainly on abstracts and bibliographic records, which were checked. Two sources, Carpenter and Trull (2013) and Staebler and colleagues (2011), were also read in full, and the text was adjusted to reflect the qualifications they state. The claims drawn from them are stated at the level the abstracts support, and the effect sizes quoted come from those abstracts. Readers who intend to rely on a particular result should consult the full source.

Fourth, there is a risk of misreading in two directions. The framework could be heard as permission to withdraw sympathy from people with psychiatric conditions, which is the opposite of the intent. The review of stigma documents how readily these diagnoses are attached to the idea of deliberate misbehavior (Sheehan et al., 2016), and the framework is designed to avoid dependence on any claim about motive. It could also be heard as a general suspicion of charity, which is also unintended. Charity is indispensable, and the argument is for charity with a memory.

Fifth, the essay is not clinical guidance, and relatives in a serious situation should seek help from clinicians and from family programs. Risk to life is outside the scope of the principles and is handled in Section 12.

15 Conclusion

The thesis can be stated briefly. Understanding why someone behaves as they do is not the same as establishing what others must tolerate. Explanation is a service the mind performs for itself and for people it values, and it is usually performed well. Tolerance is a decision about conduct, cost, and capacity, and it should be taken on the evidence of conduct. When the two are fused, the observer can find that every surprising act has an explanation and that the explanations have, together, consumed the relationship.

Five results support this claim in a narrow, checkable sense. Explanatory possibility does not entail justification (Proposition 1). Default trust is rational above a threshold that the observer rarely notices crossing, and goodwill has an exchange rate (Proposition 2 and Corollary 1). Excuses chosen after the evidence arrives remove the evidence's force, while excuses fixed in advance preserve it (Proposition 3). Decisions grounded in conduct are insensitive to disagreements about cause that do not change the forecast, which spares the observer the burden of becoming an amateur psychiatrist (Proposition 4). And accommodation that becomes precedent exhausts a finite capacity in bounded expected time unless a limit is set in advance and held (Proposition 5 and Corollary 2).

Applied to borderline personality disorder and bipolar disorder, the framework asks for a particular combination of qualities. It asks for knowledge that the conditions are different, that treatment exists and works for many people, that most people improve over time, and that the stigma attached to these diagnoses is itself a harm. It asks that the person's experience be taken as real, and that the observer keep their own account of what happened. It asks for agreements that are written down, for limits that are stated in calm and held, and for a clear separation of the question of what a person is going through from the question of what an arrangement can bear. The observer who does these things has not stopped being charitable. They have made charity sustainable, which is a different and better thing from endless forbearance, because forbearance that cannot be sustained ends in a rupture that serves no one, including the person it was meant to protect.

References

American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Association Publishing, Washington, DC, 5th edition, text revision, 2022.

American Psychiatric Association. *The American Psychiatric Association Practice Guideline for the Treatment of Patients With Borderline Personality Disorder*. American Psychiatric Association Publishing, 2024. doi:10.1176/appi.books.9780890428009.

American Psychiatric Association. Updated borderline personality disorder guideline. News release, 10 December 2024.

H. R. Arkes and C. Blumer. The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, 35(1):124–140, 1985.

R. Axelrod and W. D. Hamilton. The evolution of cooperation. *Science*, 211(4489):1390–1396, 1981.

R. C. Bailey and B. F. S. Grenyer. Burden and support needs of carers of persons with borderline personality disorder: A systematic review. *Harvard Review of Psychiatry*, 21(5):248–258, 2013.

A. Bateman and P. Fonagy. Randomized controlled trial of outpatient mentalization-based treatment versus structured clinical management for borderline personality disorder. *American Journal of Psychiatry*, 166(12):1355–1364, 2009.

R. W. Carpenter and T. J. Trull. Components of emotion dysregulation in borderline personality disorder: A review. *Current Psychiatry Reports*, 15(1):335, 2013.

I. Choi, R. E. Nisbett, and A. Norenzayan. Causal attribution across cultures: Variation and universality. *Psychological Bulletin*, 125(1):47–63, 1999.

D. Davidson. Radical interpretation. *Dialectica*, 27(3–4):313–328, 1973.

D. T. Gilbert and P. S. Malone. The correspondence bias. *Psychological Bulletin*, 117(1):21–38, 1995.

P. D. Hoffman, A. E. Fruzzetti, E. Buteau, E. R. Neiditch, D. Penney, et al. Family Connections: A program for relatives of persons with borderline personality disorder. *Family Process*, 44(2):217–225, 2005.

P. Johansson, L. Hall, S. Sikström, and A. Olsson. Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, 310(5745):116–119, 2005.

E. E. Jones and V. A. Harris. The attribution of attitudes. *Journal of Experimental Social Psychology*, 3(1):1–24, 1967.

M. M. Linehan. *Cognitive-Behavioral Treatment of Borderline Personality Disorder*. Guilford Press, New York, 1993.

M. M. Linehan, K. A. Comtois, A. M. Murray, M. Z. Brown, R. J. Gallop, H. L. Heard, et al. Two-year randomized controlled trial and follow-up of dialectical behavior therapy vs therapy by experts for suicidal behaviors and borderline personality disorder. *Archives of General Psychiatry*, 63(7):757–766, 2006.

B. F. Malle. The actor-observer asymmetry in attribution: A (surprising) meta-analysis. *Psychological Bulletin*, 132(6):895–919, 2006.

D. J. Miklowitz, E. L. George, J. A. Richards, T. L. Simoneau, and R. L. Suddath. A randomized study of family-focused psychoeducation and pharmacotherapy in the outpatient management of bipolar disorder. *Archives of General Psychiatry*, 60(9):904–912, 2003.

S. L. Murray and J. G. Holmes. Seeing virtues in faults: Negativity and the transformation of interpersonal narratives in close relationships. *Journal of Personality and Social Psychology*, 65(4):707–722, 1993.

S. L. Murray, J. G. Holmes, and D. W. Griffin. The benefits of positive illusions: Idealization and the construction of satisfaction in close relationships. *Journal of Personality and Social Psychology*, 70(1):79–98, 1996.

National Institute for Health and Care Excellence. *Bipolar disorder: assessment and management*. Clinical guideline CG185, 2014. Updated February 2016.

R. E. Nisbett and T. D. Wilson. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3):231–259, 1977.

L. Ross. The intuitive psychologist and his shortcomings: Distortions in the attribution process.

In L. Berkowitz, editor, *Advances in Experimental Social Psychology*, volume 10, pages 173–220. Academic Press, New York, 1977.

J. Paris, J. Gunderson, and I. Weinberg. The interface between borderline personality disorder and bipolar spectrum disorders. *Comprehensive Psychiatry*, 48(2):145–154, 2007.

L. Sheehan, K. Nieweglowski, and P. Corrigan. The stigma of personality disorders. *Current Psychiatry Reports*, 18(1):11, 2016.

D. Sperber, F. Clément, C. Heintz, O. Mascaro, H. Mercier, G. Origgi, and D. Wilson. Epistemic vigilance. *Mind & Language*, 25(4):359–393, 2010.

K. Staebler, E. Helbing, J. Rosenbach, and B. Renneberg. Rejection sensitivity and borderline personality disorder. *Clinical Psychology & Psychotherapy*, 18(4):275–283, 2011.

O. J. Storebø, J. M. Stoffers-Winterling, B. A. Völlm, M. S. Kongerslev, J. T. Mattivi, M. S. Jørgensen, et al. Psychological therapies for people with borderline personality disorder. *Cochrane Database of Systematic Reviews*, (11):CD012955.pub2, 2020.

J. W. Swanson, M. S. Swartz, E. B. Elbogen, R. A. Van Dorn, J. Ferron, et al. Facilitated psychiatric advance directives: A randomized trial of an intervention to foster advance treatment planning among persons with severe mental illness. *American Journal of Psychiatry*, 163(11):1943–1951, 2006.

E. Vieta, M. Berk, T. G. Schulze, A. F. Carvalho, T. Suppes, et al. Bipolar disorders. *Nature Reviews Disease Primers*, 4:18008, 2018.

M. C. Zanarini, F. R. Frankenburg, J. Reich, and G. Fitzmaurice. Attainment and stability of sustained symptomatic remission and recovery among patients with borderline personality disorder and axis II comparison subjects: A 16-year prospective follow-up study. *American Journal of Psychiatry*, 169(5):476–483, 2012.

Claim	Source	Evidential status
On average, psychotherapy reduces BPD severity compared with usual care	Storebø et al. (2020)	Review of 75 trials; moderate-quality evidence for severity, low-quality for self-harm and suicide outcomes
Structured psychotherapy recommended; no therapy a gold standard; medication time-limited and alongside psychotherapy, none found effective for core symptoms	American Psychiatric Association (2024a,b)	Practice guideline; chair's statement in the announcement
DBT outperformed community expert treatment on suicide attempts	Linehan et al. (2006)	One randomized trial, 101 women
Remission common, recovery less so, in BPD	Zanarini et al. (2012)	Prospective cohort of inpatients over 16 years
Family-focused therapy plus medication reduced bipolar relapse	Miklowitz et al. (2003)	One randomized trial, 101 patients
Carer burden is elevated	Bailey and Grenyer (2013)	Systematic review, six studies, 465 carers
Family Connections reduced grief and burden	Hoffman et al. (2005)	Uncontrolled pre-post, 44 participants, self-report
Higher rejection sensitivity in BPD samples	Staebler et al. (2011)	Cross-sectional, self-report, 46 BPD patients; authors call it preliminary
Advance directives can be completed with facilitation	Swanson et al. (2006)	One randomized trial, 469 patients; tested completion, not episode outcomes
Actor–observer asymmetry is weak overall	Malle (2006)	Meta-analysis of 173 studies
Propositions 1 to 4	This essay	Proved, conditional on stated model assumptions
Proposition 5 and Corollary 2	This essay	Proved under stronger assumptions: precedent never decays and demands never fall below it
Practical principles	This essay	Reasoned judgment, not tested

Table 3: Evidential status of the principal claims.