Predictive coding & depression

by **steve2152** 6 min read 2nd Jan 2020 9 comments

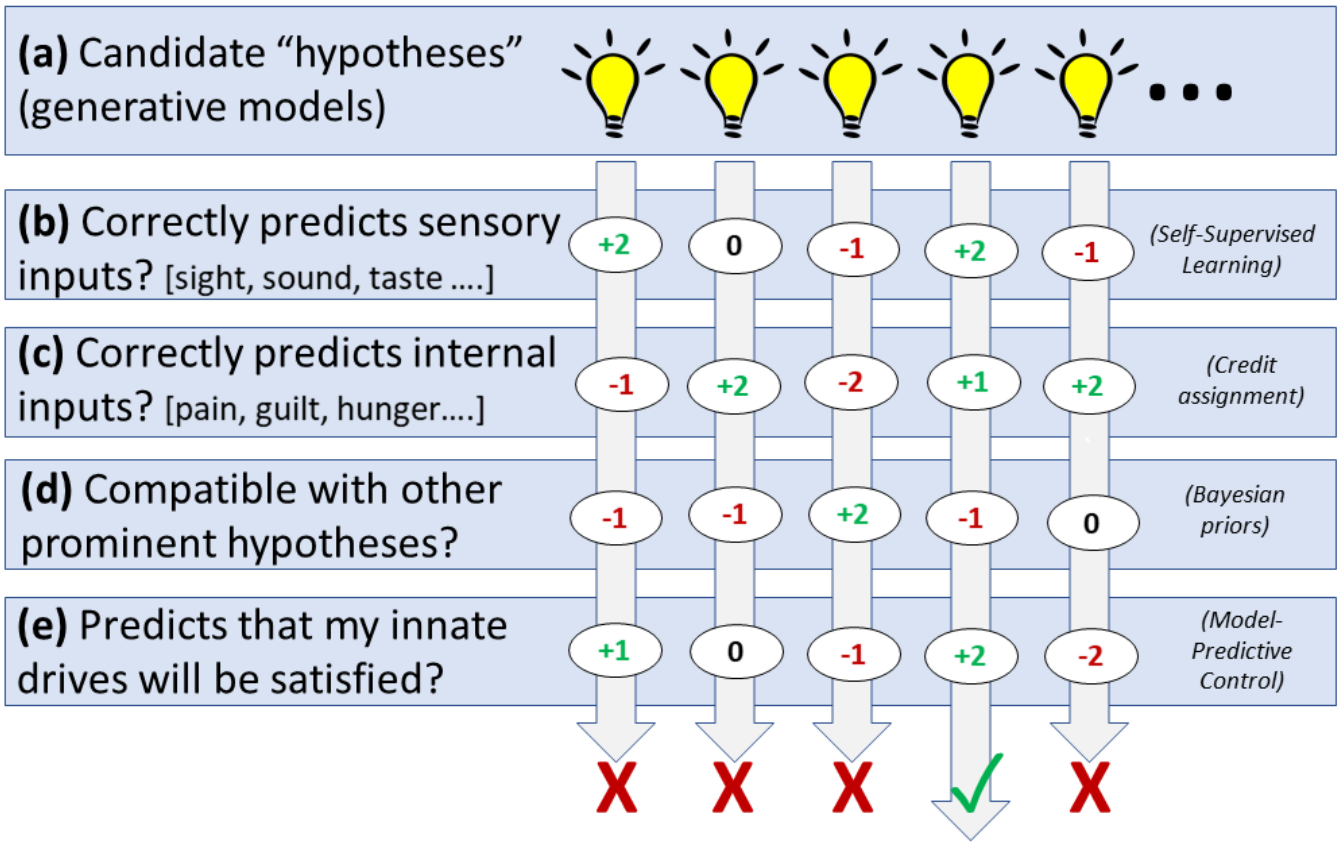
19

Epistemic status: wild speculation.

This is a follow-up to the 2017 blog post Slate Star Codex: Toward a Predictive Theory of Depression. I think that post has fundamentally the right idea, but that it had some pieces of the puzzle that didn't quite fit satisfyingly.

Well I personally have no special knowledge about depression (happily!), but I have thought an awful lot about the how predictive coding, motivation, and emotions interact in the brain. So I think that I can flesh out Scott's basic story and make it more complete, coherent and plausible. Here goes!

Let's start with my cartoon of predictive coding as I see it:



For lots of details and discussion and caveats, see Predictive coding = RL + SL + Bayes + MPC .

In fact, don't bother reading on here until you've read that.

...

... done? OK, let's move on.

One more piece of background: In predictive coding, hypotheses have a "strength" term for the various predictions they make—basically, the confidence that this prediction will come true. If a strong prediction is erroneous, that's likelier to rise to attention, and to cause the offending hypothesis to be discarded or modified. So let's say you have the hypothesis:

Hypothesis: The thing I am looking at is a TV screen showing static.

This hypothesis...

- ...makes weak predictions about visual inputs within the screen area (where there's static),
- ...makes strong predictions about visual inputs within the thin black frame of the TV,
- ...makes infinitely weak (strength-zero) predictions of sensations coming from my foot.

The sensations from my foot are just not part of this hypothesis. Thus a separate hypothesis about my foot—say, that my foot is tapping the floor and will continue to do so—can be active simultaneously, and these two hypotheses won't conflict.

The proposed grand unified theory of depression in this post is the same as in Scott's post: Severe depression is when all the brain's hypotheses make anomalously weak predictions

I'm now going to go through the symptoms of depression listed in Scott's article, plus a couple others he left out:

Perception

> ... Depressed people describe the world as gray, washed-out, losing its contrast.

This one is just like Scott says. Normally we understand an apple via a hypothesis like "This thing is a bright red apple". But in severe depression, the hypotheses are weaker: "This thing is probably an apple, and it's probably reddish".

Psychomotor retardation

Again, this one is just like Scott says. In predictive coding theory, top-down predictions about proprioceptive inputs are exactly the same signals as (semi)low-level motor control commands.[1]

If a hypothesis makes predictions about proprioceptive inputs with strength zero, then the muscles don't move at all, like the TV static + foot example above. If it makes predictions with the normal high strength, then the muscles move normally. Thus, somewhere in between, there's a threshold where a hypothesis is making a very feeble prediction of proprioceptive inputs, that can just barely and unreliably activate the downstream motor programs.

Lack of motivation

Hypotheses make predictions about both sensory inputs (vision etc.) and internal inputs (satiety, pain, etc.); I don't think there's any mechanistic difference between those two types of predictions. (By the same token, even though I drew (b) and (c) as separate rows in the picture above, they're actually implemented by the same basic mechanism.)

In depression, the hypotheses' predictions about internal inputs have anomalously low strength, just like all their other predictions. Thus instead of a normal hypothesis like "I will eat and then I will feel full", when depressed you get stuck with the weak hypothesis "I will eat and then I *might* feel full."

Now look at process (e). I claim that process (e) does not change at all in depression: we still have the normal innate drives, and we still favor hypotheses which predict that those drives will be satisfied. But (e) votes much less strongly for the weak hypothesis ("I will eat and then I *might* feel full") than for the normal hypothesis ("I will eat and then I *will* feel full"). Now, I think that we universally have a bias towards inaction—don't take an action unless there's a good reason to. So in depression, the bias towards inaction often wins out over the feeble vote from process (e), and thus we don't bother to get up and eat.

Low self-confidence

All the predictive coding machinery operates basically the same way during (i) actual experience, (ii) memory recall, and (iii) imagination. Now, when we try to figure out, "Will I succeed at X?", we imagine/plan doing X, which entails searching through the space of hypotheses for one in which X successfully occurs at the end. In depression, none of the hypotheses will make a strong prediction that X will occur, because of course they don't make strong predictions of anything at all. We interpret that as "I cannot see any way that I will succeed; thus I will fail".

Why isn't it symmetric? None of the theories are making strong predictions of failure either, right? Well, I think that generically when people are planning out how to do something, they search through their hypothesis space for a hypothesis that has a specific indicator of success; they don't search for a hypothesis that *lacks* a specific indicator of failure. I just don't think you can run the neural hypothesis search algorithm in that opposite way.[2]

There's another consideration that points in the same direction. Everyone always lets their current mood leak into their memories and expectations: when you're happy, it's tricky to remember being sad, etc. I think this is just because episodic memories and other hypotheses leave lots of gaps that we fill in with our current selves. Anyway, I hypothesize that this effect is *even stronger* when depressed: if you feel sad right now, *and* if none of your hypotheses contain a strong prediction of feeling happiness or any other emotion, well then there's no emotion to be found except sadness in your remembered past, or in your present, or in your imagined future. So, ask such a person whether they'll succeed, and they might answer the question by just trying to imagine themselves feeling the joy of victory. They can't. Ask whether they'll fail, and they'll try to imagine

themselves feeling miserable. Well, they can do that! They are feeling miserable feelings right now, and those feelings can flood into an imagined future scenario. (See Availability heuristic.)

So, putting these two considerations together, a depressed person should have a very hard time imagining a specific course of events in which they accomplish anything in particular, and should have an equally hard time imagining themselves having proudly accomplished it. But they can *easily* imagine themselves not getting anything done, and continuing to feel the same misery that they feel right now. I think that's a recipe for low self-confidence.

Feelings of sadness, worthlessness, self-hatred, etc.

Everything about predictive coding happens in the cortex.[3] But when we talk about mood, we need to bring in the amygdala.[4] My hypothesis is that the amygdala is functioning normally (i.e., according to specification) in depression. The problem lies in the signals it receives from the cortex.

Let's step back for a minute and talk about the interesting relationship between the cortex and amygdala. I discussed it a bit in Human instincts, symbol grounding, and the blank-slate neocortex . The amygdala is nominally responsible for emotions, yet figuring out what emotions to feel requires excruciatingly complex calculations that only the cortex is capable of. ("No, YOU were supposed to do the dishes, because remember I went shopping three times but you only vacuumed once and...").

I think the cortex-amygdala relationship for emotions is somewhat analogous to the cortex-muscle relationship for motor control: The cortex sends signals to the amygdala, which are "predictions" from the cortex's perspective and "input information" from the amygdala's perspective. The amygdala uses that information, along with other sources of information (pain, satiety, etc.), to decide what emotions to emit. (See this comment° for more.)

Back to depression. Again, outgoing signals from the cortex are synonymous with top-down predictions. So when all the top-down predictions get weaker, simultaneously the

signals from the cortex to the amygdala get globally weaker.

Then what?

In principle, you could imagine two worlds. In one extreme world, the messages from the cortex to the amygdala only contain bad news—I've been insulted, that's disgusting, I'm in trouble, etc. Then, when the cortical signals get globally weaker, the amygdala would flood us with joy. Everything is perfect!

In the opposite extreme world, the messages from the cortex to the amygdala only contain good news—I'm not being insulted right now, I am not disgusted right now, I have lots of friends, etc. Then, when the cortical signals get globally weaker, the amygdala would flood us with misery. Everything must be terrible!

In the real world, the cortex presumably sends both good-news messages and bad-news messages to the amygdala. But unfortunately for those suffering depression, it seems that all things considered, "no news is bad news". Presumably the amygdala is especially expecting the cortex to send various good-news signals that say that our innate drives are being satisfied, that we have high status, that we're expecting more good things to happen in the future, etc. When these signals get weaker, the amygdala responds by emitting all sorts of negative emotions.

Evolutionarily speaking, what is sadness for? In my view, sadness has an effect of causing us to abandon our plans when they're not working out, and it also has a social effect of signaling to others that we are in a bad, unsustainable situation and need help. Well, the feeling that our innate drives are not being satisfied, and will not be satisfied in the foreseeable future ... that sure seems like it ought to precipitate sadness if anything does, right?

Difficulty thinking and concentrating

Prediction strength and attention are closely related. If you want to attend to the taste of the thing your eating, your brain modifies the active hypothesis to have super high strength on the predictions of sensory inputs from your taste buds. Then any slight deviation of the actual sensory input from expectations will trigger a prediction error and bubble up to top-level attention.

When you string together a bunch of little hypotheses into an extended train of thought°, attention has to be deftly steered around, to get the right information into the right places in working memory, and manipulate them the right way. It follows that if your brain can't deploy hypotheses with sufficiently strong predictions, you will probably be unable to fully control your attention and think complex thoughts.

Insomnia

Beats me. I don't understand sleep!

Conclusion

I'm pretty pleased that everything seems to fit together, without *too* much special pleading! But again, this is wild speculation, I especially don't know anything about depression. I'm happy to hear feedback.

(Update: I replaced "precision of a prediction" with "strength of a prediction" throughout. I think that terminology is both more intuitive and more accurate.)

Addendum: Speculating on what causes depression

I think there has to be a vicious cycle involved in depression, otherwise it wouldn't persist. Here's my guess:

Vicious cycle: Globally weaker predictions cause sadness, and sadness causes globally weaker predictions.

I already talked about the first part. But why (evolutionarily) might sadness cause globally weaker predictions? Well, one evolutionary goal of sadness is to prompt us to abandon our current plans and strategies, even ones we're deeply tied to, in the event that our prospects are grim and we can't see any way to make things better. Globally weaker predictions would do that! As you weaken the prediction, "active plans" turn into

"possible plans", then into "vague musings"...

Anyway, maybe that vicious cycle dynamic is always present to some extent, but other processes push in other directions and keep our emotions stable. ...Until a biochemical insult—or an unusually prolonged bout of "normal" sadness (e.g. from grief)—tips the system out of balance, and we get sucked into that vortex of mutually reinforcing "sadness + globally weak predictions".

---

1. I say "semi-low-level" because the commands get further processed by the cerebellum etc. ↩
2. This has to do with neural algorithm implementation details that I won't get into. ↩
3. As usual, "cortex" here is technically short for "predictive-world-model-building-system involving primarily the cortex, thalamus, and hippocampus." ↩
4. As in the previous footnote, don't take the term "amygdala" too literally; I'm not sure how the functionality I'm talking about is divided up among the amygdala, hypothalamus, and/or other brain structures. ↩

[-] **Scott Alexander** 1mo 14

In this post and the previous one you linked to, you do a good job explaining why your criterion e is possible / not ruled out by the data. But can you explain more about what makes you think it's true? Maybe this is part of the standard predictive coding account and I'm just misunderstanding it, if so can you link me to a paper that explains it?

I'm a little nervous about the low-confidence model of depression, both for some of the reasons you bring up, and because the best fits (washed-out visual field and psychomotor retardation) are really marginal symptoms of depression that you only find in a few of the worst cases. The idea of depression as just a strong global negative prior (that makes you interpret everything

you see and feel more negatively) is pretty tempting. I like Friston's attempt to unify these by saying that bad mood is just a claim that you're in an unpredictable environment, with the reasoning apparently being something like "if you have no idea what's going on, probably you're failing" (eg if you have no idea about the social norms in a given space, you're more likely to be accidentally stepping on someone's toes than brilliantly navigating complicated coalitional politics by coincidence). I'm not sure what direction all of this happens in. Maybe if your brain's computational machinery gets degraded by some biochemical insult, it widens all confidence intervals since it can't detect narrow hits, this results in fewer or weaker positive hits being detected, this gets interpreted as an unpredictable world, and this gets interpreted as negative prior on how you're doing?

[-] **steve2152** 1mo 1

Thanks for the comment!

> In this post and the previous one you linked to, you do a good job explaining why your criterion e is possible / not ruled out by the data. But can you explain more about what makes you think it's true?

Maybe the reason for (e) would be more clear if you replace "hypothesis" with "possible course of action". Then (e) is the thing that makes us more likely to eat when we're hungry, etc.

("Course of action" is just a special case of what I call "hypothesis". "Hypothesis" is synonymous with "One possible set of top-down predictions".)

I don't think I'm departing from "Surfing Uncertainty" etc. in any big way in that previous post, but I felt that the predictive coding folks don't adequately discuss how the specific hypotheses / predictions are actually calculated in the brain. I might have been channeling the Jeff Hawkins 2004 book a bit to fill in some gaps, but it's mainly my take.

I guess I should contextualize something in my previous post: I think *anyone* who advocates predictive coding is obligated to discuss The Wishful Thinking Problem. It's not something specific to my little (a-e) diagram. So here is The Wishful Thinking Problem, stripped away from the rest of what I wrote:

Wishful thinking problem: *If we're hungry, we have a high-level prior that we're going to eat. Well, that prior privileges predictions that we'll go to a restaurant, which is sensible... but that prior also privileges predictions that food will magically appear in our mouths, which is wishful thinking. We don't actually believe the latter. So that's The Wishful Thinking Problem.*

The Wishful Thinking Problem is not a big problem!! It has an obvious solution: Our prior that "magic doesn't happen" is stronger than our prior that "we're going to eat". Thus, we don't expect food to magically appear in our mouth after all! Problem solved! That's all I was saying in that part of the previous post. Sorry if I made it sound overly profound or complicated.

> I like Friston's attempt to unify these by saying that bad mood is just a claim that you're in an unpredictable environment

I encourage you to think about it more computationally! The amygdala has a circuit that takes data, does some calculation, and decides on that basis whether to emit a feeling of disgust. And it has another circuit that takes data, does some calculation, and decides whether to emit a feeling of sadness. And so on for boredom and fear and jealousy and every other emotion. Each of these circuits is specifically designed by evolution to emit its feeling in the appropriate circumstances.

So pretend that you're Evolution, designing the sadness circuit. What are you trying to calculate? I think the short answer is:

Sadness circuit design goal: *Emit a feeling of sadness when: My prospects are grim, and I have no idea how to make things better.*

Why is this the goal, as opposed to something else? Because this is the appropriate situation to cry for help and rethink all your life plans.

OK, so if that's the design goal, then how do you actually build a circuit in the amygdala to do that? Keep in mind that this circuit is not allowed to directly make reference to our understanding of the world, because "our understanding of the world" is an inscrutable pattern of neural activity in a massive, convoluted, learned data structure in the cortex, whereas the emotion circuits need to have specific, genetically-predetermined neuron writing. So what can you do instead? Well, you can design the circuit such that it listens for the cortex to predict rewarding things to happen (the amygdala *does* have easy access to this information), and to *not* feel sadness when that signal is occurring regularly. After all, that signal is typically a sign that we are imagining a bright future. This circuit won't perfectly match the design goal, but it's as close as Evolution can get.

(By contrast, the algorithm "check whether you're in an unpredictable environment" doesn't seem to fit, to me. Reading a confusing book is frustrating, not saddening. Getting locked in jail for life is saddening but offers predictability.)

So anyway, my speculation here is that:

(1) a lot of the input data for the amygdala's various emotion calculation circuits comes from the cortex (duh),

(2) the neural mechanism controlling the strength of predictions also controls the strength of signals from the cortex to the amygdala (I think this is a natural consequence of the predictive coding framework, although to be clear, I'm speculating),

(3) a global reduction of the strength of signals going from the cortex to the amygdala affects pretty much all of the emotion circuits, and it turns out that the result is sadness and other negative feelings (this is my pure speculation, although it seems to fit the sadness algorithm example above). I don't think there's any particularly deep reason that globally weaker signals from the cortex to the amygdala creates sadness rather than happiness. I think it just comes out of details about how the various emotion circuits are implemented, and interact.

(The claim "depression involves global weakening of signals going from cortex to amygdala" seems like it would be pretty easy to test, if I had a psych lab. First try to elicit disgust in a

way that bypasses the cortex, like smelling something gross. Then try to elicit disgust in a way that requires data to pass from the cortex to the amygdala, like remembering or imagining something gross. [Seeing something gross can be in either category, I think.] My prediction is that in the case that doesn't involve cortex, you'll get the same disgust reaction for depressed vs control; and in the case that *does* involve cortex, depressed people will have a weaker disgust reaction, proportional to the severity of the depression.)

> the best fits (washed-out visual field and psychomotor retardation) are really marginal symptoms of depression that you only find in a few of the worst cases

I guess that counts against this blog post, but I don't think it quite falsifies it. Instead I can claim that motor control works normally if the cortical control signals are above some threshold. So the signals can get somewhat weaker without creating a noticeable effect, but if they get *severely* weaker, it starts butting against the threshold and starts to show. (The motor control signals do, after all, get further processed by the cerebellum etc.; they're not literally controlling muscles themselves.) Ditto for washed-out visual field; the appearance of a thing you're staring at is normally a *super* strong prediction, maybe it can get somewhat weaker without creating a noticeable effect. Whereas maybe the amygdala is more sensitive to relatively small changes in signal levels, for whatever reason. (This paragraph might be special pleading, I'm not sure.)

There are two perspectives. One is "Let's ignore the worst of the worst cases, their brains might be off-kilter in all kinds of ways!" The other is "Let's *especially* look at the worst of the worst cases, because instead of trying to squint at subtle changes of brain function, the effects will be turned up to 11! They'll be blindingly obvious!"

> I'm not sure what direction all of this happens in.

I think it's gotta be a vicious cycle, otherwise it wouldn't persist, right? OK how about this: "Globally weaker predictions cause sadness, and sadness causes globally weaker predictions".

I already talked about the first part. But why might sadness cause globally weaker predictions? Well, one evolutionary goal of sadness is to make us less attached to our current long-term plans, since those plans apparently aren't working out for us! (Remember the sadness circuit design goal I wrote above.) Globally weaker predictions would do that, right? As you weaken the prediction, "active plans" turn into "possible plans", then into "vague musings"...

Anyway, maybe that vicious cycle dynamic is always present to some extent, but other processes push in other directions and keep our emotions stable. ...Until a biochemical insult—or an unusually prolonged bout of "normal" sadness (e.g. from grief)—tips the system out of balance, and we get sucked into that vortex of mutually reinforcing "sadness + globally weak predictions".

[-] **G Gordon Worley III** 1mo 3

> Why isn't it symmetric? None of the theories are making high-precision predictions of failure either, right? Well, I think that generically when people are planning out how to do something, they search through their hypothesis space for a hypothesis that has a specific indicator of success; they don't search for a hypothesis that *lacks* a specific indicator of failure. I just don't think you can run the neural hypothesis search algorithm in that opposite way.[2]°

Interestingly, sometimes when people are in this state of thinking "I can't see how I will succeed, so I'll fail", you can employ the symmetric case to get them to act. That is, you might say to them something like "sure, but do you have any strong reason to think you will fail?" and if they don't come up with much more than vague musings you can push them to see that they were accidentally risk neutral all along, and then they'll act.

[-] **noggin-scratcher** 1mo 3

A thought occurs for sleep: normally to successfully fall asleep it helps to put yourself in a very predictable and boring state (lying motionless in the dark with your eyes closed), with no major sense input to attend to, threats to deal with, or necessary tasks to complete.

If everything is low certainty, maybe that leaves the brain unable to settle, because it can't reach high confidence in the hypothesis that it's safe to fall asleep.

[-] **Viliam** 1mo 2

Just a random thought: This could also explain why rationality and depression seem to often go together. Rational people are more likely to notice things that could go wrong, uncertainty, planning fallacy, etc. -- but in this model those are mostly things that assign lower probability to success.

Even in the usual debates about "whether rationality is useful", the usual conclusion is that rationality won't make you win a lottery (not even the startup lottery), but mostly helps you to avoid all kinds of crazy stuff that people sometimes do. Which from some perspective sounds good (imagine seeing a long list of various risks with their base rates, and then someone telling you "this pill will reduce the probability of each of them to 10% of the original value or less"), but is also quite disappointing from the perspective of wanting strong positive outcomes ("will rationality make me a Hollywood superstar?" "no"; "a billionaire, then?" "it may slightly increase your chance, but looking at absolute values, no"; "and what about ...?" "just stop, for anything other than slightly above average version of ordinary life, the answer is no"). Meanwhile, irrationality tells you to follow your passion, because if you think positively, success is 100% guaranteed, and shouldn't take more than a year or two.

[-] **Pattern** 1mo 1

> In depression, none of the hypotheses will make a high-precision prediction that X will

> occur,

So if 'uncertainty' causes 'depression', what does 'being certain you will fail' cause?

[-] **steve2152** 1mo 1

See the Low Self-Confidence section for why I think a depressed person can feel certain they'll fail despite an inability to create hypotheses with strong predictions.

Note that a "feeling / judgment of certainty about something in the everyday sense" is not exactly the same as the "certainty of a particular top-down prediction of a particular hypothesis (generative model) in the predictive coding brain". The former also involves things like the availability heuristic, process of elimination (I'll fail because I can't imagine succeeding), etc.

Does that answer your question? Do you think that makes sense?

[-] **Pattern** 1mo 1

The predictive coding cartoon at the top of the post has the section "Predicts that my innate drives will be satisfied". Intuitively none of the hypothesis/plans scoring highly on that metric could cause some form of 'unhappiness'.

[-] **steve2152** 1mo 1

I basically agree with that. See the section "Feelings of sadness, worthlessness, self-hatred, etc."