Implicit extortion

Paul Christiano Follow
Apr 13, 2018 · 8 min read

In this post I describe a pattern of behavior I call "implicit extortion." RL agents are particularly susceptible to implicit extortion, in a way that is likely to be problematic for high-stakes applications in open-ended strategic environments.

I expect that many people have made this point before. My goal is just to highlight the issue and to explore it a little bit more carefully.

Basic setup

Consider two actors, the target (T) and manipulator (M), such that:

- M wants T to perform some *target action*—e.g. make a payment, leak information, buy a particular product, handicap itself…
- M can take *destructive actions* that hurts both M and T—e.g. spreading rumors about T, undercutting T in a marketplace, physically attacking T…

In *explicit extortion*, M threatens to take the destructive action unless T performs the target action. Then a naive T reasons: "if I don't take the target action, something bad will happen, so I better take the target action."

In *implicit extortion*, M simply performs the destructive action whenever T doesn't perform the target action. Then a naive T eventually learns that failure to take the target action is associated with something bad happening, and so learns to take the target action.

Implicit extortion is very similar to explicit extortion:

- T would prefer not be the kind of person who is vulnerable to extortion, so that bad things don't happen to them.
- Extortion doesn't necessarily cost M very much, if they don't follow through on the threat very often.

However, implicit extortion can be particularly hard to avoid:

- It can be effective without T realizing that it's happening, which makes it hard for them to respond appropriately even if they do have defenses.
- It affects simple RL algorithms (which don't have defenses against extortion, and can't be easily modified to include such defenses).

Example

The most extreme and blatant example would be for M to send T a daily request for $100. On any day when T fails to pay, M launches a costly cyberattack against T. A human would immediately recognize this behavior as extortion and would respond appropriately, but an RL algorithm might simply notice that paying is the best strategy and therefore decide to pay.

Implicit extortion can be much harder to detect, while still being effective. Suppose that every time T tries to change their product, M runs a grassroots smear campaign. It might not be possible for T to distinguish the situations "M is attempting to manipulate me into not changing my product" and "Every time I change the product people get really unhappy, so I should do so sparingly."

Details

How expensive is this for the manipulator?

Suppose that T is using an RL algorithm, and M is trying to manipulate them. How expensive is this for M? How likely is it to be worthwhile?

At equilibrium: T learns to always perform the target action; so only fails to take the target action while exploring. The long-term cost to M depends entirely on the target's exploration policy.

If T uses ε-exploration, then they take the target action $(1 - \varepsilon)$ of the time. So M only needs to pay the cost of the destructive action on an ε fraction of trials.

For complex high-level actions, the effective ε can't be *too* high—it's not a good idea to "try something crazy" 10% of the time just to see what happens. But let's be conservative and suppose that $\varepsilon=0.1$ anyway.

Suppose that M is trying to directly extract money from T, $10 at a time, and that it costs M $50 of value in order to cause $15 of trouble for T.

If M asks for $10 on 10 occasions, T will refuse to pay only once as an exploration. Then M needs to pay that $50 cost only once, thereby ensuring that the cost of paying (=$10) is smaller than the average cost of refusing to pay (=$15). Meanwhile, M makes $90, pocketing $40 of profit.

In general, M can make a profit whenever the product of (payment efficiency) * (destructive efficiency) > ε, where "payment efficiency" is the benefit to M divided by the cost to T of the target action, and "destructive efficiency" is the cost to T divided by the cost to M of the destructive action.

In practice I think it's not too uncommon for payment efficiency to be ~1, and for destructive efficiency to be >1, such that extortion is possible regardless of ε. Small values of ε make extortion considerably easier and more cost-effective, and make it much harder to prevent.

During learning: the analysis above only applies when the agent has already learned to consistently take the target action. Earlier in learning, the target action may only occur rarely and so punishment may be very expensive. This could be worth it over the long term but may be a major hurdle.

Fortunately for M, they can simply start by rewarding the target behavior, and then gradually shift to punishment once the target behavior is common. From the perspective of the RL agent, the benefit of the target action is the same whether it's getting a reward or avoiding a punishment.

In the cash payment example, M could start by paying T $20 every time that T sends $10. Once T notices that paying works well, M can gradually reduce the payment towards $10 (but leaving a profit so that the behavior becomes more and more entrenched). Once T is consistently paying, M can start scaling up the cost of not paying while it gradually reduces the benefits of paying.

Analyzing the error

Paying off a (committed) extortionist typically has the best consequences and so is recommended by causal decision theory, but

having the policy of paying off extortionists is a bad mistake.

Even if our decision theory would avoid caving in to extortion, it can probably only avoid implicit extortion if it recognizes it. For example, UDT typically avoids extortion because of the logical link from "I cave to extortion" → "I get extorted." There is a similar logical link from "I cave to implicit extortion" → "I get implicitly extorted." But if we aren't aware that an empirical correlation is due to implicit extortion, we won't recognize this link and so it can't inform our decision.

In practice the target is only in trouble if would-be manipulators know that they are inclined to comply with extortion. If manipulators base that judgment on past behavior, then taking actions that "look like what someone vulnerable to extortion would do" is itself a bad decision that even a causal decision theorist would avoid. Unfortunately, it's basically impossible for an RL algorithm to learn to avoid this, because the negative consequences only appear over a very long timescale. In fact, the timescale for the negative consequences is longer than the timescale over which the RL agent adjusts its policy— which is too long for a traditional RL system to possibly do the credit assignment.

Other learning systems

What algorithms are vulnerable?

At first glance the problem may seem distinctive to policy gradient RL algorithms, where we take actions randomly and then reinforce whatever actions are associated with a high reward.

But the same problem afflicts any kind of RL. For example, a model-based agent would simply learn the model "not doing what the manipulator wants causes <bad thing X> to happen," and using that model for planning would have exactly the same effect as using policy gradients.

More broadly, the problem is with the algorithm: "learn an opaque causal model and use it to inform decisions." That's an incredibly general algorithm. If you aren't willing to use that algorithm, then you are at a significant competitive disadvantage, since the world contains lots of complicated causal processes that we can learn about by experiment but can't model explicitly. So it seems like everyone just has to live with the risk of implicit extortion.

I describe the problem as afflicting "algorithms," but it can also afflict humans or organizations. For example, any organization that is compelled by arguments like "X has always worked out poorly in the past, even though we're not quite sure why, so let's stop doing it" is potentially vulnerable to implicit extortion.

What about human learning?

Humans have heuristics like vindictiveness that help prevent us from being manipulated by extortion, and which seem particularly effective against implicit extortion. Modern humans are also capable of doing explicit reasoning to recognize the costs of giving in to extortion.

Of course, we can only be robust to implicit extortion when we recognize it is occurring. Humans do have some general heuristics of caution when acting on the basis of opaque empirical correlations, or in situations where they feel they might be manipulable. However, it still seems pretty clear that human learning is vulnerable to implicit extortion in practice. (Imagine a social network which subtly punishes users, e.g. by modulating social feedback, for failing to visit the site regularly.)

Evolution?

Evolution itself doesn't have any check against extortion, and it operates entirely by empirical correlations, so why isn't it exploited in this way?

Manipulating evolution requires the manipulator to have a time horizon that is many times the generation length of the target. There aren't many agents with long enough time horizons, or sophisticated enough behavior, to exploit the evolutionary learning dynamic (and in particular, evolution can't easily learn to exploit it).

When we do have such a large gap in time horizons and sophistication—for example, when humans square off against bacteria with very rapid evolution—we do start to see implicit extortion.

For example, when a population of bacteria develop resistance to antibiotic A, we take extra pains to totally eradicate them with antibiotic B, even though we could not afford to use that strategy if A-resistance spread more broadly through the bacteria population. This is effectively implicit extortion to prevent bacteria from developing A-resistance. It would continue to be worthwhile for humanity even if

the side effects of antibiotic B were much worse than the infection itself, though we probably wouldn't do it in that case since it's a hard coordination problem (and there are lots of other complications).

Conclusion

There are many ways that an AI can fail to do the right thing. Implicit extortion is a simple one that is pretty likely to come up in practice, and which may seriously affect the applicability of RL in some contexts.

I don't think there is any "silver bullet" or simple decision-theoretic remedy to implicit extortion, we just need to think about the details of the real world, who might manipulate us in what ways, what their incentives and leverage are, and how to manage the risk on a case-by-case basis.

I think we need to define "alignment" narrowly enough that it is consistent with implicit extortion, just like we define alignment narrowly enough that it's consistent with losing at chess. I've found understanding implicit extortion helpful for alignment because it's one of many conditions under which an aligned agent may end up effectively optimizing for the "wrong" preferences, and I'd like to understand those cases in order to understand what we are actually trying to do with alignment.

I don't believe implicit extortion is an existential risk. It's just another kind of conflict between agents, that will divert resources from other problems but should "wash out in the long run." In particular, every agent can engage in implicit extortion and so it doesn't seem to shift the relative balance of influence amongst competing agents. (Unlike alignment problems, which shift influence from human values to whatever values unaligned AI systems end up pursuing.)