Against In Defense of Unreliability

hivewired.wordpress.com/2018/07/02/against-in-defense-of-unreliability/

Saede Riordan July 2, 2018

Epistemic Status: Game Theory. Consequences.
Content Warning: Game Theory. Consequences.

So back in June of last year Zvi wrote about Duncan's Dragon Army Barracks. That post is long, goes over a ton of stuff, and is responding to Duncan's planning post about Dragon Army. I don't live in the Bay Area, and the trials and tribulations of a particular group house is really only of interest to me as a policy wonk, and I don't have a lot of interest in involving myself in the actual group house politics of the Bay. However, Zvi's post spawned another post by Ozy, which is what we'll be responding to here today.

Let's start with the original bit of Zvi's post that Ozy is responding to.

> I also strongly endorse that the *default* level of reliability needs to be *much, much higher* than the standard default level of reliability, especially in The Bay. Things there are really bad.
>
> When I make a plan with a friend in The Bay, I *never* assume the plan will actually happen. There is actual no onethere I feel I can count on to be on time and not flake. I would come to visit more often if plans could actually be made. Instead, suggestions can be made, and half the time things go more or less the way you planned them. This is a terrible, very bad, no good equilibrium. Are there people I want to see badly enough to put up with a 50% reliability rate? Yes, but there are not many, and I get much less than half the utility out of those friendships than I would otherwise get.

So Zvi says that "actually, reliability is kind of important" and Ozy extracts that bit and writes a post saying "no, actually, unreliability is important too"

> First of all, I'd like to say that nothing in my post should be construed as saying Zvi's desire for reliable friends is invalid or wrong. It's disappointing to expect a friend to come over and then they don't. If you're a busy person, on vacation or otherwise limited in time, a friend's canceled plans may mean that you've missed out on an important opportunity to do something productive and/or fun. It is very reasonable to want to befriend people who will reliably show up places they said they will on time. However, I do want to explain why I myself am quite unreliable and how I benefit from a social norm in which this unreliability is acceptable.

So here *I* want to say first that nothing in *my* post should be construed as saying that *Ozy's* desire for permission and allowance for being unreliable and flakey is invalid or wrong. There's nothing wrong with giving affordance to specific people around this issue on a case by case basis.

However, Ozy goes one step further, in that they're *advocating for a norm,* and well… that's where I have to step in and say something.

The next part of Ozy's post is…well I could uncharitably call it *excuses* but I'll be a bit more charitable and call it *contingent circumstances*.

It's kind of long, so I've baked Ozy's premise down as much as I can.

- Ozy's time is valuable too, If it's bad for someone to show up ten minutes late because the person is waiting around being bored, then it is also bad for Ozy to show up ten minutes early so *they* have to wait around and be bored.
- In many cases, showing up early is just as inconvenient for others as showing up late.

Now, I personally don't consider "shows up at some point within a 30-minute window" to be too big of a deal for some things, but the importance of timeliness quickly rises when you factor in other things, which Ozy doesn't and so I will, with some general heuristics on how important it is to be on time to particular things with respect to benefits and tradeoffs.

- Sure, your time has value, but if you are at an event with 30 people, and they are counting on you, then your time is only worth 1/30th the value of the time of everyone forced to wait for you (assuming that everyone's time is equally valuable, Ozy may be valuing their own time more highly than they value other people's and I can't say I entirely blame them, it's useful in our society to be at least somewhat biased in favor of yourself, because if you're not, chances are no one will be.) Even with that bias though, the more people attending a Thing, and the more the timing matters on your part.
    > Corollary: this applies in a matrix to how valued your time is vs the time of those around you. You may be comfortable showing up to a show a few minutes late and slipping into the back only disturbing a few people a little as a patron, but as an actor, you really need to be actually on time to the performance and the time value matrix looks very different.
- Punctuality signals that you respect the time of your friend as much as you respect your own time.
- Some things are *actually* on inflexible schedules (Ozy notes doctor's appointments but also anything to do with children is generally fairly inflexible, and of course most work environments take a dim view of casually wandering in and out.) Generally, if there is a waiting room associated with the activity, it's because it's expected that you are there early.
- Showing up a bit late vs a bit early depends on the event. You want to show up to things with fixed start times, such that you actually hit that fixed start time even if it means being bored for a few minutes beforehand. I would classify things like movies and plays, talks, Rationality Reading Group, doctor's appointments, court dates, and interactions with anyone whose time you value to fall into this category.

- Conversely, a lot of things have some flex to them, and in many of the cases where showing up early is worse than showing up late, this flex is present. Parties tend to phase slowly into existence, such that showing up "fashionably late" has become an idiom and actually becomes a status play. You're demonstrating that your time is valuable by waiting to show up until you know the party will be *really* in-swing.

However, as Ozy points out,

> Zvi didn't just talk about being on time: he also talked about flaking. My local corner of the Bay seems to have less of a flaking problem than his corner. I, a diagnosed agoraphobe, still manage to make the **majority** of the social events I agree to go to, and many people of my acquaintance make as much as **ninety or ninety-five percent.**

Bolding mine. This seems reasonable. This should be your goal, if you can't make a commitment, then don't plan for it and tell other people to plan for it. However, Ozy is advocating for, as they say,

> a social norm in which this unreliability is acceptable.

And *this* is kind of unacceptable. I want to avoid coming off as attacking Ozy too harshly, and I'm fine with certain people being unreliable and having a reputation as such, but I want to strongly denounce this sort of thing as a norm. We can make exceptions for individual people with contingent circumstances, but the *norm* should be to *honor your fucking commitments* or suffer the social repercussions thereof, and *yes,* this is *actually kind of important*.

Here's some more of Ozy,

> This means that when I make social arrangements a lot of the time I won't end up actually going to them because I will be too scared of leaving my house. Whether I'm going to have a good mental health day or a bad mental health day is hard to predict even a week in advance, because it depends on short-term triggers like whether I've fought with a close friend, whether the assholes across the street have decided to set off fireworks, whether a person has said something unpleasant about me on the Internet, whether I've been doing a good job of remembering that in spite of what my brain tells me doing things will make me feel better and not doing things will make me feel worse, and so on. **So the only way I can achieve any sort of reliability in social arrangements is by not making them.**

Bolding mine, I sympathize with Ozy's particular case, but wonder if it wouldn't be possible for them to at the very least provide some sort of probabilistic estimate? The issue is that Ozy is essentially saying they have no control or ability to negotiate with their future self. This makes them relatively useless as far as being an ally under game theory. If you can't count on them to keep their word then you have to assume everything they say is faulty, and that they don't value their own word very highly.

I know, let's make a chart, everyone loves charts.

	I Predict: Reliability	I Predict: Unreliability
You Behave: Reliably	Reliably Reliable	Reliably Unreliable
You Behave: Unreliably	Unreliably Reliable	Unreliably Unreliable

There's actually a third dimension of this chart where you replace the leftmost column with "your predictions." So the true axes are "your prediction" vs "my prediction" vs "your behavior."

So we have a bunch of different wants here

- I want to accurately predict what future you will do
- You want to accurately predict what future you will do.
- You want to portray yourself in a positive light
- You want to not unduly impose upon your or my future selves
- I want to not unduly impose on your or my future selves

So for my own part I can:

1. Predict you will behave reliably and have you behave reliably. This is being "reliably reliable"
2. Predict you will behave reliably, only for you to behave unreliably, this is a danger zone. This is being "Unreliably Reliable"
3. Predict you will behave unreliably, only for you to behave more reliably than I think you will, and be pleasantly surprised, this is being "Reliably Unreliable"
4. Predict you will behave unreliably, and you behave unreliably, this is being "unreliably unreliable"

I really want to avoid quadrant two. That's basically letting myself be betrayed, so if I think you're going to behave unreliably than I *really* want to predict that you will behave unreliably. In this sense, Ozy is doing us a huge favor in getting out ahead of things and warning us that they're not that reliable in advance so we can plan accordingly, which is huge and allows us to avoid the biggest danger zone and makes them a much safer person for that alone. They don't lie about their reliability to upsell their reputation.

But people who fall into categories 3 and 4 are useless outside a very shallow class of friendships, and I think that this is largely Ozy's privilege and bias as a relatively well-off programmer-type showing.

> a friend's canceled plans may mean that you've missed out on an important opportunity to do something productive and/or fun.

As it turns out, productive and fun are not the only vectors of friendship for a lot of people. What Ozy's talking about is the sort of milquetoast, atomized, easily dropped friendships that are endemic to Western American culture. But that sort of friendship,

in addition to just being flat out kind of shitty as a mode of interaction from my point of view, is also the sort of relationship that is a privilege of people who are high in slack and can afford to make friends with people who are game-theoretically useless.

There's no pressure on Ozy to optimize their friendships for people who will be useful to them because their critical needs are already met. Shikashi, if you're not a well-to-do bay area programmer, then you might want to *actually* be able to rely on your friends in a pinch, and if you predict someone will go "sorry, I understand that your car broke down and that you're going to get fired and lose your job and then go homeless, but I'm having a bad mental health day so I'm not going to help you" you're probably not going to want to invest as much time and energy into the friendship, compared to someone who would willingly come pick you up on the side of the road at 7:30 am.

Ozy says it themselves.

> I do not want to not make social arrangements. Social isolation makes my mental health worse. And doing *literally anything* tends to make me less depressed. I am also informed that some people would occasionally like to talk to me [citation needed]. So therefore I have decided to make plans *anyway*, **and push onto my friends the negative consequences of dealing with my flakiness.**

Bolding mine. This is only a viable course of action if your friends have the slack to absorb this, and the consequence is that you're selecting against anyone in a low slack situation, with constraints on their time or resources. It's basically reinforcing the Personal Filter Bubble effect. Your flakiness means that you're selecting against vulnerable people and anyone who *needs* to rely on their social network for survival, basically ensuring the only people you talk to are other well-to-do programmers who can afford to eat the negative consequences of your flakiness without being crippled by it. Vulnerable people that can't afford the negative consequences will be forced to choose other friends, even if they really like you as a person because the risk that you just vanish when they need you the most is too great to invest energy into the friendship.

Friendship can't be unrequited unless you have the slack to not need those friends. You wouldn't go a doctor who might decide to not show up to your appointments because they were having a bad mental health day or a firefighter who will only save your house if they're in a good mood, and by using this model of "my friends don't need to actually count on me for anything" Ozy is defining friendship as a sort of positive relationship with people they don't need or expect anything from. Fair-weather friends, basically.

Once again, it's fine if a few people are like this, and especially if they signal "I'm like this, please don't hold it against me" in advance like Ozy does, I'm pretty willing to accept it, but again, the *problem* is proposing that this should be a *norm*, and as a *norm,* it's terrible and has the side effect of driving vulnerable people out of our communities.

It's saying "if you don't have the slack to invest time and energy into me with no expectation of anything in return then you shouldn't try to be friends with me," and as a norm, this basically means that poors need not apply and that the rationalist community is only for people who don't need anything from the communities they participate in. I don't know if this is a dominant attitude, but if it is, it really explains why the Less Wrong community seems to be so biased towards people in a relatively secure life position (white, cisgendered, upper class, straight, males in particular).

If we want to have a community that is safe for people in more vulnerable life circumstances, then we need to be willing to actually do things for each other and support our friends, and not just pay lip service to diversity while creating a community environment that is cruel and unsafe to marginalized individuals. If you need reliable friends who you can call on in a pinch in order to avoid homelessness or ruin than Ozy's post is essentially telling you to fuck off, and that the rationalist community is for people who only need fairweather friends. I don't think Ozy *means* to say this, but the subtext is there nonetheless.

Their entire post is honestly written from a place of tremendous privilege, such that it was rather cringe-inducing for me to read. I'm willing to give affordance to Ozy as a person because I agree with their assessment that their time is valuable and they've contributed a lot to the community over the years, however, as a norm, it's classist as fuck, and I'd really rather we didn't.

But there's a whole other dimension of this outside of the interactions between me and you and my model of future you. There's also *your* model of future you, and your ability to negotiate with that person, trade with them, do things for them, and extract commitments from them.

There's a whole chunk of Zvi's post that is talking about *why* reliability and meeting commitments reliably are so important which Ozy never provides a counterargument to. Not for other people, but as a tool to be able to actually negotiate with your future self.

> When you are learning to play the piano, you are effectively deciding each day whether to stick with it or to quit, and you only learn to play the piano if you never decide to quit (you can obviously miss a day and recover, but I think the toy model gives the key insights and is good enough). You can reliably predict that there will be variation (some random, some predictable) in your motivation from day to day and week to week, and over longer time frames, so if you give yourself a veto every day (or every week) then by default you will quit far too often.
>
> If every few years, you hold a vote on whether to leave the European Union and destroy your economy, or to end your democracy and appoint a dictator, eventually the answer will be yes. It will not be the 'will of the people' so much as the 'whim of the people' and you want protection against that. The one-person case is no different.
>
> The ejector seat is important. If things are going sufficiently badly, there needs to be a way out, because the alternatives are to either stick with the thing, or to eject anyway and destroy your ability to commit to future things. Even when you eject for good reasons using the agreed upon procedures, it still damages your ability to commit. The key is to calibrate the threshold for the seat, in terms of requirements and costs, such that it being used implies that the decision to eject was over-determined, but with a bar no higher than is necessary for that to be true.
>
> For most commitments, your ability to commit to things is far more valuable than anything else at stake. Even when the other stakes are big, that also means the commitment stakes are also big. This means that once you commit, you should follow through almost all the time even when you realize that agreeing to commit was a mistake. That in turn means one should think *very carefully* about when to commit to things, and not committing if you think you are likely to quit in a way that is damaging to your commitment abilities.

When you make a commitment and then fail to uphold it, you're essentially betraying your past self, and telling the "Past you" that "future you" is an untrustworthy bargaining partner. Every time you fail to uphold a commitment, it gets easier to perform that betrayal and the less the commitment means anything. In the worst cases of this, your sense of self completely truncates down into the moment, and you can't bargain with your past or future self at all. Someone in this state is going to feel unable to commit to anything because they know their future self might betray that commitment, and so the magical power the commitment holds evaporates, and you lose the ability to force the commitment through aversive conditions.

> I think that if anything, Duncan under-states the importance of reliable commitment. His statements above about marriage are a good example of that, even despite the corrective words he writes about the subject later on. Agreeing to stay together for a year is a sea change from no commitment at all, and there are some big benefits to the year, *but that is not remotely like the benefits of a real marriage.* Giving an agreement an end point, at which the parties will re-negotiate, fundamentally changes the nature of the relationship. Longer term plans and trades, which are extremely valuable, cannot be made without worrying about incentive compatibility, and both sides have to worry about their future negotiating positions and market value. Even if both parties want things to continue, each year both parties have to worry about their negotiating position, and plan for their future negotiating positions.
>
> You get to move from a world in which you need to play both for the team and for yourself, and where you get to play only for the team. This changes everything.
>
> It also means that you do not get the insurance benefits. This isn't formal, pay-you-money insurance. This is the insurance of having someone there for you even when you have gone sick or insane or depressed, or other similar thing, and you have nothing to offer them, and they will be there for you anyway. We need that. We need to *count* on that.
>
> I could say a lot more, but it would be beyond scope.

Saying more might have been beyond the scope of Zvi's post, but it's what this post is *about*.

Humans *need* support networks that they can count on. No man is an island, everyone needs people they can lean on in times of crisis. If in crisis, all our friends evaporate like the morning dew and leave us alone in our struggles, then we're much worse off. These fair-weather friends are in a sense social leeches because they're stealing time and energy from you that they won't pay back. Like the insurance company stiffing you after you've gotten into an accident.

It might feel slightly gross to view friendship in these sorts of terms, surely "My friends value me for me! As a person! I don't want to only invest energy into a friendship when I think it'll give me a payout!" and that's *good* in a sense, but it only holds if the majority of the people in a social space share that norm and support each other and are actually there for each other regardless of their anticipated payout, which can only happen when people behave in a reliable and predictable manner and do the things they say they've committed to doing.

I'm going to repeat a chunk of what Zvi said above because it's *really* important and I want to sort of hammer it in if I can.

> Longer term plans and trades, which are extremely valuable, cannot be made without worrying about incentive compatibility, and both sides have to worry about their future negotiating positions and market value. Even if both parties want things to continue, each year both parties have to worry about their negotiating position, and plan for their future negotiating positions.
>
> You get to move from a world in which you need to play both for the team and for yourself, and where you get to play only for the team. **This changes everything.**

Bolding mine. True commitment and reliability change the nature of the game. They make a competitive game cooperative, they allow you to focus on what you can contribute to the relationship, instead of looking at what you can extract from it, and you can't move from extractive to cooperative if there's a set point in the future when the relationship is going to terminate, or if there's no real honor or strength to the commitments associated with that relationship.

I think that people in this community are far too willing in general to defect, both on their past selves, and on their relationships. Everything I hear about the bay is that it's become something of a trainwreck of late, and this seems like a major component in why that's happening, and why the rationalist community, in general, is not as effective at getting things done as we could be. It's too easy to defect with relatively minor repercussions, and so there's this constant jockeying for position and infighting in the community. From where I'm standing the *last* thing we need right now is a *defense* of unreliability.

What we need are actually reliable people who actually honor their commitments. Don't say you're going to do something if you don't think you're going to do it, or I will update towards you not valuing your own truthfulness or my time very highly, contingent circumstances notwithstanding, and will be less likely to want you as a friend. I want my friend group to be made of people who I know would help me in an emergency, and yes this *does* mean that *I* need to be able to be there for *them* in similar emergencies to the best of my ability, even if I *am* having a bad mental health day.