Privacy Practices

"Can you keep this confidential? How do you know?"

Pet peeve about privacy: I think people are woefully inadequate at asking, and answering, "Can you keep this confidential?"

Disclosure: I am not inherently great at keeping information private. By default, if a topic came up in conversation, I would accidentally sometimes say my thoughts before I had time to realize "oh, right, this was private information I shouldn't share."

I've worked over the past few years to become better at this – I've learned several specific skills and habits that make it easier. But I didn't learn those skills in school, and no one even really suggested I was supposed to learn them. People seemed to just assume "people can keep secrets, and it's low cost for them to do so."

And... maybe this is just me. But, people say to me "hey, can you keep this private?", in a tone that implies I'm not really supposed to say no. And that's the *best* case. I've also observed things like...

...people saying "hey, this is confidential", and then just saying the thing without checking in.

...people saying "sign this NDA", without really checking I have the skills to honor that agreement, and if I were to *not* sign, I'd... probably get fired? Unclear.

...people gathering for a Circle or other private safe space, and saying (best case) "do we all agree to keep things here confidential? Raise you hand?" and worst case, just flatly asserting "This is a safe space, things are confidential here". (And I have seen at least one instance where someone I actively trusted later betrayed that trust)

...people saying "You can report things to our [org / HR department / point-person], and they will keep things confidential." But, I know that in the hiring process for that org or department, no one ever checked that people actually had privacy skills.

And meanwhile, I have almost never heard anyone say something like "I have been given 10 bits of private-info over the past few years, and I accidentally leaked two of them", or even "I have paid any attention at all to how leaky I am with regards to confidential information."

What is a secret, even?

Meanwhile, people seem to vary in what they even mean by "secret" or "private information." Some people take them as serious oaths, some people just kinda sorta try to keep the R0 of the info lower than 1. Sometimes it seems to mean "carry this information to your grave", and sometimes it means "I dunno keep this on the down-low for awhile until the current controversy blows over."

Some people reading this might be surprised this is even a big deal. I gave a lightning-talk version of this blogpost last weekend, and one person asked "does this really matter that much, outside of major company NDAs or state-secrets?" Another person expressed similar skepticism.

I think it varies. The problem is exactly that *most* of the time, secrets aren't that big of a deal. But people don't seem to take time to get on the same page of exactly how big a deal they are, which is a recipe for mismatched expectations.

It's a bigger deal for me, because I live in social and professional circles adjacent to EA Grantmaking where line between the personal and professional is (perhaps unfortunately) a bit blurry. Sometimes, I talk to people exploring ideas that are legit infohazardous. Sometimes, people are hesitant to talk because they're worried it may affect their career.

It's also important to me from a Robust Agency standpoint – I'd like to be a reliable agent that people can coordinate with in complicated domains. Many other people in the x-risk ecosystem also seem interested in that. I think "the ability to exchange information, or reliably not exchange it" is a key skill, and worth cultivating because it enables higher order strategies.

What to do with all this?

I don't have a clear next action with all this. Right now, there's a vague social norm that you're supposed to be able to keep secrets, and that certain types of information tend to be private-by-default, but outside of things like "your social security number", there's not much agreement on *what.*

What I've personally taken to doing is giving myself a TAP, where as soon as I notice that a conversation or relationship is moving in the direction where someone might want to give me private information (or vice versa), I say "hey, I'd like to have a little meta-discussion about privacy."

And then we have a chat. If the conversation literally *just* broached the idea that one of us share private info, I try to avoid face-to-face contact to avoid micro-expressions revealing information. (Someone else recently suggested leaving more pauses in the conversation, so that reaction-time didn't reveal information either).

Then, I ask some questions like:

Can you keep a secret?

How do you know?

What exactly do you mean by secret?

Meanwhile, acknowledging: "Hey, so, in the past few years I've leaked at least one important bit of private-info. I haven't kept track of how much private info I *didn't* leak. But, I've also been working on gaining skills that make me more reliable at keeping things private, and making it lower cost for myself to take on confidential information. I'm fairly confident I can keep things private if I have to, but it's still a moderate cost to myself and I have to choose to do it on purpose. So please don't assume I'm keeping anything private unless I've specifically told you so."

I think it'd be good if such meta-conversations became more common.

I think they most importantly should be common if you are *creating an organization* that relies a lot on confidentiality. If you're promising to your clients that their information is private, but you aren't actually checking that your employees can keep

confidence, you're creating integrity debt for yourself. You will need to pay it down sooner or later.

This is (hopefully) the first post in the Privacy Practices sequences. The next post will (probably) be "Parameters of Privacy."

Parameters of Privacy

In my last post, I argued that people should probably have more meta-discussions about privacy (rather than simply assuming everyone was on the same page about how seriously to take confidentiality)

What might that conversation entail?

First, it's worth first checking **"Is this actually all that important? Do you want me to try very hard to keep this private?"** Much of the time, many people don't care that strongly. They just don't want you to go around blabbing publicly, and would prefer if you err on the side of not spreading it if you can.

Simply confirming that the stakes are low may be all that's needed, and it's good check that first to avoid spending unnecessary effort.

(As I said in the comments last time: the reason I think it's *useful* to check if the stakes are actually low, is that a) people sometimes have different expectations, b) sometimes, the ambiguity about "how seriously am I supposed to take privacy here?" can become ammunition in a power game, and I'd prefer to remove that ambiguity)

But if the stakes are moderate-to-high, you might talk through some parameters before revealing more information.

Note: I'm using "secret" and "private information" somewhat interchangeably here, because "secret" is a shorter noun that's easier to work into sentences. I think there are actually some distinctions between them, but those distinctions aren't the point of this essay.

Frames of Privacy: Ownership vs Caution

Ownership

One model of privacy is ownership based – I have some information, I'm considering sharing the information with you, but want to "retain ownership" over the information, such that you only use it in ways I endorse.

This could include "my social security number", or "my private feelings about a matter." It could also include *other* people's private information that I've been "loaned" (Alice shared her social security number or private feelings with me, and they belong to Alice, but in this circumstance I'm confident Alice would be fine with me sharing them with you so long as you agree to the same privacy terms that I did)

Caution

But, a different model here is about "proper caution." Say I'm a physicist who discovers how to build nuclear reactors. As a scientist, I might generally desire to share information and educate people. I don't are about "ownership" of the idea.

But nuclear reactors are dangerous, and I don't want it to fall into the wrong hands. If I share it with other people, I might want to check: will they misuse the information? Will they share it with other people who might misuse the information?

Sometimes the danger comes from incomplete information – Carla overhears Alice and Bob conducting an improve scene, where Bob is insulting Alice. If Carla were to tell someone "Bob insulted Alice" but not "Bob and Alice were in an improv scene where the insults were completely consensual", she'd be spreading misinformation that harms Bob.

If Carla is considering telling Dave about the information, she might care about is whether Dave will make sure that if he tells anyone else, he conveys the *full* story, not just "Bob insulted Alice."

In general this frame is less concerned about ownership, but about good judgment, which might be domain specific. (You trust Joe not to reveal secrets about nuclear reactors, but might not trust his judgment in sharing personal information that people might misinterpret)

Other Frames

There may be other frames for privacy. I think it's good to at least be aware that you and your colleague might be operating in different frames, and which come with different assumptions about what's important.

With that in mind, what are some specific parameters you might fine-tune for a given exchange of private-information?

Parameters

i. Am I making a promise?

Privacy is an important tool for coordination.

Another useful tool for coordination is the specific tech of "making a promise" – committing to definitely make sure to get something done (or not done). If I do not successfully do the thing, you are right to judge me, and trust me less in the future. Breaking a promise has longterm consequences.

I think it's quite important to be able to make promises, and to be able to rely on people who make them.

Consequently, I think it's important that our social norms *not* require people to casually make promises that they can't actually keep. Doing so erodes the tool of promise-making. And it fosters an environment where most people are guilty, but can be selectively punished.

So I think it's useful to more explicitly distinguish "private information that you're making a reasonably good faith effort to contain" and "private information you're making a promise to contain."

I generally don't think it makes sense to make promises by default.

ii. Who am I keeping this secret from, and to what degree?

One might want any of the following:

1. Never reveal *any* information that allows anyone to make updates about the secret, including microexpressions. (This is quite hard, and I don't think should generally be expected)
2. Don’t reveal more than “I can’t talk about that because of confidentiality"ere
3. Don’t tell anyone directly about the secret
4. Don’t spread the secret more than N degrees
5. Make sure the information doesn’t spread to a particular person.
6. Make sure the secret doesn’t reach people who might use it to hurt the ingroup.
7. You can talk about the secret, but not reveal the particulars.
8. Any of the above, but you can *have a confidant.*

I want to draw particular attention to that last point. One thing I've found fairly burdensome about privacy is not having someone who can help me think through the ramifications of a situation.

iii. What Skills Am I Expected to Have?

Depending on the previous question, you might need to have particular skills:

For not revealing information:

- Remembering not to tell the secret in the first place.
- Gracefully segueing a conversation so as not to reveal that you almost revealed something
- Thinking fast enough to respond to direct interrogation without revealing information. Or...
- ...removing yourself from the conversation (awkwardly if necessary).
- Control over your microexpressions.

Attention to context:

- Awareness of which people talk to each other (i.e. if tell Carla, is she likely to tell Bob?)
- Good judgment about the object-level content of the information. Under what sorts of circumstances might it be harmful for someone to know it, or spread it?
- Good judgment about which other people are okay to tell (including whether *they* have any of these skills)

Psychological safety:

- Some secrets are hard to bear alone, or even sometimes in a small group. Do I have the resilience to hold the secret without feeling isolated and stressed?
- If a secret is a literal infohazard that might harm the listener, do I actually have the skills to think through that infohazard safely? (this includes lots of subskills which depends on the infohazard)

Then, there's self-awareness about how good you are at each of these skills.

iv. Duration

How long do you need to keep the secret? Literally until the day you die? Until some current controversy has blown over, or some product launched?

Most of why I'm *averse* to keeping secrets has to do with the cognitive overhead of tracking multiple secrets that accumulate over time. Time-limited secrets avoid secret-creep.

v. Escape Clauses

There are some circumstances where I might end up regretting having made an all-encompassing promise. If a secret is important to someone, I try to talk through

Two clusters of reasons are:

Costs/Benefits in local situations

Sometimes a secret isn't that *that* big a deal, and meanwhile, a situation comes up where it's hard for me to have a conversation with Bob without inadvertently revealing some facts that relate to a secret Alice told me.

It'd be quite valuable to have the conversation openly with Bob, and meanwhile I'm pretty confident it wouldn't harm Alice or anyone else to tell Bob.

Now, this is the sort of judgment call that results in mismatched expectations and feelings of betrayal, and I'm *not* advocating that people unilaterally decide to share information whenever it feels convenient. But, I do think people underestimate the costs when agreeing to keep something private in the first place. If you were trying seriously to keep a secret, often that means keeping a lot of related details secret, and that ends up making it really hard to have what would otherwise be an innocuous conversation.

So, before agreeing to keep something private, I try to get a sense of how important it actually is to the person, and to talk through this consideration explicitly.

Patterns of Manipulation

I'll have a whole other blogpost about this. But quickly noting for now: one major issue with privacy is that it can be used to protect bad actors.

I've met a couple people who exploited my willingness-to-keep-things-confidential, which made it harder for me to share notes about them with other people. They generally pushed for confidentiality in a way that dampened an entire community's ability to notice that they were harming people.

I'm still figuring out exactly how to think about this. But, I know have a general escape clause in all privacy promises: If I come to believe that you are manipulating and harming people, I may reveal some things you told me in confidence (in as controlled and honorable a way I can think of). If you choose to tell me the secret, you're trusting that a) I have good judgment about that, and b) that the secret is not something I'm likely to perceive as part of a manipulative pattern.

Sensible Defaults

Negotiating all that individually each time is a pain, and you probably don't want to do it each time. Also, most people don't enjoy meta-discussion as much as I do, and you probably don't want to dump 1-3 blogposts worth of material on most people the first time private-information comes up.

Obviously people may vary in what defaults make sense. In a future post, I'll lay out my entire privacy policy more explicitly, but for now it seemed good to list as an example my most common defaults.

My defaults:

1. Try to notice when people are sharing information that is commonly-coded "private", and get the person to stop until we've chatted at least briefly about all this.
2. **Don't** make promises by default.
3. **Do** offer "low effort not-too-reliable pseudoprivacy" (clearly labeled as such)
4. **Do** offer "reasonable good faith effort to keep things private" (without promising) fairly easily to people who need it.
5. When higher degrees of privacy are required, **almost always have a confidant** who is able to offer an outside perspective on the situation.
6. If I take on semi-private information, and need to share it with others, try to have the others take on a higher degree of privacy than I did, to limit it's spread. (i.e it's often basically fine for some info to spread *a little*, Alice just didn't want it spreading across the whole internet)

Privacy and Manipulation

Previously:

- “Can you keep that confidential? How do you know?”
- Parameters of Privacy
- Norm Innovation and Theory of Mind

My parents taught me the norm of keeping my promises.

My vague societal culture taught me a norm of automatically treat certain types of information as private.

My vague rationalist culture taught me norms that include:

- noticing when I'm confused
- noticing when societal norms were inadequate to handle my situation
- being very honest, and having strong theoretical underpinnings for handling situations where I felt it was inadvisable to be honest.

Eliezer's post about meta-honesty was one of the most influential posts I've read in the past few years, and among the posts that inspired the coordination frontier. I was impressed that Eliezer looked at ethical edgecases, and wasn't content to make a reasonable judgment call and declare himself done.

He went on to think through the ramifications of various policies, devise a potential new norm/protocol, examine reasons that protocol might work or not work. He noted considerations like [paraphrased] "It matters that the norm be simple enough that people can reliably understand and use it." Or, quoted directly: "This norm is too subtle for Twitter. It might be too subtle for us, too."

From this post, I derived a (not-quite-spelled-out) norm of "when you try to navigate the edge cases of your norms, try thinking through the underlying principles. But don't try to be too clever, and consider the ways your edge-case-handling may fail to scale."

With this in mind, I want to walk through one of the ethical dilemmas I faced that I reflected on, when writing Norm Innovation and Theory of Mind. This is more of an object-level post, primarily a followup to my Privacy Practices sequence. But it seemed like a useful illustrative example for the Coordination Frontier concept.

Privacy norms can be wielded as an obfuscating weapon

Sometimes, privacy is wielded as a tool to enable manipulation.

I’ve run into a couple people who exploited my good faith / willingness to keep things confidential, as part of an overall manipulative pattern. Unfortunately, I don't feel comfortable going too far into the details here (please don't speculate in the comments), which makes it a bit harder to sanity check.

"Manipulation" is a tricky to define. It's a blurry line between "communicating normally" and "communicating in a way that systematically distorts another people's thinking and controls their behavior against their wishes". I'd like it say "it's hard to define but you know it when you see it", but often it's hard to see it because manipulation systematically tries not to be seen.

I've met some people seemed deliberately manipulative, and some people who might have been well intentioned, but in the end it didn't matter. They interacted with me (and others) in a way that felt increasingly uncomfortable, which seemed to be harming people. They skirted lines, wove narratives that made it feel awkward for me to criticize them or think clearly.

One of the key strategies they employed was to make it feel awkward to get help from other people to think or sanity check things. And one tactic in that strategy was pushing for confidentiality – sometimes explicitly, sometimes implicitly.

Explicit promises I regret

One person (call them Dave) asked for explicit promises of confidentiality on a number of occasions, sometimes after-the-fact. We once had a long conversation about their worldview and worries they had, which ended with me saying "so, wait, is this meant to be confidential?". They responded, somewhat alarmed-seeming, with "oh definitely I would never have told you all this if I thought you might share it."

At the time I found that somewhat annoying, but agreed to keep it confidential and didn't think much of it. (Nowadays try to notice when a conversation is veering into sensitive topics, and have a quick meta-conversation about confidentiality preferences in advance).

The most frustrating thing came under ideal-privacy-conditions: Dave asked me to make a specific promise of confidentiality before telling me something. I agreed. Then they told me some stories that included somebody harming themselves as a result of interaction with Dave.

Later on, a number of people turned out to be having bad interactions with Dave. Many of them had had similar conversations with him. Some of those conversations had included promises of confidentiality. Others had not. It gradually became clear that Dave was not being honest.

What became really frustrating was that a) it was actually important to figure out whether Dave was harmful, and it was much harder to do without sharing notes. b) more infuriatingly, most of the information had been given to some people *without* conditions of confidentiality, but it was still hard to talk about openly about without betraying promises.

I think it's important to take promises seriously. But in this case I think many of the promises had been a mistake. Part of this is because I think people should generally make fewer privacy promises in the first place.

At the time, I decided to reveal some bits of the information when it seemed really important, and acknowledging to myself that this made me a less trustworthy person, in some ways. This seemed worth it, because if I hadn't revealed the information, I'd be revealing myself to be untrustworthy in *other* ways – I'd be the sort of person who

was vulnerable to manipulative attacks. Integrity isn't just about being honest, it's about being functional and robust. Sometimes it involves hard tradeoffs in no-win scenarios.

It feels important to me that I internalize that hit to my integrity. That's part of why I'm writing this blogpost – it's sometimes necessary to violate your internal moral code (including keeping promises). But, when I do, I want people to know that I take it seriously. And I want people to have an accurate model of me.

But in this case, at least, the solution *going forwards* is pretty simple: I now try to avoid making such promises in the first place.

Instead, I include a clause saying "in rare circumstances, if I come to believe that this was a part of a manipulative pattern that is harming people, I may carefully share some of the information with other people." Most of the time this is fine, because most private-information is obviously not the sort of thing that's likely to be interpreted as part of a manipulative pattern (assuming you trust me to have remotely sane judgment).

There *are* some cases where you have information that you'd like to share with me, that I actually want to hear, which *is* the sort of thing that could be easily construed as manipulative and/or harmful, and which requires more trust than you currently trust my judgment. (Dave would have recognized this to be true about the conversation he was asking confidentiality about).

I am not sure what to do in those cases.

I think I would never commit to 100% reliable confidentiality. But, if the conversation seemed important, I'd first have a lengthy conversation about meta-honesty and meta-privacy. I might ask Alice for ~2 confidants that we both trust (from different parts of the social graph), who I might go to to get help evaluating whether Alice is manipulating me.

Implicit confidentiality and incriminating evidence

Another person (call them Carla) never extracted a promise of confidentiality from me. But they took advantage of a vague background norm. Normally, if someone comes to me talking about something troubling them, I try to keep it private by default. If someone is hurting and expresses vulnerability, I want them to feel safe talking through a problem with me (whether they're just venting, or trying to devise a solution).

Sometimes, this includes them talking about times they screwed up, or ways they harmed people. And in most cases (that I have experienced) it still seemed correct to keep that confidential by default – the harm was relatively minor. Meanwhile, there was value in helping someone with a "Am I the asshole?" kind of question.

But some of my talks with Carla veered into "man, this is actually a red flag that should have prompted me to a higher level of attention", where I should have considered not just how to help Carla, but how to investigate whether Carla was harming others and what to do about it.

In one notable case, the conversation broached a subject that might have been explicitly damning of Carla. I asked a clarifying question about it. She said something evasive, avoided answering the question. I let it slide.

If I had paid more attention, I think I could have updated on Carla not being trustworthy much sooner. (In fact, it was another few years before I made the update, and Carla is no longer welcome in my day-to-day life). I now believe Carla had a conscious strategy of revealing different parts of herself with different people, making people feel awkward for violating her trust, and using that to get away with harmful behavior in relatively-plain-sight.

I'm a bit unsure how harshly to judge myself. Noticing manipulation, evasion, and adversarial action is legitimately hard. My guess is that at the time, it was *a little* beyond my skillset to have noticed and taken appropriate action. It's not useful to judge yourself harshly for things you couldn't really have done better.

But it wasn't *unreasonably* beyond my skillset-at-the-time. And in any case, by this point, it *is* within my skillset. I hold myself now to the standard of paying attention if someone is skirting the line of confessing something harmful.

What do you do if someone *does* confess something harmful, though?

It's still generally good to have a norm of "people can come to each other expressing vulnerabilities." It's bad if Alice has to worry "but, if I express a vulnerability that Bob decides is *actually* bad, Bob will reveal it to other people and I will get hurt."

Most of the time, I think it is quite good for Alice to feel safe coming to me, even if I think she's being a bit of a jerk to someone. It's only in rare cases that I think it makes sense to get a sanity-check from someone else.

I don't think I ever owed Carla the security of a promise. But, it still matters whether people can generally expect to feel safe sharing vulnerable information with me.

Reneging on Confidentiality Pro-Socially

I don't have a great solution. But, here is my current algorithm for how to handle this class of situation:

First, put a lot of upfront effort into talking publicly about my privacy policies, so that they're already in the water and ideally, Alice already knows about them.

Second, notice *as soon as Alice starts sharing something vulnerable*, and say "hey, is this something you want me to keep confidential? If so I'd like to chat a little about how I do confidentiality." (What happens next depends on the exact situation. But at the very least I convey that I'm not promising confidentiality yet. And if that means Alice isn't comfortable sharing, she should stop and we should talk about it more at the meta level)

Third, if I think that Alice is manipulating me and I'm not sure what to do, get help from a confidant who promises a high degree of confidentiality. Share as little information as possible with them so that they can help me form my judgment about the situation. Try to get as much clarity as I can, while violating as little implicit or explicit expectations of privacy as possible.

Fourth, ????. Maybe I decide the red flag is actually just a yellow flag, and Alice is fine, and I continue to help Alice with their problem. If I believe Alice's behavior is manipulative and harmful, but that she's mostly acting in good faith, maybe I talk directly to her about it.

And, in (I hope rare?) cases, ask a couple other confidants, and if everyone seems to agree that Alice is acting adversarially, maybe start treating her as an adversary.

The Catch-All Escape Clause

Once upon a time, I didn't have a special clause in my privacy policy for "manipulative patterns". I made promises that didn't caveat that potentiality. I had the moral-unluck to have to deal with some situations without having thought them through in advance, and took a hit to my integrity because of that.

It seems quite plausible this will not be the last time I discover a vulnerability in my privacy practices, or my general commitment-making practices.

So, it currently seems like I should include more general escape clauses. If you are trusting me with something important, you are necessarily trusting my future judgment. I *can* promise that, if I need to renege on a promise, I will do so as pro-socially as I can. (i.e. try to internalize as much of the cost as I can, and try to adjust my overall policies to avoid having to renege further in the future)

Communities of Robust Agents

These are my current practices. I'm not confident they are best practices. I think this is a domain where it is particularly important that a social network has shared assumptions (or at least, common knowledge of divergent assumptions).

It matters whether a community is a safe space to vulnerably reveal challenges you're facing.

It matters whether a community can sanely discuss literal infohazards.

It also matters whether a community can notice when someone is using the guise of vulnerability or infohazards to obfuscate a pattern of harm, or power grab.

I aspire to be a robust agent, and I hope that my community can be a robust community. The social circles I run in are trying to do complicated things, for which there is no common wisdom.

They require norms that are intelligently designed, not culturally evolved. I want to have norms that are stable upon reflection, that are possible to talk about publicly, that people can inspect and agree "yes, these norms are the best tradeoffs we could make given the circumstances."