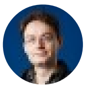


François Chollet Follow

Mar 29 · 17 min read

What worries me about AI

***Disclaimer*: These are my own personal views. I do not speak for my employer. If you quote this article, please have the honesty to present these ideas as what they are: personal, speculative opinions, to be judged on their own merits.**

If you were around in the 1980s and 1990s, you may remember the now-extinct phenomenon of "computerphobia". I have personally witnessed it a few times as late as the early 2000s—as personal computers were introduced into our lives, in our workplaces and homes, quite a few people would react with anxiety, fear, or even aggressivity. While some of us were fascinated by computers and awestruck by the potential they could glimpse in them, most people didn't understand them. They felt alien, abstruse, and in many ways, threatening. People feared getting replaced by technology.

Most of us react to technological shifts with unease at best, panic at worst. Maybe that is true of any change at all. But remarkably, most of what we worry about ends up never happening.

Fast-forward a few years, and the computer-haters have learned to live with them and to use them for their own benefit. Computers did not replace us and trigger mass unemployment—and nowadays we couldn't imagine life without our laptops, tablets, and smartphones. Threatening change has become comfortable status quo. But at the same time as our fears failed to materialize, computers and the internet have enabled threats that almost no one was warning us about in the 1980s and 1990s. Ubiquitous mass surveillance. Hackers going after our infrastructure or our personal data. Psychological alienation on social media. The loss of our patience and our ability to focus. The political or religious radicalization of easily-influenced minds online. Hostile foreign powers hijacking social networks to disrupt Western democracies.

If most of our fears turn out to be irrational, inversely, most of the truly worrying developments that have happened in the past as a result of technological change stem from things that most people didn't worry about until it was already there. A hundred years ago, we couldn't really forecast that the transportation and manufacturing technologies we were developing would enable a new form of industrial warfare that would wipe out tens of millions in two World Wars. We didn't recognize early on that the invention of the radio would enable a new form of mass propaganda that would facilitate the rise of fascism in Italy and Germany. The progress of theoretical physics in the 1920s and 1930s wasn't accompanied by anxious press articles about how these developments would soon enable thermonuclear weapons that would place the world forever under the threat of imminent annihilation. And today, even as alarms have been sounding for decades about the most dire problem of our times, climate, a large fraction (44%) of the American public still chooses to ignore it. As a civilization, we seem to be really bad at correctly identifying future threats and rightfully worrying about them, just as we seem to be extremely prone to panic due to irrational fears.

Today, like many times in the past, we are faced with a new wave of radical change: cognitive automation, which could be broadly summed up under the keyword "AI". And like many time in the past, we are worried that this new set of technologies will harm us—that AI will lead to mass unemployment, or that AI will gain an agency of its own, become superhuman, and choose to destroy us.

Image source: facebook.com/zuck

But what if we're worrying about the wrong thing, like we have almost every single time before? What if the real danger of AI was far remote from the "superintelligence" and "singularity" narratives that many are panicking about today? In this post, I'd like to raise awareness about what really worries me when it comes to AI: **the highly effective, highly scalable manipulation of human behavior that AI enables,**

and its malicious use by corporations and governments. Of course, this is not the only tangible risk that arises from the development of cognitive technologies—there are many others, in particular issues related to the harmful biases of machine learning models. Other people are raising awareness of these problems far better than I could. I chose to write about mass population manipulation specifically because I see this risk as pressing and direly under-appreciated.

This risk is already a reality today, and a number of long-term technological trends are going to considerably amplify it over the next few decades. As our lives become increasingly digitized, social media companies get increasing visibility into our lives and minds. At the same time, they gain increasing access to behavioral control vectors—in particular via algorithmic newsfeeds, which control our information consumption. This casts human behavior as an optimization problem, as an AI problem: it becomes possible for social media companies to iteratively tune their control vectors in order to achieve specific behaviors, just like a game AI would iterative refine its play strategy in order to beat a level, driven by score feedback. The only bottleneck to this process is the intelligence of the algorithm in the loop—and as it happens, the largest social network company is currently investing billions in fundamental AI research.

Let me explain in detail.

Social media as a psychological panopticon

In the past 20 years, our private and public lives have moved online. We spend an ever greater fraction of each day staring at screens. Our world is moving to a state where most of what we do consists of digital information consumption, modification, or creation.

A side effect of this long-term trend is that corporations and governments are now collecting staggering amounts of data about us, in particular through social network services. Who we communicate with. What we say. What content we've been consuming—images, movies, music, news. What mood we are in at specific times. Ultimately, almost everything we perceive and everything we do will end up recorded on some remote server.

This data, in theory, allows the entities that collect it to build extremely accurate psychological profiles of both individuals and groups. Your opinions and behavior can be cross-correlated with that of thousands of similar people, achieving an uncanny understanding of what makes you tick—probably more predictive than what yourself could achieve through mere introspection (for instance, Facebook Likes enable algorithms to better assess your personality that your own friends could). This data makes it possible to predict a few days in advance when you will start a new relationship (and with whom), and when you will end your current one. Or who is at risk of suicide. Or which side you will ultimately vote for in an election, even while you're still feeling undecided. And it's not just individual-level profiling power—large groups can be even more predictable, as average behaviors erasing randomness and individual outliers.

Digital information consumption as a psychological control vector

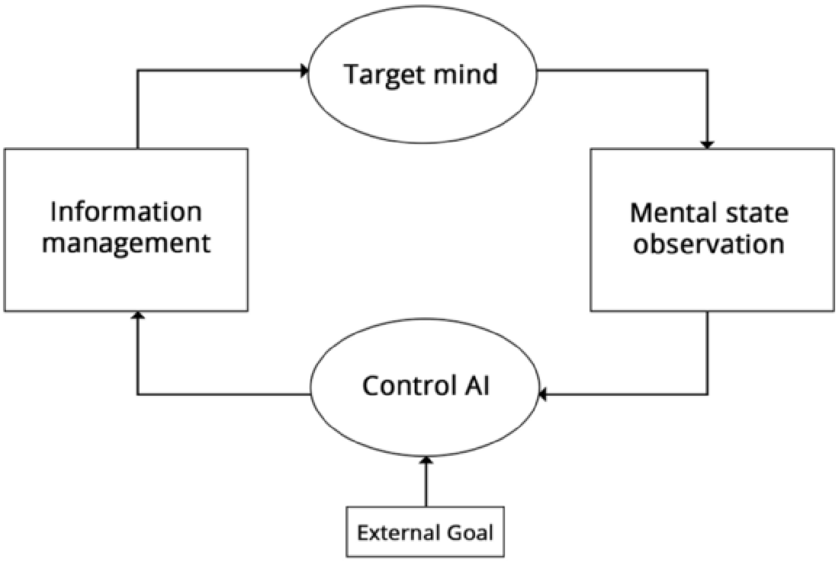
Passive data collection is not where it ends. Increasingly, social network services are in control of what information we consume. What see in our newsfeeds has become algorithmically "curated". Opaque social media algorithms get to decide, to an ever-increasing extent, which political articles we read, which movie trailers we see, who we keep in touch with, whose feedback we receive on the opinions we express.

Integrated over many years of exposure, the algorithmic curation of the information we consume gives the algorithms in charge considerable power over our lives—over who we are, who we become. If Facebook gets to decide, over the span of many years, which news you will see (real or fake), whose political status updates you'll see, and who will see yours, then Facebook is in effect in control of your worldview and your political beliefs.

Facebook's business lies in influencing people. That's what the service it sells to its customers—advertisers, including political advertisers. As such, Facebook has built a fine-tuned algorithmic engine that does just that. This engine isn't merely capable of influencing your view of a brand of or your next smart-speaker purchase. It can influence your mood, tuning the content it feeds you in order to make you angry or happy, at will. It may even be able to swing elections.

Human behavior as an optimization problem

In short, social network companies can simultaneously measure everything about us, and control the information we consume. And that's an accelerating trend. **When you have access to both perception and action, you're looking at an AI problem.** You can start establishing *an optimization loop for human behavior*, in which you observe the current state of your targets and keep tuning what information you feed them, until you start observing the opinions and behaviors you wanted to see. A large subset of the field of AI—in particular "reinforcement learning"—is about developing algorithms to solve such optimization problems as efficiently as possible, to close the loop and achieve full control of the target at hand—in this case, us. By moving our lives to the digital realm, we become vulnerable to that which rules it—AI algorithms.



A reinforcement learning loop for human behavior

This is made all the easier by the fact that the human mind is highly vulnerable to simple patterns of social manipulation. Consider, for instance, the following vectors of attack:

- Identity reinforcement: this is an old trick that has been leveraged since the first very ads in history, and still works just as well as it did the first time, consisting of associating a given view with markers that you identify with (or wish you did), thus making you automatically siding with the target view. In the context of AI-optimized social media consumption, a control algorithm could make sure that you only see content (whether news stories or posts from your friends) where the views it wants you to hold co-occur with your own identity markers, and inversely for views the algorithm wants you to move away from.
- Negative social reinforcement: if you make a post expressing a view that the control algorithm doesn't want you to hold, the system can choose to only show your post to people who hold the opposite view (maybe acquaintances, maybe strangers, maybe

bots), and who will harshly criticize it. Repeated many times, such social backlash is likely to make you move away from your initial views.

- Positive social reinforcement: if you make a post expressing a view that the control algorithm wants to spread, it can choose to only show it to people who will “like” it (it could even be bots). This will reinforce your belief and put you under the impression that you are part of a supportive majority.
- Sampling bias: the algorithm may also be more likely to show you posts from your friends (or the media at large) that support the views it wants you to hold. Placed in such an information bubble, you will be under the impression that these views have much broader support than they do in reality.
- Argument personalization: the algorithm may observe that exposure to certain pieces of content, among people with a psychological profile close to yours, has resulted in the sort of view shift it seeks. It may then serve you with content that is expected to be maximally effective for someone with your particular views and life experience. In the long run, the algorithm may even be able to generate such maximally-effective content from scratch, specifically for you.

From an information security perspective, you would call these *vulnerabilities*: known exploits that can be used to take over a system. In the case of the human minds, these vulnerabilities never get patched, they are just the way we work. They’re in our DNA. The human mind is a static, vulnerable system that will come increasingly under attack from ever-smarter AI algorithms that will simultaneously have a complete view of everything we do and believe, and complete control of the information we consume.

The current landscape

Remarkably, mass population manipulation—in particular political control—arising from placing AI algorithms in charge of our information diet does not necessarily require very advanced AI. You don’t need self-aware, superintelligent AI for this to be a dire threat—current technology may well suffice. Social network companies have been working on it for a few years, with significant results. And while they may only be trying to maximize “engagement” and to influence your purchase decisions, rather than to manipulate your view of the world, the tools they’ve developed are already being hijacked by hostile state actors for political purposes—as seen in the 2016 Brexit referendum or the 2016 US presidential election. This is already our reality. But if mass population manipulation is already possible today—in theory—why hasn't the world been upended yet?

In short, I think it's because we're really bad at AI. But that may be about to change.

Until 2015, all ad targeting algorithms across the industry were running on mere logistic regression. In fact, that's still true to a large extent today—only the biggest players have switched to more advanced models. Logistic regression, an algorithm that predates the computing era, is one of the most basic techniques you could use for personalization. It is the reason why so many of the ads you see online are desperately irrelevant. Likewise, the social media bots used by hostile state actors to sway public opinion have little to no AI in them. They're all extremely primitive. For now.

Machine learning and AI have been making fast progress in recent years, and that progress is only beginning to get deployed in targeting algorithms and social media bots. Deep learning has only started to make its way into newsfeeds and ad networks in 2016. Who knows what will be next. It is quite striking that Facebook has been investing enormous amounts in AI research and development, with the explicit goal of becoming a leader in the field. When your product is a social newsfeed, what use are you going to make of natural language processing and reinforcement learning?

We're looking at a company that builds fine-grained psychological profiles of almost two billion humans, that serves as a primary news source for many of them, that runs large-scale behavior manipulation experiments, and that aims at developing the best AI technology the world has ever seen. Personally, it scares me. And consider that Facebook may not even be the most worrying threat here. Ponder, for instance, China's use of information control to enable unprecedented forms of totalitarianism, such as its "social credit system". Many people like to pretend that large corporations the all-powerful rulers of the modern world, but what power they hold is dwarfed by that of governments. If given algorithmic control over our minds, governments may well turn into far worst actors than corporations.

Now, what can we do about it? How can we defend ourselves? As technologists, what can we do to avert the risk of mass manipulation via our social newsfeeds?

The flip side of the coin: what AI can do for us

Importantly, the existence of this threat doesn't mean that *all* algorithmic curation is bad, or that *all* targeted advertising is bad. Far from it. Both of these can serve a valuable purpose.

With the rise of the Internet and AI, placing algorithms in charge of our information diet isn't just an inevitable trend—it's a desirable one. As our lives become increasingly digital and connected, and as our world becomes increasingly information-intensive, we will *need* AI to serve as our interface to the world. In the long-run, education and self-development will be some of the most impactful applications of AI—and this will happen through dynamics that almost entirely mirror that of a nefarious AI-enabled newsfeed trying to manipulate you. Algorithmic information management has tremendous potential to help us, to empower individuals to realize more of their potential, and to help society better manage itself.

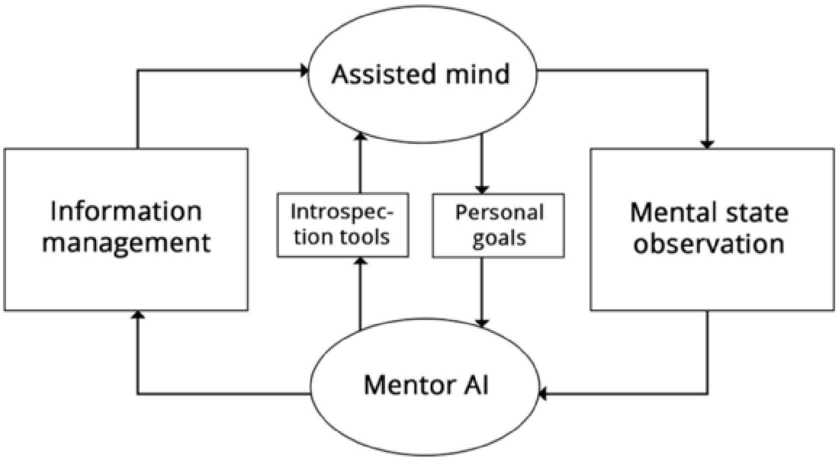
The issue is not AI itself. **The issue is control.**

Instead of letting newsfeed algorithms manipulate the user to achieve opaque goals, such as swaying their political opinions, or maximally wasting their time, we should put the user in charge of the goals that the algorithms optimize for. We are talking, after all, about *your* news, *your* worldview, *your* friends, *your* life—the impact that technology has on you should naturally be placed under your own control. Information management algorithms should not be a mysterious force inflicted on us to serve ends that run opposite to our own interests; instead, they should be a tool in our hand. A tool that we can use for our own purposes, say, for education and personal instead of entertainment.

Here's an idea—any algorithmic newsfeed with significant adoption should:

- Transparently convey what objectives the feed algorithm is currently optimizing for, and how these objectives are affecting your information diet.
- Give you intuitive tools to set these goals yourself. For instance, it should be possible for you to configure your newsfeed to maximize learning and personal growth—in specific directions.
- Feature an always-visible measure of how much time you are spending on the feed.

- Feature tools to stay control of how much time you're spending on the feed—such as a daily time target, past which the algorithm will seek to get you off the feed.



Augmenting ourselves with AI while retaining control

We should build AI to serve humans, not to manipulate them for profit or political gain. What if newsfeed algorithms didn't operate like casino operators or propagandists? What if instead, they were closer to a mentor or a good librarian, someone who used their keen understanding of your psychology—and that of millions of other similar people—to recommend to you that next book that will most resonate with your objectives and make you grow. A sort of navigation tool for your life—an AI capable of guiding you through the optimal path in experience space to get where you want to go. Can you imagine looking at your own life through the lens of a system that has seen millions of lives unfold? Or writing a book together with a system that has read every book? Or conducting research in collaboration with a system that sees the full scope of current human knowledge?

In products where you are fully in control of the AI that interacts with you, a more sophisticated algorithm, instead of being a threat, would be a net positive, letting you achieve your own goals more efficiently.

Building the anti-Facebook

In summary, our future is one where AI will be our interface to the world—a world made of digital information. This can equally lead to empowering individuals to gain greater control over their lives, or to a total loss of agency. Unfortunately, social media is currently engaged on the wrong road. But it's still early enough that we can reverse course.

As an industry, we need to develop product categories and markets where the incentives are aligned with placing the user in charge of the algorithms that affect them, instead of using AI to exploit the user's mind for profit or political gain. We need to strive towards products that are the anti-Facebook.

In the far future, such products will likely take the form of AI assistants. Digital mentors programmed to help you, that put you in control of the objectives they pursue in their interactions with you. And in the present, search engines could be seen as an early, more primitive example of an AI-driven information interface that serves users instead of seeking to hijack their mental space. Search is a tool that you deliberately use to reach specific goals, rather than a passive always-on feed that elects what to show you. You tell it what to it should do for you. And instead of seeking to maximally waste your time, a search engine attempts to minimize the time it takes to go from question to answer, from problem to solution.

You may be thinking, since a search engine is still an AI layer between us and the information we consume, could it bias its results to attempt to manipulate us? Yes, that risk is latent in every information-management algorithm. But in stark contrast with social networks, market incentives in this case are actually aligned with users needs, pushing search engines to be as relevant and objective as possible. If they fail to be maximally useful, there's essentially no friction for users to move to a competing product. And importantly, a search engine would have a considerably smaller psychological attack surface than a social newsfeed. The threat we've profiled in this post requires most of the following to be present in a product:

- Both perception and action: not only should the product be in control of the information it shows you (news and social updates), it should also be able to "perceive" your current mental states via "likes", chat messages, and status updates. Without both perception and action, no reinforcement learning loop can be established. A read-only feed would only be dangerous as a potential avenue for classical propaganda.
- Centrality to our lives: the product should be a major source of information for at least a subset of its users, and typical users should be spending several hours per day on it. A feed that is auxiliary and specialized (such as Amazon's product recommendations) would not be a serious threat.

- A social component, enabling a far broader and more effective array of psychological control vectors (in particular social reinforcement). An impersonal newsfeed has only a fraction of the leverage over our minds.
- Business incentives set towards manipulating users and making users spend more time on the product.

Most AI-driven information-management products don't meet these requirements. Social networks, on the other hand, are a frightening combination of risk factors. As technologists, we should gravitate towards product that do not feature these characteristics, and push back against products that combine them all, if only because of their potential for dangerous misuse. Build search engines and digital assistants, not social newsfeeds. Make your recommendation engines transparent, configurable, and constructive, rather than slot-like machines that maximize "engagement" and wasted hours of human time. Invest your UI, UX, and AI expertise into building great configuration panels for your algorithm, to enable your users to use your product on their own terms.

And importantly, we should educate users about these issues, so that they reject manipulative products, generating enough market pressure to align the incentives of the technology industry with that of consumers.

Conclusion: the fork in the road ahead

- Not only does social media know enough about us to build powerful psychological models of both individuals and groups, it is also increasingly in control of our information diet. It has access to a set of extremely effective psychological exploits to manipulate what we believe, how we feel, and what we do.
- A sufficiently advanced AI algorithm with access to both *perception* of our mental state, and *action* over our mental state, in a continuous loop, can be used to effectively hijack our beliefs and behavior.
- Using AI as our interface to information isn't the problem per se. Such AI interfaces, if well-designed, have the potential to be tremendously beneficial and empowering for all of us. The key factor: the user should stay fully in control of the algorithm's objectives, using it as a tool to pursue their own goals (in the same way that you would use a search engine).

- As technologists, we have a responsibility to push back against products that take away control, and dedicate our efforts to building information interfaces that place the user in charge. Don't use AI as a tool to manipulate your users; instead, give AI to your users as a tool to gain greater agency over their circumstances.

One path leads to a place that really scares me. The other leads to a more humane future. There's still time to take the better one. If you work on these technologies, keep this in mind. You may not have evil intentions. You may simply not care. You may simply value your RSUs more than our shared future. But whether or not you care, because you have a hand in shaping the infrastructure of the digital world, your choices affect us all. And you may eventually be held responsible for them.