To Serve and/or Protect?

AI and vulnerability in financial services: Principles and Questions to help firms Understand and Mitigate Risks, and Identify and Pursue Opportunities

Dan Holloway, *Rogue Interrobang, WhatWeNeed.Support, University of Oxford*

Table of Contents

Context

There is widespread vulnerability among the customers of firms in financial services[1]. These customers are more at risk of experiencing negative outcomes in their dealings with financial services companies. Harms such as exclusion from essential services, lack of access to suitable financial products, higher pricing, and more complex procedures can impinge on other areas of life. Many of those vulnerable to these harms have less bandwidth than others to start with, use more bandwidth to access essential services, and those services have a higher bandwidth cost for them than they do for others[2].

Yet much of that vulnerability remains undisclosed and/or unidentified[3]. AI tools offer the potential to reduce that identification gap, enabling firms to tailor journeys and put safeguards in place, thereby helping customers avoid the harmful outcomes to which they would otherwise be vulnerable.

This comes at the same time as widely publicised and debated concerns around the use of AI. Within communities with a high prevalence of vulnerability, AI-related worries may amplify existing concerns such as those around transparency, autonomy, and loss of access to services available to others. This means that as well as having the potential to transform people's financial lives for the better, the use of AI has the potential to erode trust in financial services firms, making the disparity of outcomes even more unequal.

The background against which this plays out was established by the Sunak Government's *A pro-innovation approach to AI regulation* in 2023. At the heart of this approach are five principles for regulating AI:

- Safety, security & robustness
- Appropriate transparency and explainability
- Fairness
- Accountability and governance
- Contestability and redress

The Financial Conduct Authority (FCA) issued an update in April 2024[4] outlining its approach to regulation based on these principles. That approach is built on a pro-innovation platform, with mitigation and risk assessment/management as the preferred mechanisms for protecting customers rather than red lines. It is "tech positive" in the sense that it sees technology primarily

---

[1] The 2025 Financial Lives Survey, carried out by the Financial Conduct Authority (FCA) found that 49% of UK adults had at least one characteristic of vulnerability https://www.fca.org.uk/financial-lives/financial-lives-2024

[2] See for example shafir and Mullainathan, *Scarcity*, Times Books, 2013

[3] Research by the Money and Mental Health Policy Institute found only 14% of people with a mental health condition had disclosed this to their financial services provider, whilst financial services firms routinely report a disclosure rate of substantially under 5% of their customers https://www.moneyandmentalhealth.org/wp-content/uploads/2022/11/Discosure-Guide-1-Disclosure-Environments.pdf

[4] https://www.fca.org.uk/publication/corporate/ai-update.pdf

as a tool for growth *and* for good, and it continues to create environments such as sandboxes and sprints to help firms explore possibilities and seek assurance.

At the same time, the FCA's Consumer Duty remains the essential framework for ensuring that firms do not provide worse outcomes for vulnerable customers, and that they avoid subjecting customers in vulnerable circumstances to foreseeable harms.

I would characterise this regulatory landscape as having a prior presumption in favour of seeing AI as a way of preventing harms to vulnerable consumers (by helping identify and mitigate possible harm) rather than as a possible cause of such harms.

Purpose

This paper addresses that context.

The principles outlined here do not take issue with the potential of AI to reduce harm. But they come from a belief that failing to understand the possible harms to which AI might expose vulnerable customers as well as those it might reduce or remove could present a degree of risk not just to consumers but to firms.

The principles and questions can be seen as tools for understanding and then managing that risk.

With that fundamental aim of risk management in mind, the intention of this paper is to help firms in the financial services sector use AI better in the context of serving vulnerable customers. For the purposes of writing, I am defining the use of AI as "better" if it does one or more of the following[5]

- Engages more consumers with a firm's services;
- Enables consumers to use services with less friction, AND/OR to do more in the rest of their lives because the lack of friction in those services reduces the exhaustion they would otherwise experience;
- Helps firms comply better with regulation;
- Enables firms to increase profit without shedding any of their customer base (especially the most vulnerable);
- Enables customers to avoid financial harms;
- Enables firms to spend more human time on those of its customers who most need it;
- Makes firms' customers trust them more, dislike them less, share criticism less and be more likely to recommend them.

[5] In each case the base for comparison is either "the status quo" or "using AI in a different way"

A Unique Perspective

The discussion points and practical suggestions in this paper are based on the following overlapping areas of experience and expertise.

Some of these are areas I share with many others. The combination of them is sufficiently unusual, I hope, to enable me to provide a perspective that is not just unique but unique, practical, and valuable.

- I have lived experience of bipolar disorder, ADHD and dyspraxia. This led to financial difficulties as a student, and have been exacerbated by lack of accessible communication channels from the financial services companies I have dealt with. Consistent inaccessibility of communication, platforms, and services continues to leave me vulnerable to less favourable experiences in the years since. I also care for a disabled spouse, and often act as a point of communication on their behalf.
- I have spent two decades working as an advocate, speaker, and consultant in the area of financial services and inclusion. I have worked with the FCA, Credit UK, Money Advice Trust, Money and Mental Health Policy Institute, and with Experian helping with the design and implementation of their Support Hub portal.
- I am founder and CEO of Rogue Interrobang, a spinout company from the University of Oxford. We help institutions and individuals to live more creative lives. Half of what we do is rooted in my research into and experience of creative thinking (as a writer, curator, performance poet, and 5 times Creative Thinking World Champion). And half is based on eliminating barriers to accessibility. At the root of both is a belief that creativity and accessibility are two sides of the same coin. I have worked with organisations from schools to intelligence analysts.
- Since 1995 I have been researching phenomenology, narrative, and disability. My academic background is in theology, with a focus on mediaeval and early modern history of ideas, in particular ideas about what it means to think, to know, to imagine, and to create. I am interested in the way these activities are categorised, and the way they are experienced (and most of all in how the way they are categorised influences the way they are experienced and vice versa).
  More recently, through the Futures Thinking Network at The Oxford Research Centre for the Humanities (TORCH) I have directed this approach specifically to the way discourse about technology and/or utopias affects the experience of disabled people using that technology and our ability as disabled people to imagine ourselves living in these utopias.

Vulnerability: An Introductory Note

Remember that it is not just customers who can be vulnerable. As firms, you can be vulnerable as well. Like consumer vulnerability, the vulnerability of a firm can shift over time. And like the vulnerability of your customers, the key question is, “vulnerable to what?”

The principles and questions that follow focus on those moments when a firm is particularly vulnerable: to reputational harm, to customer disengagement, to the regulators’ ire, to sinking large amounts of money into implementing the wrong product or solving the wrong problem. Rather like the work firms are trying to do for customers, these principles and questions are designed to provide early intervention at times of vulnerability, mitigating potentially damaging outcomes.

Principles

This is a summary of some key principles for firms to consider.

Other sets of principles exist addressing the use of AI in the context of vulnerability. Many more will exist. This list has overlaps, but it has deliberate differences and omissions. That is not to disparage other principles. Rather it acknowledges the value of seeking many perspectives. And the value of recognising that to navigate a very human space such as the relationship a firm has with its vulnerable customers requires balancing many perspectives, reflecting the complexity of being human, rather than seeking a single exhaustive checklist.

This list is written from the intersection of perspectives and experience I outlined in the previous section.

1. Beware the Future consequences of information asymmetry

“Customers are happy with AI if it gives them an experience that works for them.” This is something I hear a lot from firms. And it’s true. It’s as true of me as it is of my disabled peers, and I have spent more than a decade advocating that “we just want it to work.” So when I add a huge caveat, it doesn’t negate that.

The caveat is simple. The history of business scandals is littered with instances where customers ticked boxes and expressed happiness with a process because it gave them an experience they valued. And what each of those cases, from PPI to Cambridge Analytica, has in common is that the initial proposition involved a huge information asymmetry. Which became a problem when people eventually saw behind the curtain of that asymmetry. Consumers will generally express comfort with a risk when they do not feel an experiential connection to its consequence (that is, when they don’t fully “get” the consequence OR when they don’t FEEL that the down-the-line consequence will directly result from the action they are taking right now, because they only feel the connection to the good bit that comes straightaway).

That doesn’t mean the answer is endless reams of explanatory notes that create huge frictions. And it doesn’t mean taking away people’s right to such services. I am, after all, a frequent

advocate against paternalism and for the notion that autonomy involves the freedom to make bad choices. But it means being aware that information asymmetries create a structural advantage for the industry side of a contract, and that structural advantage wraps up a risk that at some point in the future a court or a regulator or other body will decide it pushed the edges of contract law too far.

This is why so many of these principles are based on focusing firms' attention on how they respond to scenarios in the future, which may play out in environments that may be less tech positive. It has, after all, happened again and again in the past, and if it happens in the future, the firms that saw it coming will be at a huge advantage. Whether or not that outweighs the competitive disadvantage of planning for it is a matter for individual firms, but not realising that the choice exists would be egregious. "The opinion of regulators, governments, and consumers may vary over time" should be the industry-side equivalent of the warnings the industry provides to customers that prices may fall as well as rise.

2. AI should first serve accessible design

Accessible design will solve many problems for many vulnerable consumers. Many of these are simple steps such as multiple communication channels, or offering options for length and timing of meetings and calls. These don't require gathering special category (indeed, any) data or using training data in ways that customers may find unpalatable. If you decide to use AI, first do so to provide multiple ways of undertaking a single process, such as taking a form and converting it into multiple file formats, multiple methods of input, multiple levels of detail or technical vocabulary, and different languages and media.

This will greatly reduce the number of customers you struggle to serve. You can then make an informed decision on whether you need a complex, controversial and costly tool to identify and serve those still missing: or just a well-paid. slightly expanded, and time-rich customer service team.

3. Everything is part of a story

At the most fundamental level, all your customers' behaviour, whether you acknowledge it or not, is part of not just one but many stories: stories they tell about themselves; stories society and family and friend groups and companies tell about them; stories in which your platforms and product have one, or many, parts to play. To understand and leverage this will lead to greater customer engagement and satisfaction; more pitfalls avoided at the inexpensive prototyping stage; less anxiety about low key audiences will react to new products or policies.

This requires involving at least one perspective in any design and build phase that is grounded in the Humanities.

Each (potential) customer will experience their encounter with everything you write and everything you do not as "learning a fact" but as a part of one or more of their own personal

stories, every one with a history, with emotions attached to it, and with hopes and fears for how it plays out next.

Everything you say and do around AI matters because it is the part of the story that gets your customers, in their head, from whatever their past is to some kind of future. It boils down to this: after interacting with something you have said or done, does your customer feel closer to their hopes or to their fears? Could saying or doing it differently nudge them more towards the former and less the latter? This is the hardest concept to grasp in this list, but only because it's unfamiliar to many of us in our business life (though rarely our personal life).

This is not about AI Ethics. AI Ethics is vital, but it is about whether you should or should not deploy things and how you should do so. The approach here is a methodological one that encompasses the process of de[ployment. It's the actual doing of the thing.

The points I made in principle 1 about customers FEELING a connection to consequence and how important that is for decision making is an example of how Humanities-rooted thinking can help inform a firm's understanding of consumer behaviour, how it can change over time, and how that affects the firm's handling of risk. In this case, it comes from my own discipline, phenomenology.

4. Focus on how you respond to your customer's WTF moments

We have all experienced moments when we start browsing and it feels like our computer's been listening in on us. As detailed in the film The great Hack about Cambridge Analytica, that's because, essentially, it has. And when people realise that, that realisation can come with a sense of violation and betrayal, and create a general feeling of anger and distrust.

It is not the immediate experience that creates these negative feelings. The immediate experience might be one of ease, convenience, relief, even gratefulness at a lack of friction or how easily one has been able to access a journey that feels personally curated.

Rather, the negative reaction comes later when people realise how that personal curation happened, and what else it means about the use of the information driving it. Because the personal information that helps provide an online journey that you identify the kind of short that would go really well at next week's wedding, or even get a sense check before making a financial mistake can be used for a whole lot of things you might be less pleased about.

And the real kicker is when someone realises they don't remember being asked whether they wanted this.

This principle is not an instruction to tell everyone everything at the time. It is a recommendation to think about people's reaction when they find out you have done something without telling them. And decide whether, and how, you mitigate that reaction (in the moment, or before it arises).

5. Decide on your red line and build actions out from it

This builds on the previous principle. That deals with your response to a customer's discovery of something you have done. This deals with your response to a customer's request for something you won't do.

I recommend every firm ask, "What is the thing you won't do if your customer asks you to?" This is your red line when it comes to consent, and it is the point at which you are most vulnerable to negative reactions and viral criticism.

Once you have identified this (it is most likely to be some kind of safeguarding), ask the following.

What is the reason for your doing this irrespective of consent?

And ask yourself

What do you tell your customers is the reason for not telling them and doing it anyway when they ask?

Any gap between the two is your vulnerability.

6. In the future firms will have AI tools. So will customers. So will bad actors

AI offers many people the prospect, in the next few years, of being able to access services that have been inaccessible to them all their lives, or for which they have been reliant on the support or intervention of others. But this help comes with great risk.

And just as AI tools help firms to create better interactions with customers, so customers will use AI tools to create better interactions with firms. This will include tools that can not only find but open accounts involving multiple steps for those of us who find executive function hard, and providing AI-generated copies of our voices to make calls on behalf of those of us who can't use the phone.

These tools of ours will present significant risks to firms. In part because just as customers must be aware that the data flowing into and out of company-side AI is at risk from interception and use by bad actors, so firms will be aware of the risk of bad actors posing as customers.

Meanwhile AI-generated voice, assistance, and video will lower the usefulness of AI tools designed to detect vulnerability in changes of tone or appearance or screen navigation or behaviour, and may ultimately render the datasets behind those tools useless.

This is something firms and customers need to tackle together and urgently.

7. Many of the most vociferous vulnerable consumers are among the most, and most strongly, AI-sceptical

Some of the most vulnerable customer groups participate in the maker economy, whether that be crafting on Etsy, self-publishing on KDP select, freelancing on Fiverr, or making content for DeviantArt or Soundcloud. For many, this is the only means of generating income available to them. And these communities are the ones who have most felt and most been concerned about the training of generative AI platforms on copyrighted material scraped from the internet. The scale of this abuse can be seen in the ongoing class action against Anthropic which has identified a class of nearly 500,000 pirated versions of copyrighted books used to train the large language model Claude[6].

This means that among a significant portion of the vulnerable community firms want to assist by using AI tools there is an anger about, awareness of, and suspicion around AI not present to the same degree elsewhere.

Again, this principle is not intended to recommend using or not using AI tools. It is rather a reminder to firms of things to consider before they decide on their actions and most important how those actions are communicated.

8. **Understand ALL the risks of false positives**

Even the best AI will sometimes identify vulnerability where there is none, or misdirect either a human operative or a customer into undertaking something inappropriate to the circumstances.

Most people, in most cases, when offered something based on the assumption that they might be vulnerable to harm, would shrug it off, or at worst mutter under their breath about overcautious coddling.

But consider someone who has, say, a complex set of needs in one area but is not at risk in the setting of finding the best car insurance, for example. They may have a significant history of firms making assumptions about them, being denied access to products, or being denied opportunities more widely as a result of incorrect assumptions about what they are and are not able to cope with or achieve.

Find ways to identify the risk to particular groups of customers (see principle 11 on using lived experience). The consequences extend beyond financial harm to feelings of trauma and betrayal.

---

[6] See for example the 25 September 2025 statement from the Authors Guild of America https://authorsguild.org/advocacy/artificial-intelligence/what-authors-need-to-know-about-the-anthropic-settlement/

Identify what actions may follow from those harms and reactions. And identify how they would impact you as a company, such as media or social media explosions and resultant reputational risk, or even changes to the regulatory landscape.

9. **Understand how past trauma can shape future interactions**.

This is really a subsection of principle 2, but such an important one I have given it a principle of its own. Your customer's reactions shouldn't catch you off guard. If you understand what causes negative reactions, you can avoid many of them, and have the right response ready for those you can't avoid.

10. Know where your balance between efficiency and experience lies, and understand the process by which you calculate it

Are you freeing up time to improve the bottom line or are you doing so to increase the amount of time and attention you can provide to those who need it most? There is a business case for doing both. Complex customers don't equal unprofitable customers. They are far more likely to do so where seemingly "affordable" tools that do not meet their needs are employed, leading to poor outcomes and greater administrative input down the line.

11. **Involve lived experience**

It is vital to involve people with lived experience in your thought and design processes or you run the risk of solving the wrong problem, of being surprised by outcomes you should have predicted, and of finding things out after you have sunk costs into development rather than during prototyping.

Disabled people are in the best position to tell you their own stories, the experiences that shape their responses, and the things they need from any product, policy, process, or platform. So involve us in designing and building your project. Getting our input at the earliest stage will minimise your cost and maximise our communities' buy-in. There are whole papers that could be and indeed are written on how to do this but overarching principles are:

- Combine breadth and depth. Work for the duration of the project or long parts thereof with a few people who understand and believe in your product and can bring experience relative to vulnerable communities and your sector so that they understand the need for pragmatism and authenticity alike. And work in short bursts with a wide variety of people, asking specific questions.

- Make the design and build process an act of co-production with lived experience fully involved (in the room), enabled (given the space and tools needed to make meaningful contributions) and empowered (given a meaningful vote in key decisions), not just consulted.

12. Learn from history

Recent history is full of examples where processes or policies that seemed reasonable, even beneficial, and profitable have had catastrophic (and in some cases tragic) outcomes for customers. And where those outcomes for customers has led to vast financial loss or a change in legislation that has been met with horror by the industry oin which it was imposed.

I am not suggesting that using AI to assist with vulnerable customers will cause tragedy or catastrophe. Nor am I suggesting firms are ill-intentioned. But being ill-informed about the past and therefore not spending time identifying where it **could** repeat would be very unwise. However low the risk, the consequence for the financial services industry is huge.

The following all provide cases that would benefit from study.

- The alleged role of social media companies allowing children to access harmful material and the introduction of the Online Safety Act;
- PPI misselling;
- Cambridge Analytica;
- Tabloid phone hacking (an interesting control case given the lack of consequence following Leveson);

Focus in particular on the following elements of each case:

- The consequence for individuals that led to action being taken against firms;
- How much firms were aware of these consequences before taking action;
- Whether warning had been given either internally or externally before action was taken (a different slant from the previous consideration);
- What elements came together to create the pressure that led consequences to become action (why were policy makers, media, pressure groups engaged specifically in these instances and how were they able to turn that engagement into widespread traction?). Focus particularly on the human elements of the stories told by those seeking action;
- The extent to which a route for early redress had been offered and ignored;
- The extent to which a firm's initial and subsequent response, in media, in silence or statement, and in court, moved the dial of opinion or action.

13. Red team. Repeatedly

Red teaming is a form of scripted role-playing or scenario testing. It allows you to explore how you would respond to potential catastrophes you have identified in principle 12, and to do so in a safe sandpit environment. Typically red teaming involves bringing in a facilitator who will role

play all external stakeholders, from customers to media, politicians and regulators., It is an exceptionally good way of stress testing in a no-blame exploratory setting.

Many of the principles in this paper, as befits a Humanities-based approach, call on a risk-consequence analysis to be carried out, where the measure of risk and of consequence has a qualitative element. Red teaming is an ideal way to explore this.

Questions

These questions flow naturally from the principles above. Answering them with those principles in mind will provide the same benefits as those principles: increased customer engagement and trust; better outcomes for customers; lower costs because issues are sidestepped at the prototyping stage; lower risk of reputational damage and customer dissatisfaction.

1. How do transparency and consent relate in this context?

Transparency is not only encouraged by the FCA. It is a sound principle of accessible design and good user experience more generally. Whilst not everyone wants to know what's "under the hood" of the products and services they rely on, it is good practice for the information to be available for those that do.

Transparency serves many purposes, but almost all of them can be expressed in a single word. Choice

When you choose something after seeing under the hood, you go in eyes wide open. And you can properly be said to have consented to the choice you made. This is why whenever firms ask for consent, they spell out exactly what you are consenting to.

But meaningfully consenting to a choice you make, implies that you could have made a different choice. And there are two facets of financial products that mean this is not necessarily the case.

First, the more general point that if not consenting to the use of AI means you cannot offer a product, then being transparent about your AI use has not provided real choice.

Second, more specifically, financial products and other essential services may be things a person needs, possibly desperately needs. This creates an asymmetry in which the desire or need for the product can override misgivings about the conditions of acquiring it. Often no amount of transparency will override that desperation, raising questions about how much of a "choice" was made, transparency or no.

This is not to imply that transparency is anything other than a good thing (though information overload in the name of transparency should be avoided, with the ideal being different levels of information being offered according to the preferences of individual customers). Rather, it is

important that ticking the transparency box not be used as an excuse for not offering meaningful consent. There are things that you as a firm will do whatever a customer says or does. And when that is the case, that is the thing to be transparent about. At least it offers the customer a chance to avoid a product or firm altogether. Though when a whole sector does this, that might be an issue for regulators.

2. Which existing inequalities might AI widen?

Existing inequalities within a firm's customer base could be widened by either lack of access to the digital spaces where AI is used to improve outcomes, through bias encoded within platforms; or through biased interpretation of the prompts to staff generated by platforms.

3. How will the AI you use improve?

This question and the next go together.

Will your customer data be part of the improvement process? Will that improvement be walled off and limited to you, or will it enable the company providing it to improve the overall model? If the latter, there are data and trust issue; if the former, there will be performance issues

4. How will you know if your use of AI has been successful?

"What is on the right hand side of the graph?" is one of the most fundamental questions any project manager can ask. Without knowing where they want to get to, it is almost impossible for any project manager to assess their progress along the way.

Firms acknowledge the difficulty of figuring out what to measure or how to measure it.

But they are less adept at recognising this poses fundamental questions about the way they approach the endeavour of creating better outcomes for consumers. Because it makes it almost impossible to assess whether the method they are undertaking is actually improving anything, and whether an alternative would do a better job. It makes it impossible to create go/no go points for any undertaking, essentially meaning the project's future will depend on the psychological impact of the sunk cost on decision makers.

5. **Does your AI know your customers better than they do**

This is not a question I want you to spend hours researching. Rather, it is a question about a very basic belief that you have as a company. I have saved it for the end because in a way it wraps up everything else here. How you answer this question will determine how you proceed in almost every other situation. So once you have answered it, go back and revisit everything else

on this list in the light of that answer. Just as you might revisit a film in the light of an unexpected twist in the ending.

Why do you need to answer this question? In part because so much of everything above relies upon there being an answer. But in part because when an AI says one thing and a human says another, each constitutes evidence against the other. And, as the entity that needs to take meaningful action based on which statement is true, you need to know which evidence to weight more highly.

To correspond on the subjects in this paper or to discuss working with me to help explore what the principles and questions in the paper mean for you, your products, platforms, and customers, email me at rogueinterrobang@gmail.com