FINTRUST

Newcastle University

Engineering Trust in Financial AI

A Trust Engineering Toolkit for the Financial Industry

FinTrust Project // 2022

https://fintrustresearch.com/

Contents

Cite this report
Mitchell Finnigan, S., Copilah-Ali, J., Deng, J., Huan, R., Ma, Z., Wu, H., Spiliotopoulos, T., Aitken, M., Horsfall, D., Ng, M., Toreini, E., Coopamootoo, K., Elliott, K., and van Moorsel, A. (2022). *Engineering Trust in Financial AI*. Newcastle University, Newcastle upon Tyne, United Kingdom. p.1-22. Available at: https://fintrustresearch.com/assets/pdf/FinTrustToolkit.pdf

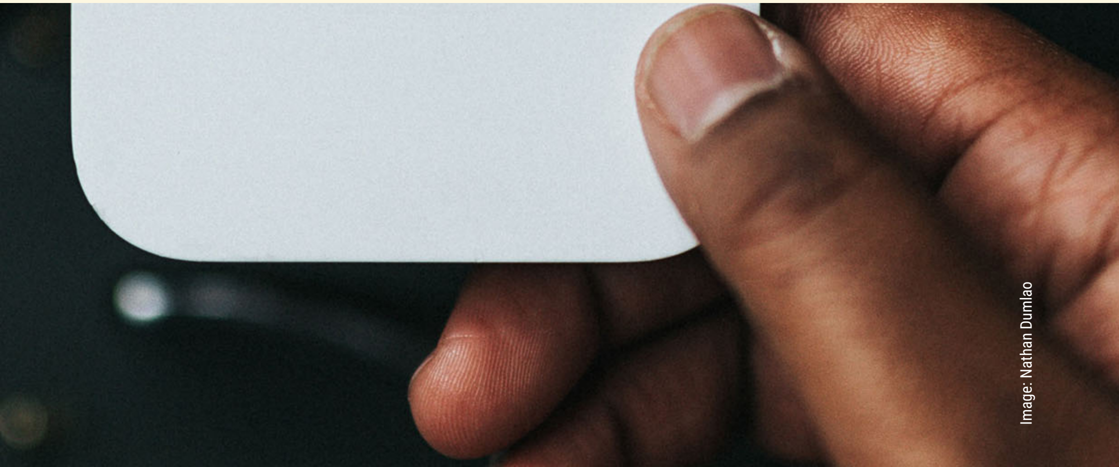


Image: Nathan Dumlao

Are your customers excluded?

We all need to be able to trust the financial services we use.

But in the age of new digital banking technologies and "FinTech", understanding how we should build trustworthy financial services is a topic of public concern. This toolkit explores such problems inherent in engineering trust into financial services, and discusses potential tools and solutions arising through new research undertaken at Newcastle University in the FinTrust project.

Digital banking can pose challenges for citizens, raising vulnerability concerns around trust in banking institutions: branch closures, accessibility of online services, and digital skills and understanding all play a role in the inclusion and participation of people in the digital society.

FinTrust investigates digitization and its impact on society. Our Trust Engineering Tool Kit translates understandings and techniques developed through the FinTrust project to help your business to better understand trust and trustworthy systems within the increasing automation of financial services.

1. Vulnerability & Finclusion

Image: Tim Mossholder

Vulnerability: Why should we care?

Can blockchain technologies support vulnerable and financially excluded customers?

Some of the most vulnerable members of our society are marginalised by advances in technology. In thinking about "finclusion", we have looked at some of the drivers of this exclusion. This section examines the contexts where trustworthy technologies are of specific importance, with respect to communities and individuals who are typically marginalised within the financial industry.

Things to consider:

- Are you serving customers who meet the FCA's vulnerability drivers?
- Do you know which of your customers are in a vulnerable situation?
- Do you check and update the vulnerability status of your customers?

We found that:

- Over 53% of the UK population sit under the four FCA drivers of vulnerability: health, life events, resilience and capabilities[1].
- Most technology is not built to adapt to vulnerabilities.
- Other forms of digital identity can meet KYC (Know Your Customer) FCA regulations to drive financial inclusion, and can help meet ESG (Environmental Social Governance) and SDG (Sustainable Development Goals) UN directives.

1.1 Trustworthy Digital Identity

In this project, we implemented a prototype decentralised identity system using the Microsoft ION platform, to allow customers to retain ownership of their personal details and to share them securely.

This project was part of the Trustworthy Digital Infrastructure for Identity Systems project, led by the Turing Institute and funded through a grant from the Bill & Melinda Gates Foundation (INV-001309).

[1] https://www.fca.org.uk/publication/finalised-guidance/fg21-1.pdf (p.9)

For more information, our final project report on our verifiable credentials Finclusion project can be found at our website.

1.2 Using Verifiable Credentials to Identify Vulnerable Customers

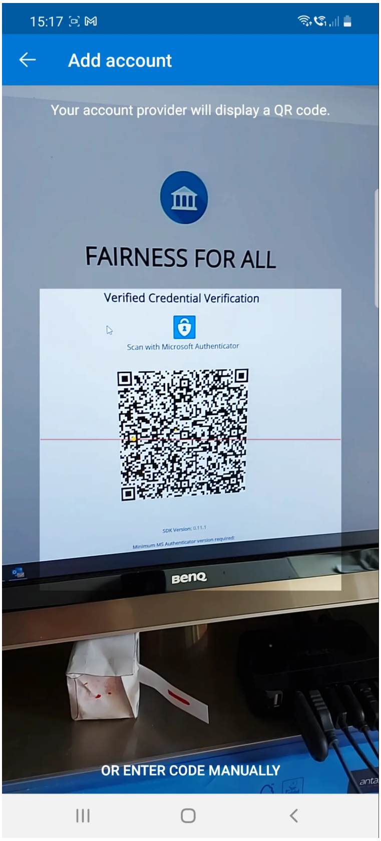
Digital identity systems are used worldwide, from "digital passports" to online log-in systems, but this project sought to investigate how "trustworthiness" might be designed-in to such systems. Trust is characterised here through several characteristics, including security, privacy, ethics, resilience, robustness and reliability, yet there remain significant challenges in designing systems which embody these concepts.

Vulnerable customers within the financial sector are particularly important to consider within this framework, and such a trustworthy identity system should not further exclude people from participation in the financial industry. In the United Kingdom, the FCA (Financial Conduct Authority) has issued guidance in this respect, which strongly encourages fair treatment of vulnerable customers, but financial institutions often lack a coherent strategy to the identification of client vulnerability.

The sociotechnical challenges in this space include the (often manual) disclosure and handling of vulnerabilities, integration with support processes, and risks for collusion and fraud. A detailed look at this problem space can be read in our publication (Spiliotopoulos, 2021).

Decentralised Identifiers (DIDs) and Verifiable Credentials (VCs) hold potential for improving the identification and disclosure process for such vulnerable customers, and allowing the provision of tailored financial services and products.

We produced a design specification for such a



Your account provider will display a QR code.
FAIRNESS FOR ALL
Verified Credential Verification
Scan with Microsoft Authenticator
SDK Version: 0.11.1
Minimum MS Authenticator version required:
BenQ
OR ENTER CODE MANUALLY

Fig. 1: The credentials issuance process.

system in 2021, which provides an implementation of the World Wide Web Consortium (W3C) standards for DIDs v1.06 and Verifiable Credentials Data Model 1.07 in a Microsoft Azure environment. This drove our discussion and evaluation of potential solutions for the use case of vulnerability in finance.

Finally, we produced and deployed a software prototype based on this specification, for evaluation in interviews, workshops and focus groups.

1.3 Code

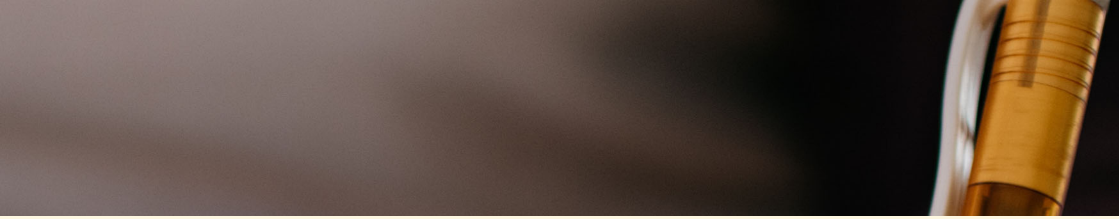
The prototype project is split into two code repositories: the issuer and the verifier. These can be accessed at:

- https://github.com/NewcastleRSE/fintrust-vc-issuer
- https://github.com/NewcastleRSE/fintrust-vc-verifier

For issues, please use the GitHub issues feature, or feel free to get in touch by emailing fintrust@newcastle.ac.uk

1.4 Relevant publications

- Spiliotopoulos, T., Horsfall, D., Ng, M., Coopamootoo, K., van Moorsel, A. and Elliott, K., 2021. Identifying and Supporting Financially Vulnerable Consumers in a Privacy-Preserving Manner: A Use Case Using Decentralised Identifiers and Verifiable Credentials. ACM CHI 2021 workshop on Designing for New Forms of Vulnerability. https://arxiv.org/abs/2106.06053
- Elliott, K., Price, R., Shaw, P. et al. Towards an Equitable Digital Society: Artificial Intelligence (AI) and Corporate Digital Responsibility (CDR). Soc 58, 179–188 (2021). https://doi.org/10.1007/s12115-021-00594-8
- Aitken, M. et al. (2020) 'Establishing a social licence for Financial Technology: Reflections on the role of the private sector in pursuing ethical data practices', Big Data & Society. https://doi.org/10.1177/2053951720908892



2. Trustworthy Interfaces

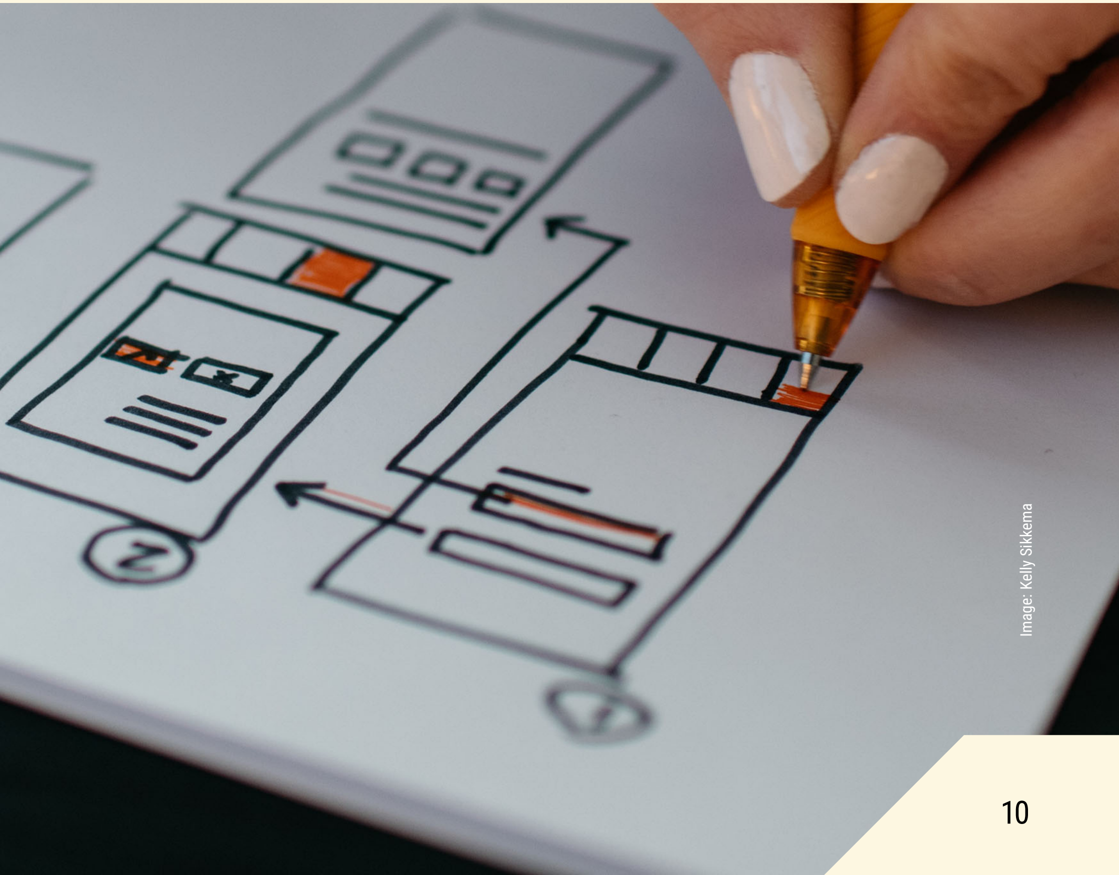


Image: Kelly Sikkema

How do we trust interfaces?

Chatbots are increasing in prevalence, sometimes even equipped to mimic human social rules, expectations and norms, decreasing the necessity for human-to-human interaction. As banks and financial advisory platforms move towards creating bots that enhance the current state of consumer trust and adoption rates, we investigated the effects of chatbot vignettes with and without socio-emotional features on intention to use the chatbot for financial support purposes.

Things to consider:

- Do you need a chatbot?
- Have you asked your customers what they prefer?

We found that:

- Your customers may find it easier to trust a chatbot which seems less human
- Hybrid chatbots which give an option to chat to a human may be preferred

2.1 Attitudes to Artificial Intelligence

Within this project, we explored peoples' attitudes to managing their financial assets using a chatbot. Given the results of this work, we prototyped a conversational interface using the Google DialogFlow conversational analysis AI.

A chatbot is an artificial intelligence application that can imitate a real conversation with a human in their natural language. Chatbots enable communication via text or audio on websites, messaging applications, mobile apps, or telephone, leading to potential for the inclusion of groups of people who would otherwise be excluded by existing online financial technologies.

While many of us are familiar with chatbot technologies, research into these virtual agents indicates that how they are presented strongly influences what kinds of information end users are trust them with. Our project publication by Ng et al. (2020) shows how important the psychology of trust is for designing human versus machine-like chatbot interfaces.

One approach to increasing trust of computer systems lies in designing human-like chat interfaces which display social and emotional intelligence: for example, by simulating politeness; demonstrating active listening skills; generating empathetic responses; and personalising to the individual by using their preferred name. However, it wasn't clear whether it would be possible to apply this to the finance and FinTech context.

We compared two hypothetical chatbots: "Emma" described with human-like emotive features, and "XRO23", a system described in more mechanical and impersonal terms. Both chatbots were described as secure and reliable in vignettes outlining their use as an automated financial adviser. Our study found that people were more willing to disclose sensitive financial information such as an account number and sort code to the impersonal "XRO23" system, without humanistic traits. This suggests that socio-emotional features in chatbots designed exclusively for automated financial support have little advantage in FinTech.

2.2 Chatbot Prototype

Following this research, the decision was made to create a prototype chatbot using Google DialogFlow. In order to recognise user commands in typed natural language phrases, a language model is required to 'understand' what the end user wants. This is called natural language processing, or 'NLP'. We can leverage existing commercial technologies for this purpose, and our technology review document covers existing technologies and emerging trends in conversational interfaces, as well as social and ethical considerations for their use.

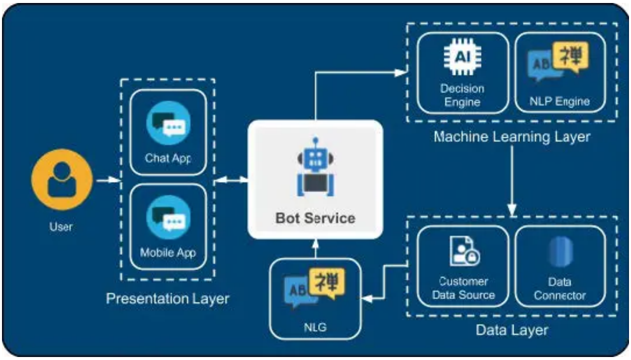


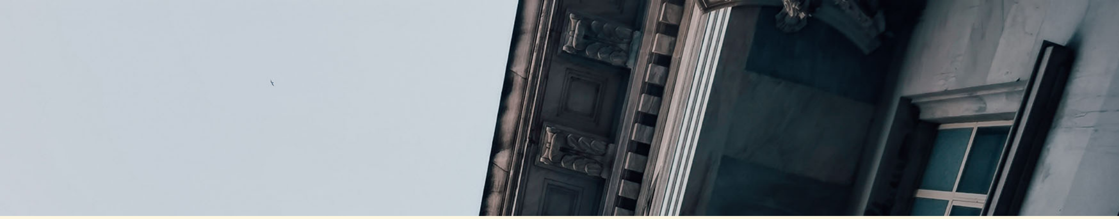
Fig. 2: A block diagram for a chatbot application design using a natural-language model to understand requests.

We selected Google DialogFlow for our chatbot prototype, following an analysis of privacy considerations under UK data protection law. A Laravel-based server application was developed to connect to this API.

By prototyping this application we hope to gain understandings of the implementation issues surrounding financial chatbot systems in-the-wild. The application is currently in a developmental state and therefore remains closed-source. For access, please contact us.

2.3 Relevant publications

- Aitken, M., Ng, M., Horsfall, D., Coopamootoo, K.P., van Moorsel, A, and Elliott, K., 2021, August. In pursuit of socially-minded data-intensive innovation in banking: A focus group study of public expectations of digital innovation in banking. In Technology in Society vol. 66 (pp. 101666). Elsevier. https://doi.org/10.1016/j.techsoc.2021.101666
- Ng, M., Coopamootoo, K.P., Toreini, E., Aitken, M., Elliott, K. and van Moorsel, A., 2020, September. Simulating the effects of social presence on trust, privacy concerns & usage intentions in automated bots for finance. In 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW) (pp. 190-199). IEEE. https://arxiv.org/abs/2006.15449



3. Checking Bias in Artificial Intelligence

Image: Iro Klg

Can we trust algorithms?

How we talk about trust is subject to assumptions on what ’trust’ means: how we understand it at a social level is not the same as trustworthiness at the systems level.

Things to consider:

- How does your company understand human versus technical trust?
- How do your customers or service users perceive trust?
- How does your organisation measure trust and build trustworthy technologies?

We found that:

- People’s trust perceptions and beliefs surrounds that an organisation or system is honest, reliable, or safe.
- "Trust" means different things to different people.
- For engineers and designers, "trustworthiness" is concerned with the "FEAS" properties of machine learning systems: *Fairness, Explainability, Auditability and Safety*.

3.1 Understanding the Difference Between "Trust" and "Trustworthiness"

How we talk about trust is subject to assumptions on what "trust" means: how we understand it at a social level is not the same as trustworthiness at the systems level. To us, "trust" means peoples’ perception or belief that an organisation or system is honest, reliable, or safe, but it is nebulous in that each of us will imagine what it means to trust in slightly different ways. When we say "trustworthiness", we are talking about the property of machine learning systems. This is what our research examines, and there is a need to pin down exactly what "trustworthiness" means in the context of AI and ML systems. In other words, when talking about trust, we have to ensure that the language in use is understood as the same thing by all parties.

3.2 FEAS: Fairness, Explainability, Auditability and Safety

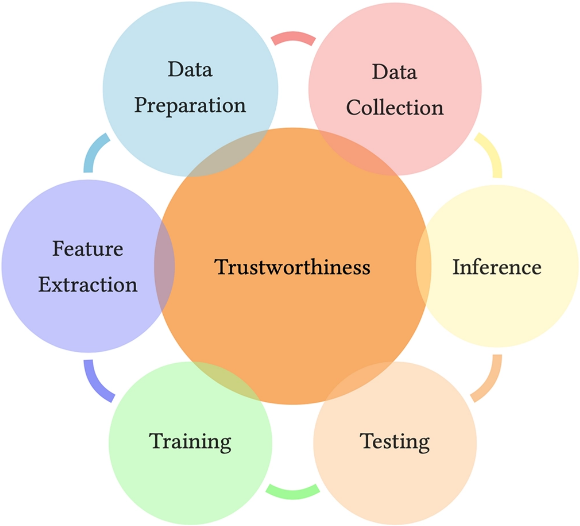


Fig. 3: Stages in the machine learning pipeline, motivating the notion of a Chain of Trust. (Toreini, 2020)

Machine learning pipelines can be broken down into stages: the data-centric stages (data collection, data preparation, feature extraction), and the model-centric stages (training, testing, and inference). To enable trust in the whole system, trustworthiness must be considered in every part of the pipeline. In our research, we call this the "Chain of Trust" (Fig.

3). Further, we have defined a set of properties which must be observed to achieve trustworthiness. Machine learning system developers must understand these four properties: Fairness, Explainability, Auditability and Safety (FEAS).

Fairness

Fairness is the property of non-discrimination. Machine learning algorithms are not deliberately built with planned bias: they are not inherently fair or unfair. However, if unfairness and discrimination from the real world is reflected in training datasets, this can result in the replication of bias against minority groups and an unfair outcome for people. Designing discrimination-free models is therefore a challenging problem.

Explainability

Explainability means the ability to understand why a result is the way it is. It is often used interchangeably with "transparency". For simple types of predictive model, explainability can be achieved through model-specific explanations. More complex models such as neural networks operate in a non-linear fashion, which means they are effectively a "black box". Various frameworks have been proposed to explain how a deep learning model reaches a result, but it is widely agreed that there is still no single consensus on how to do it.

Auditability

Auditability is about making a machine learning process transparent to stakeholders. For third-parties to verify the operation of a model, we must establish an audit trail. In the event that bias appears in model outcomes, this can help to infer whether the bias was present in the training data, and whether data preparation activities affected this. This becomes key in the implementation of regulations such as the European GDPR legislation.

Safety

Safety means the robust operation of systems in real-life situations, without loss of data, privacy leaks, or compromising cyber-security. This covers a wide variety of concerns: from driverless cars which do not crash, to safeguarding of data against privacy attacks. For example, a machine learning system might leak confidential information if such is used within its training dataset, and an attacker might exploit design weaknesses or implementation bugs within a system to expose this.

3.3 AI Policy frameworks

Governments and organisations are already implementing AI policy frameworks which cover these trustworthy technology properties: some cover all topics comprehensively, others focus on a subset. These often use different terminology: ethics, justice, fairness, accountability, transparency, which do not map to the technical literature directly. In our paper (Toreini, 2020), we collected 32 such frameworks and outline how they cover the FEAS properties laid out above. While interest from regulators is strong, the question remains to what extent the implementation of FEAS technologies, which aim to enhance trustworthiness, actually also enhance service users' trust.

3.4 Fairness as a Service

Our project has identified a challenge in creating a service to verify the fairness of machine learning algorithms. Our project has prototyped some solutions, and we expect publications on this in the future. This is an exciting space which we are still actively uncovering.

3.5 Relevant publications

- Toreini, E., Aitken, M., Coopamootoo, K.P., Elliott, K., Zelaya, V.G., Missier, P., Ng, M. and van Moorsel, A., 2020. Technologies for Trustworthy Machine Learning: A Survey in a Socio-Technical Context. arXiv preprint arXiv:2007.08911. https://arxiv.org/abs/2007.08911
- Toreini, E., Aitken, M., Coopamootoo, K., Elliott, K., Zelaya, C.G. and Van Moorsel, A., 2020, January. The relationship between trust in AI and trustworthy machine learning technologies. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 272-283). https://doi.org/10.1145/3351095.3372834

FINTRUST

Newcastle University

Atom bank

The Alan Turing Institute

UKRI Engineering and Physical Sciences Research Council

BILL & MELINDA GATES foundation

2022

https://fintrustresearch.com/

Written & Typeset by Samantha Finnigan, with assistance of Aad van Moorsel, Karen Elliott, Jehana Copilah-Ali, and others.
With thanks to Atom Bank and Catharine Powers-Freeling for their collaboration and feedback.