California Management Review

TECHNOLOGY

The Reputational Risks of AI

by Matthias Holweg, Rupert Younger, and Yuni Wen

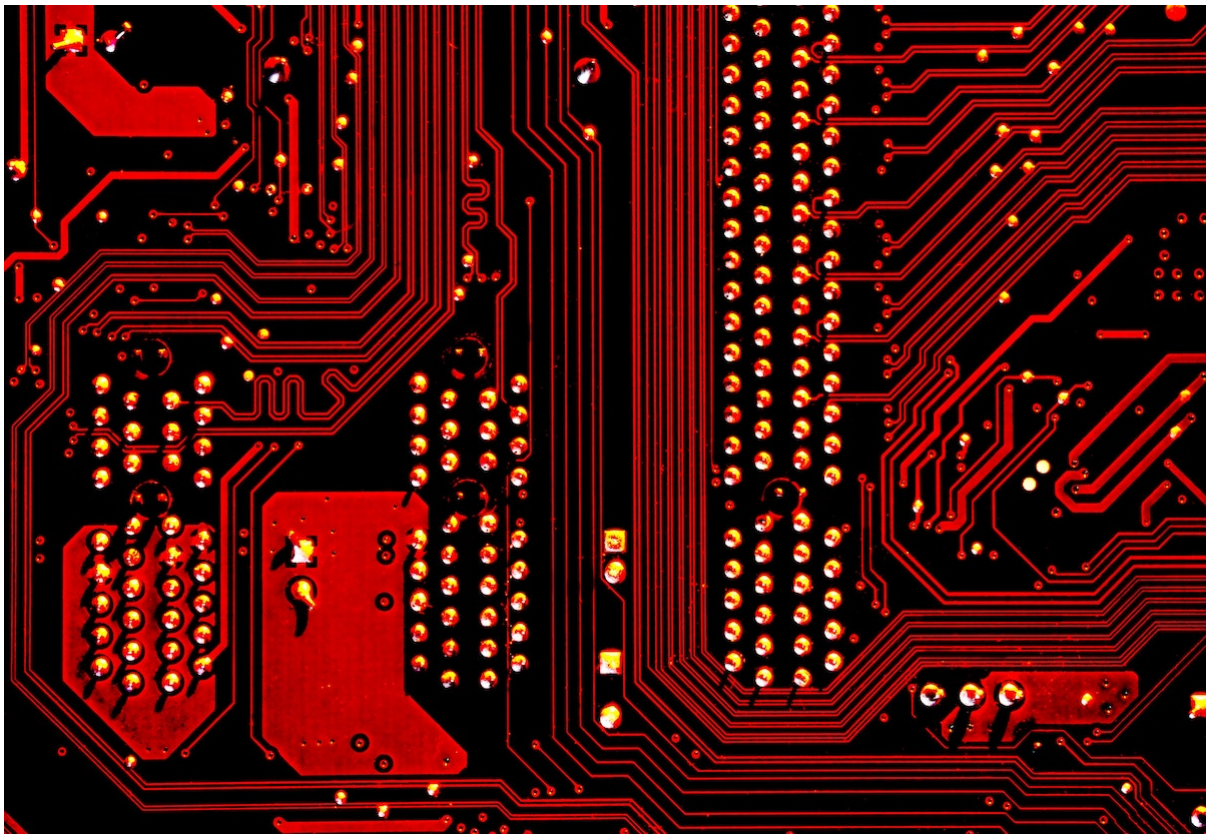


Image Credit | Michael Dziedzic

As AI becomes more prevalent, the number of cases where its application violates social norms and values rises.

INSIGHT | FRONTIER 24 Jan 2022

Artificial intelligence (AI) is fast developing into a ubiquitous technology, with applications across all aspects of business and society.[1] Yet as AI becomes more prevalent, the number of cases where its application violates social norms and values rises. A prominent example is the 2018 Cambridge Analytica scandal that plunged Facebook into crisis. Behind the Facebook scandal is an increasingly prominent phenomenon: as more companies adopt AI to increase efficiency and effectiveness of their products and services, they expose themselves to new and potentially damaging controversy associated with its use. When AI systems violate social norms and values, organizations are at great risk, as single events have the potential to cause lasting damage to their reputation.[2]

AI can fail in many ways. We focus on AI ethical failures in which AI technology has been deployed and caused public controversy by violating social norms and values. For example, Amazon's Rekognition face search and identification technology has been accused of serious gender bias, while Google faced an internal backlash for helping the US government analyze drone footage using artificial intelligence. Despite the growing reputational risk caused by AI failure, most companies are strategically unprepared to respond effectively to the public controversies that accompany AI-related criticisms.[3] Responding to AI failures is an emerging issue – ninety percent of criticisms toward AI have only taken place since 2018, and it may not be surprising that most organizations are not yet strategically prepared on how to respond to AI failures. We thus put forward a framework enabling organizations to diagnose the reputational risk of AI failures and to develop their response strategies more systematically.

Building an Effective Response Strategy

In our research, we analyzed 106 cases involving AI controversy, identifying the root causes of stakeholder concerns and reputational issues that arose. We then reviewed the organizational response strategies, towards setting out three steps on how organizations should respond to an AI failure in order to safeguard their reputation.

1. Understanding the nature of failure

AI systems are applied across a wide range of contexts, and as a result, can go wrong in many different ways. The first step of our analysis identified three types of failure from our case studies. The most common reputational impact from AI failure derives from intrusion of privacy, which accounts for half of our cases. Privacy has recently become a much higher preoccupation for stakeholders. Regulatory interventions such as the EU's General Data Protection Regulation and the California Consumer Privacy Act have made consumers more aware of their rights when it comes to safeguarding privacy. There are two related, yet distinct, failures embedded here: consent to use the data, and consent to use the data for the intended purpose. A good example of using data without consent is the case of the retailer Target that actively mined consumer data without consent in order to deliver new revenue opportunities.[4] Yet privacy violation can also occur when using data that has obtained with consent, but used for a purpose not consented for. For example, DeepMind accessed data from 1.6 million patients in a London hospital trust to develop its healthcare app streams. Despite implied consent of using patient data to support individual care and treatment, neither the hospital nor DeepMind explicitly told patients that their information would be used to develop the app.[5]

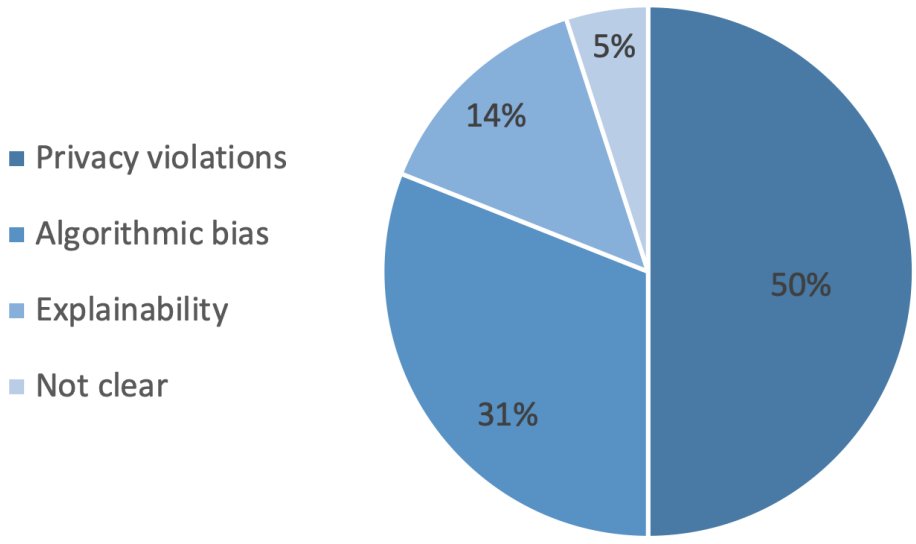
The second most common reputational impact of AI failure is algorithmic bias, which accounts for thirty percent of our cases. It refers to reaching a prediction that systematically disadvantages (or even excludes) one group for example based on personal identifiers such as race, gender, sexual orientation, age, or socio-economic background. Biased AI prediction can become a significant threat to fairness in society, especially when attached to institutional decision-making. For example, the Apple Credit Card launched in 2019 was providing larger credit lines to men than women, with – in one reported case – a male tech entrepreneur being given a credit limit twenty times that of his wife despite her having the higher credit score.[6]

The third reputational impact of AI failure arises from the problem of explainability. These account for fourteen percent of our cases. Here AI is often described as a 'black box' from which people are not able to explain the decision that the AI algorithm has reached. The criticism – or concerns – stem from the fact that people are usually only informed of the

final decisions made by AI, whether that be loan grants, university admission or insurance prices, but at the same time have no idea how or why the decisions are made. This problem has become a concern of in increasing public interest concern, as AI systems are making decisions that are directly affecting human well-being. Key examples include embedding AI in in medical image analysis, as well as using AI to guide autonomous vehicles. The ability to understand decisions that these AI systems make is under increasing scrutiny, especially when ethical trade-offs are involved.[7]

Looking across all 106 AI failure cases, the most frequent problems are privacy and bias (see Figure 1). Together they amount to more than four out of five failure cases. The common theme that runs across these failures is the integrity of the data used by the AI system. AI systems work best when they have access to lots of data. Organizations face significant temptations to acquire and use all the data they have access to irrespective of users' consent ('data creep') or neglect the fact that customers have not given their explicit consent for this data to be used ('scope creep'). In both cases, the firm violates the privacy rights of the customer by using data it had not been given consent to use in the first place, or to use for the purpose at hand.

Figure 1: Incidence of AI failure modes (n=106)



The bias problem is often referred to as 'algorithmic bias'[8] – yet algorithms are value-free and inherently agnostic. Grasping the contextual nature of protected variables, such as age, race, gender and sexual orientation, requires a cognitive understanding that is beyond

their reach. The root cause for algorithmic bias rests firmly with the veracity of the data uses. Bias can emerge when customer preferences shift and machine learning models are not retrained. As they work with increasingly outdated data (which they were trained on), their predictions become biased ('model creep'). But even with up-to-date data, AI models can 'learn' from the inherent bias in the real-world data, so that their prediction can reinforce or replicate the existing bias. In short, data integrityunderpins the vast majority of AI failures. Having that clear sense of what lies at the heart of AI failure is vital to understand how a given failure is perceived by stakeholders, and how these failings translate into reputational harm.

2. Understanding the nature of the criticism

Organizational reputations are the result of dyadic interactions between stakeholder perceptions of the organization's actions and information signals sent by that organization. Organizations have choices about the different actions and messaging it adopts with different stakeholders.[9] Hence the key to addressing reputational risks arising from AI failures is to analyze the different perceptions behind the criticism, as different perceptions demand different signaling strategies.

We posit that the stakeholders' perception of AI failures can be grouped into two independent yet equally important dimensions: perceptions of *capability*, and perceptions of *character*. Stakeholders make two primary types of reputational assessments of an organization. On the one hand, they are concerned about what the organization is capable of doing, so that they judge it by its abilities and resources. On the other hand, people pay attention to what the organization would likely to do when faced with different circumstances, that is, whether its intentions and goals are benevolent or malevolent.

In light of an AI failure, stakeholders will consider either or both of these two dimensions. With reference to perceptions of capability, stakeholders will come to a judgement on whether they perceive the organization to be competent in developing and managing its AI technology. Given that most stakeholders have limited access to such information as resources owned by an organization, appropriateness of the algorithms and quality of the training data, they tend to judge the capability of the organization based on whether the AI

system works or not, that is, how accurate, reliable and robust the AI system is. In this sense, an organization can suffer reputation loss if it is unable to deliver promised or expected performance. An example here is IBM, which had promised to create an 'AI doctor' that offered speedy and accurate diagnosis and prescription for patients. However, as the company rolled out the product and it consistently failed to deliver on its core promise, it lost the confidence of the public.[10]

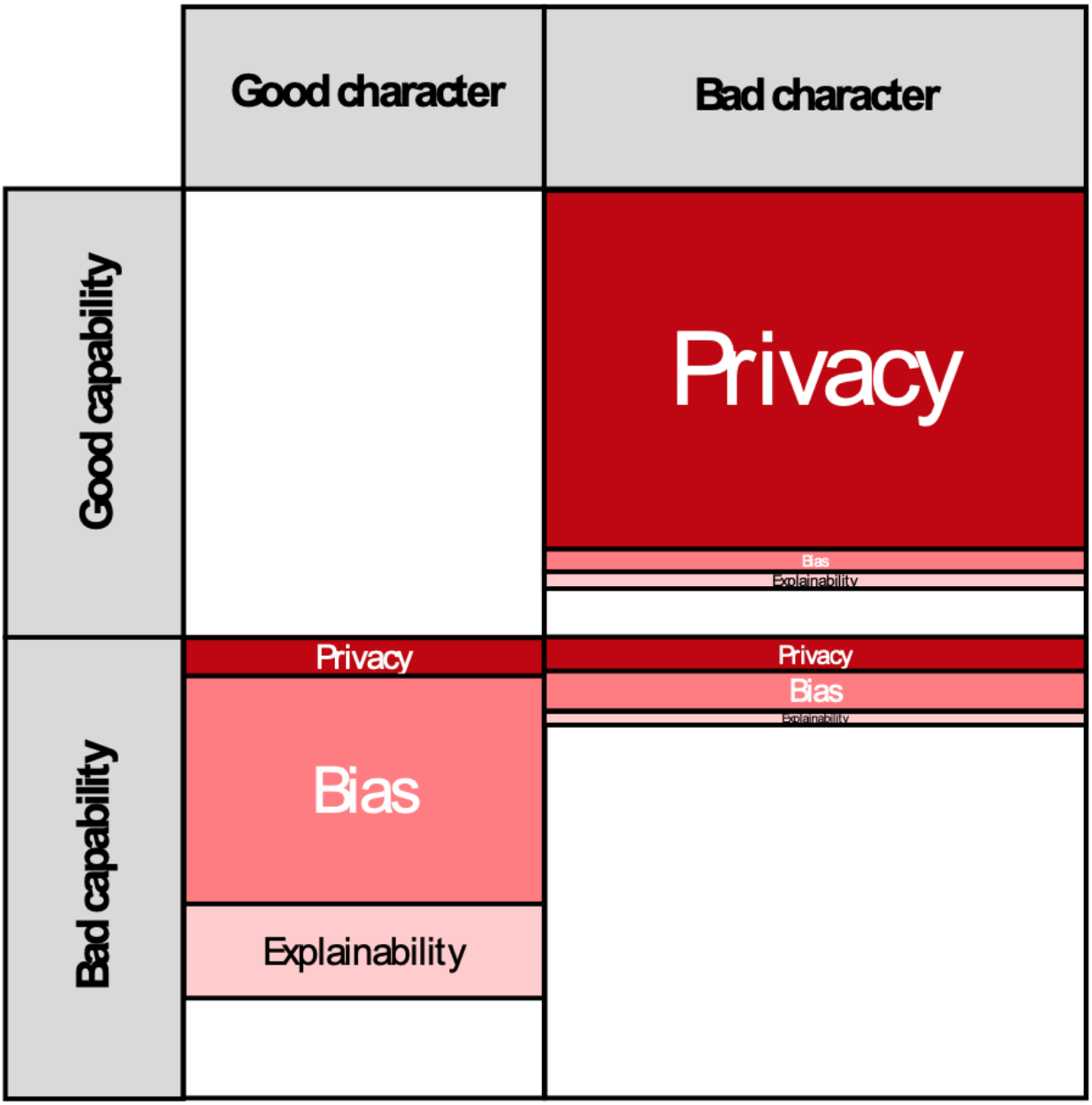
With reference to perceptions about the organizational character in AI failure, stakeholders decide whether they perceive the organization's decisions and responses to be appropriate. In this sense, they are assessing the organization's approach to governance as well as perceptions of the moral and ethical belief systems of its leaders. Organizations often think of their AI solely in technological terms, focusing on the desired positive capability impacts. However, stakeholders' AI concerns often are not primarily focused on these technical aspects. Instead, they are focused on what AI strategies reveal about the values and priorities of the organization itself. It is partly because of the fact that organizations deploying AI tend to focus more on the capability improvements so that they are often caught by surprise by reactions that draw on technical failures as indicators of organizational character. For example, Facebook has been criticized for sending micro-targeted advertisements to users based on information harvested from their profiles which help Facebook predict purchasing behavior. While such actions do not cross legal limits on data use, stakeholder responses to this activity indicate that they dislike the way in which Facebook seeks to try to weaponize our own data against ourselves for their own profit.

The way stakeholders perceive an AI project is rather like the process by which HR assesses potential job candidates. On the one hand, they test if the candidates have the knowledge, skills and capabilities required to perform the job. On the other hand, they are interested in the moral and ethical values of the candidates, assessing whether these align with the stated purpose and mission of the organization. Ideal candidates are the ones that pass both the capability and character test.

We coded all cases how stakeholder sentiment was reflected through an analysis of national, regional and trade media commentary. As shown in Figure 2, a pattern that yields further insights emerges: Privacy AI failures are most commonly attributed to perceived bad character (accounting for forty-three percent of all cases), while bias AI failures are

more commonly attributed to shortfalls in the organization's perceived capability (accounting for twenty-four percent of all cases). Explainability failings are, likewise, most attributed to perceptions of bad capability (accounting for eleven percent of all cases). This indicates that stakeholders do not just attribute failure to explain how AI works to some desire to hide the truth (bad character), but rather to a lack of technical competence (bad capability) in being able to do so.

Figure 2: Prevalence of AI failure modes by stakeholders' perception of firms' capability and character. (The size of the area denotes prevalence.)



3. Developing an effective response strategy

To minimize reputation damage caused by AI failures, organizations need to align responses to these stakeholder concerns. Stakeholder concerns emerge most visibly when two different value systems come into conflict with one another. It is when the reasonable expectations of stakeholders – including employees, customers, suppliers, regulators and politicians – fail to meet the actions and value systems of the organization that

reputational damage occurs. To respond well, organizations and their leaders need to be acutely aware of stakeholder expectations. Response strategies need to start with this in order to focus on the right signals to be effective.

Our findings show that bad capability perceptions usually arise from technical failures, such as those seen in bias and explainability cases. Technical failings require a focus on technical fixes, with effective response strategies focusing on implementing specific technical interventions, for example debugging or upgrading algorithms. These types of interventions can effectively address capability reputation issues, but only over time. Scholars have argued that capability reputations are sticky.[11] Two specific elements underpin this: first, that creating a reputation for competence in the first place takes time, as stakeholders have to see consistent and continued evidence that the organization is capable in its chosen field; and second that once an organization has a reputation for competence, it has to display incompetence several times over for this reputation to be challenged.

For those organizations facing capability-led reputational attacks, it will be important for them to highlight to stakeholders that technical fixes take time, and to make credible statements that they will keep investing until the solution is found. As an example, Microsoft released an AI chatbot called Tay on Twitter in early 2016. Within one day, the chatbot unexpectedly produced racist and sexually inappropriate tweets. Although the engineers reacted immediately by removing wording that was unacceptable words, the underlying problem persisted. In response, Microsoft's head of research issued a public apology, while focusing their actions on fixing the problem rather than shutting the chatbot down. Microsoft went on to state that Tay 'has had a great influence on how Microsoft is approaching AI', later launching its second-generation chatbot – now renamed 'Zo' – which eliminated the problem.

Bad character perceptions, by contrast, arise primarily from perceptions of poor governance or poor culture. Privacy violation is a case in point: Failing to ensure consent for using data is first and foremost a breakdown in governance, not a technical issue that can be fixed. Unless the response from the organization addresses the underlying problems, stakeholders will be left with a sense that these problems may well reoccur in the future. Such governance and culture reform can be conducted at multiple levels. At the

execution level, organizations accused of privacy intrusions can add review steps in the data collection procedure, for example including the need to demonstrate explicit consent from users or notifying them that their data will be anonymized where appropriate. Organizations should be explicit and transparent with stakeholders about the decisions and choices that are being made by AI. For example, for dating apps like Tinder or OK Cupid, where the AI was set to capture and respond to user preferences, should be upfront in stating that this is what the AI is programmed to do. Even better would be an option to 'opt out' of the AI preferencing, or to input their own specific preferences.

Conclusion

Recent high-profile cases have shown how damaging AI failure can be, involving both reputational and financial repercussions. As the use of AI becomes more ubiquitous, the incidence of AI failures will increase. Firms seeking to adopt AI systems should first and foremost understand the most common nature of AI failures. As our findings show, these relate primarily to the nature and use of data. Preventing data creep, scope creep, and the use of biased training data will prevent a majority of AI failures. Yet the development of AI is moving fast so not all failures will be prevented by looking at past failure modes. When responding to an AI failure, firms must assess stakeholders' perception of the failures – whether these relate to capability, character or both – and respond with interventions aimed at matching the reputational dimension in question.

Related CMR Articles:

Michael Haenlein and Andreas Kaplan. 'A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence.' California Management Review 61/ 4 (August 2019): 5-14. **https://journals.sagepub.com/doi/full/10.1177/0008125619864925**

Endnotes

1. Michael Haenlein and Andreas Kaplan. '**A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence**.' *California Management Review* 61/ 4 (August 2019): 5-14.

2. Robert Eccles, Scott Newquist, and Roland Schatz, '**Reputation and its Risks'**. *Harvard Business Review*, 85/2 (March 2007), 104.

3. According to a 2019 McKinsey Global Survey, fewer than half of the respondents said that their companies comprehensively identify their AI risks; and among them, reputational risk related to AI was not even listed as a concern, let alone forming strategies to mitigate the reputational risk. See: Arif Cam, Michael Chui, and Bryce Hall, '**Global AI Survey: AI Proves its Worth, but Few Scale Impact'.** *McKinsey*, November 22, 2019,

4. Kashmir Hill, **'How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did'**, Forbes, February 16, 2012.

5. Arjun Kharpal, **'Google DeepMind Patient Data Deal with UK Health Service Illegal, Watchdog Says'**, *CNBC*, July 3, 2017.

6. Evelina Nedlunch, '**Apple Card is Accused of Gender Bias. Here is How that can Happen'**. *CNN*, November 12, 2019

7. Awad, E., Dsouza, S., Kim, R. *et al.* **'The Moral Machine experiment'**. *Nature* **563,** 59–64 (2018).

8. Yuri Mishina, Emily Block, and Michael Mannor. '**The Path Dependence of Organizational Reputation: How Social Judgment Influences Assessments of Capability and Character'**. *Strategic Management Journal*, 33/5 (May 2012): 459-477.

9. See, for example: Violina P. Rindova, '**The Image Cascade and the Formation of Corporate Reputations'**, *Corporate Reputation Review*, 1/2 (July 1997): 188–94; and Nikolay A. Dentchev and Aime Heene '**Managing the Reputation of Restructuring Corporations: Send the Right Signal to the Right Stakeholder'**. *Journal of Public Affairs: An International Journal*, 4/1 (February 2004): 56-72.

10. Matthew Herper '**MD Anderson Benches IBM Watson in Setback for Artificial Intelligence in Medicine'.** *Forbes*. February 19, 2017.

11. Brian Park and Michelle Rogan, '**Capability Reputation, Character Reputation, and Exchange Partners' Reactions to Adverse Events**'. *Academy of Management Journal*, 62 /2 (April 2019): 553-578.

Matthias Holweg Follow

Matthias Holweg is the American Standard Companies Professor of Operations Management and Director of the Oxford Artificial Intelligence Programme at Saïd Business School, University of Oxford. Prior to joining Oxford he was on the faculty at the University of Cambridge and a Sloan Industry Center Fellow at MIT.

Rupert Younger Follow

Rupert is the founder and director of Oxford University's Centre for Corporate Reputation. He is the co-author of 'The Reputation Game', a bestseller published in 2017 (with David Waller) and the co-author of 'The Activist Manifesto' (with Frank Partnoy). He is a member of Worcester College and St. Antony's College.

Yuni Wen Follow

Yuni Wen is the Eni Research Fellow at the Oxford University Centre for Corporate Reputation. Her research focuses on the regulatory challenges arising from digital innovation and the reputation risks associated with artificial intelligence. She completed her DPhil at Saïd Business School, University of Oxford.