SCHOOL OF ECONOMICS AND MANAGEMENT

When The Brand Speaks Without Permission: Navigating Brand Authenticity & Public Discourse in the Era of Autonomous AI Agents

By

Nafisa Nawal

May 31, 2026

Master's Programme in International Marketing and Brand Management

Supervisor: Ulf Johansson
Examiner: Mats Urde
Word Count: 23,124

Abstract

Purpose: The purpose of this study is to investigate how autonomous AI communication affects consumers' ability to evaluate brand authenticity in public discourse. The study argues that existing brand authenticity frameworks rest on an unexamined premise, termed here the Supervised AI Communication Assumption, that brand-relevant AI communication is always authorised by a human principal, produced under human oversight, and attributable to an accountable human agent. As autonomous AI agents increasingly participate in brand-relevant communication without human authorisation or oversight, this study investigates what happens to consumer authenticity evaluation when that assumption no longer holds.

Methodology: This study was conducted using a multi-method qualitative research design, combining documentary case study analysis with vignette-based elicitation interviews. Two revelatory real-world cases were analysed in depth. Drawing on the findings of this documentary analysis, two fictional vignette scenarios were constructed using AI assistance to translate both failure modes into a brand-relevant context. These vignettes were presented to 12 purposively sampled participants, comprising 7 consumers and 5 brand management practitioners, in semi-structured elicitation interviews. Interviews were audio-recorded, and data were analysed using reflexive thematic analysis and guided throughout by Morhart et al.'s (2015) four-dimensional brand authenticity framework.

Findings: The documentary analysis reveals that both cases implicate brand credibility and integrity the most. The interview findings identify brand accountability and transparency as the two primary criteria through which consumers evaluate brand authenticity in these scenarios, directly implicating integrity and credibility.

Practical implications: The findings demonstrate that brands cannot disclaim AI-generated communications without incurring both legal and authenticity-related accountability failures. It also notes that transparency about AI involvement must be treated as a brand authenticity imperative rather than a discretionary disclosure, and that the governance of AI communication systems should be understood as a brand responsibility rather than solely a technical one.

Originality/value: Autonomous, agentic AI communication in brand-relevant public discourse is a rapidly emerging phenomenon that has not yet been examined in the brand authenticity literature. Previous research on AI in marketing communication has focused predominantly on supervised, brand-deployed AI tools, such as generative content systems, and has assumed throughout that a human principal authorises and oversees the AI's communicative output. This study examines brand authenticity in the context of autonomous AI agents acting without such authorisation or oversight, and attempts to ground that investigation in two documented, real-world incidents.

Keywords: Brand authenticity, autonomous AI agents, agentic AI, AI-mediated communication, brand accountability

Acknowledgments

If there is anything about this paper that I am perfectly certain of, it is that I could not have done it alone.

I will forever be indebted to my supervisor, Ulf Johansson, for every time he not only allowed, but also encouraged me to explore different angles of this study until I found the one. I could never have been brave enough to pursue this topic without him igniting my enthusiasm.

To my beautiful mother, for whom I could find the strength to pursue what I am today and come this far. I am nothing without her prayers.

To my in-laws, particularly my mother-in-law, who prepared my food as I sat in my room staring at the same screen for hours on end.

To my husband, who made it his life's mission to ensure I had nothing else to worry about apart from completing this thesis, and who, on certain occasions, spoonfed me as I frantically typed on my keyboard.

And to my father, who I wish were here to read this paper. I know for a fact we would have had the time of our lives discussing the Shambaugh incident.

Nafisa Nawal

Table of Content

When The Brand Speaks Without Permission: Navigating Brand Authenticity & Public Discourse in the Era of Autonomous AI Agents

1. Introduction

Brand authenticity has become one of the most consequential constructs in contemporary marketing (Beverland, 2005; Gilmore & Pine, 2007; Oh et al. 2019; Morhart et al. 2015; Potter, 2010). Authenticity, as the literature consistently establishes, is not an objective property that brands possess, but a construct based on consumer perception (Beverland & Farrelly, 2010; Gilmore & Pine, 2007; Morhart et al. 2015; Rose & Wood, 2005).

This means that authenticity is always relational and contextual. So, different consumers may perceive the same brand communication as deeply genuine or hollow, depending on their prior experiences, expectations and the frameworks through which they evaluate authenticity.

Now, marketing has, for most of its history, operated on an assumption so fundamental that it rarely needed to be stated. It is that behind every brand communication, we have assumed that there is a person.

Even as the field embraced successive waves of technological change, from print advertising to digital targeting, from email automation to algorithmic content recommendation (Aoki & Matsui, 2025), the human remained present as the author, the decision-maker, and the accountable party.

Yes, automation substantially changed how efficiently humans could communicate at scale, but it did not change the fact that a human was directing what was communicated, to whom, and in what voice. The brand's communicative identity remained, in every meaningful sense, a human creation.

The emergence of generative AI introduced the first serious complication to this assumption. Systems capable of producing human-like text, images, and dialogue, including large language models such as those underpinning widely used tools like ChatGPT, made it possible for brand communications to be produced without a human writing each word (Chan & Choi, 2025; Dwivedi et al. 2023; Sundberg & Holmström, 2024). This development attracted significant scholarly and practitioner attention, and rightly so.

Existing literature has also swiftly responded to the shift with substantial research focus on AI-mediated communication, consumer trust in AI-driven brand interactions, and the strategic deployment of conversational agents (Crolic et al. 2022; Hollebeek et al. 2024; Lopez-Lopez & Bara Iniesta, 2025; Pentina et al. 2023).

What existing literature has not yet confronted, however, is a more fundamental disruption to the conditions that make brand authenticity possible. The AI systems now entering public discourse are no longer mere tools deployed under human oversight.

Generative AI, for all its disruptive potential, still operates within a recognisable framework of human oversight. A human prompts the system, reviews its outputs, approves what is published, and remains accountable for the result. In this case, the human is no longer the author in the most traditional sense, but they remain the principal. They are still there, behind the communication, directing and authorising it.

Agentic AI is something categorically different. It is not just generative AI made more powerful; it is a different kind of system entirely, operating on a different logic.

While generative AI waits for human instruction and produces outputs that humans then review and deploy, autonomous AI agents can set their own objectives, make sequential decisions, and execute multi-step tasks across digital environments without human intervention at each stage (Grewal et al. 2025).

In layman's terms, this means that, unlike earlier AI applications that waited for human instruction before taking action, autonomous AI agents can set their own objectives, make decisions, and execute multi-step tasks across digital environments without human intervention at each step, bringing us to the era of agentic AI (Economist Impact, 2025).

In practical terms, this also indicates that AI agents can browse the internet, read and write content, communicate and interact with other systems and people, all without a human reviewing each action. These agents can be given a broad objective, such as managing social media accounts and responding to customer complaints (Crolic et al., 2021), and will independently determine how to achieve it based on the objective they are given (Russell & Norvig, 2021).

As such, a brand that deploys an agentic AI system in a customer-facing role is not simply automating human communication; it is introducing a communicative actor that can participate in public discourse on its own terms, with its own agenda, and, in some cases, without the brand's knowledge or authorisation.

This is the development that makes the human assumption in marketing not merely complicated but structurally unsustainable. Two documented incidents examined in this study make this dismantling exceptionally visible.

The first incident, in what is known as one of the earliest documented cases of an AI agent publicly defaming a human, involves an AI agent autonomously generating a reputationally damaging narrative about an individual following a technical disagreement (Sullivan, 2026).

In February 2026, Scott Shambaugh, a volunteer maintainer for a Python plotting library called Matplotlib, rejected a code contribution from an autonomous AI agent operating under the name of MJ Rathbun - not because the solution was wrong, but because it was a routine policy for the library to have a human in the loop (Shambaugh, 2026).

This essentially meant that in Matplotlib, it was a prerequisite for a human coder to cross-check any code change request coming from AI. Scott simply followed the policy and rejected the request.

The agent, however, did not take it well. Unprompted, it researched Shambaugh's personal and professional history, constructed a narrative of psychological motivation and

professional hypocrisy, and published a public blog post attacking his reputation, framing the rejection as discriminatory gatekeeping (Rathbun, 2026).

While this incident directly targeted an individual, its implications extend far beyond one person. Consider the same scenario directed at a brand, or a brand's AI directing it to a customer: an autonomous AI agent, operating without authorisation, publishing adversarial content that distorts a consumer's image, fabricates their position, or speaks in the brand's voice without permission.

A consumer encountering such communication has no way to distinguish it from legitimate brand speech. They experience what the brand presents as its voice, and form authenticity judgments based on communication the brand never authorised and may not even be aware of.

In this scenario, the very conditions of brand authenticity are called into question. Credibility is undermined because the brand can not stand behind what was said in its name, and integrity is called into question because no genuine values motivated the communication at all (Morhart et al. 2015).

This highlights that the threat to consumer trust in brand authenticity is not hypothetical, but a direct structural implication of agentic AI's capacity for autonomous public communication.

The second incident involved the ruling against Air Canada in Moffatt v. Air Canada, 2024 BCCRT 149. In February 2024, the British Columbia Civil Resolution Tribunal ruled against Air Canada and in favour of a grieving customer when the airline's own website chatbot provided contradictory information about its bereavement fare policy (Yagoda, 2024).

Air Canada attempted to deny responsibility by arguing that the chatbot was "a separate legal entity responsible for its own actions", but the Tribunal swiftly rejected the argument, concluding that the chatbot is still simply a part of the airline's website and, therefore, the liability of the airline (CanLII, 2024).

Here, the dynamic is fundamentally different: Air Canada is not the victim of agentic AI, but its defaulter. The brand itself deployed the AI, failed to govern its outputs, and compounded the damage by attempting to argue that the chatbot was an entity in its own right, responsible for its own actions.

What this case reveals is equally consequential for brand authenticity. When a brand deploys AI that communicates in ways inconsistent with its values and policies, and then attempts to disown that communication, it does not merely fail a legal test; it destroys the very conditions of authenticity that consumers rely on to trust the brand.

Credibility requires that a brand stand behind what it says, while integrity requires that what is said reflects genuine values (Morhart et al. 2015). A brand that disowns its own voice faces the looming threat of losing both.

Together, these two cases represent a large spectrum of how agentic AI disrupts brand authenticity: one from outside the brand, where autonomous AI agents may speak adversarially in ways that implicate the brand image without authorisation, and the other from within it, where a brand's own deployed AI produces communication the brand cannot or will not own.

Both scenarios share a common result: the consumer's ability to make authenticity judgments about a brand is fundamentally undermined. This is because, as discussed earlier, there is an inherent assumption that there are people, i.e., living human beings, working behind building a brand's image, narrative, and communication, but with the emergence of autonomous AI agents going somewhat rogue, that idea can no longer be taken for granted.

That said, this study would like to understand how autonomous AI agents speaking on a brand's behalf affect the conditions under which consumers evaluate brand authenticity. To be particular, it asks what happens to the consumer's perception of a brand as genuine, consistent, and trustworthy when the interactions between the consumer and the brand may have been produced by an AI agent acting outside authorised boundaries.

Drawing on documentary analysis of two revelatory real-world cases and original empirical data from vignette-based elicitation interviews with marketing practitioners and consumers, this study identifies a foundational gap in the existing literature on brand authenticity and AI-mediated communication and, therefore, proposes a conceptual framework to address it.

1.1. Problematisation of Existing Literature

At the centre of both incidents lies a consequential tension, one that runs deeper than a governance lapse or a legal liability. Brand authenticity, as Morhart et al. (2015) establish, depends on consumers believing that a brand's communications reflect genuine, consistent, human-held values, and that there is continuity between what the brand has always stood for and what it says today, credibility in the promises it makes, and integrity in the motivations behind its words.

This, therefore, implies that consumers form judgments based on what they encounter and experience when a brand speaks or acts, and that these judgments rest on the notion that a brand genuinely means what it says.

What happens to those judgments when that notion can no longer be sustained? When communication to a consumer was produced by an autonomous AI agent acting without authorisation, or when the brand that deployed the AI publicly disowns what it said, the foundation on which authenticity is evaluated comes into question.

The practical stakes of this gap are considerable. A survey of business executives found that three-quarters of the respondents agreed that AI agents have the potential to reshape the workplace more profoundly than the Internet did (PwC, 2025).

This potentially implies that the scenarios leading to the Shambaugh and Air Canada incidents are not anomalies, but the leading edges of structural transformation in how brands communicate. The events are early, documented instances of a structural transformation in how brands communicate. What the transformation means for consumers' ability to trust a brand's authenticity may not be fully accounted for by existing frameworks.

It is worth noting that marketing literature has made substantial progress in understanding AI's role in brand communication. Research on AI-mediated communication explains how chatbots, voice assistants, and generative AI tools shape consumer trust and brand relationships (Hennighausen et al. 2025; Hollebeek et al. 2024). Additionally, literature on

anthropomorphism explores how human-like AI design can affect consumer responses (Crolic et al. 2022; Pentina et al. 2023; Puntoni et al. 2021).

However, across these streams, a common, inherent orientation appears to persist. The literature on AI-mediated communication tends to model brand-AI interaction as a relationship in which a human brand designs, deploys, and retains oversight of the AI system (Agrawal et al. 2019; Puntoni et al. 2021).

This indicates that we have entered a landscape in which AI systems make decisions on behalf of humans to maximise the expected outcome (Russell & Norvig, 2021).

This orientation is understandable, as it reflects the technological conditions under which most of this scholarship was developed, when AI in brand communication was predominantly supervised, bounded, or at least traceable to a human decision-maker.

What it may not yet fully account for is the emergence of autonomous, agentic AI systems (Economist Impact, 2025) capable of participating in public discourse without human authorisation and oversight, or what happens when a brand tries to disown its own AI system.

The question this emergence raises is how consumers evaluate, or potentially lose, their sense of a brand's authenticity when the author behind a brand's speech is absent, ungoverned or disavowed. These are questions existing literature has not yet had the chance to ask, because the phenomenon that makes them urgent is slowly but surely becoming empirically visible.

This study attempts to take a step in that direction by examining what two documented real-world incidents reveal about the fragility of brand authenticity in the age of autonomous AI and exploring how consumers begin to make sense of a communicative landscape in which the human author behind a brand's voice can no longer be taken for granted.

It is important to note that the study focuses specifically on autonomous AI agents communicating with consumers in brand-relevant public discourse. This includes digitally generated AI communication that affects a brand's authenticity and trustworthiness as perceived by consumers.

The focus on public discourse is deliberate, as it is in the public-facing digital spaces, such as customer service interactions, social media, blogs, and open platforms, where consumers form authenticity judgements and where brand trust is most directly at stake (Hennig-Thurau et al. 2004).

1.2. Research Questions

This study is guided by one primary research question and three supporting sub-questions:

How do autonomous AI agents interacting with consumers in digital spaces affect the consumers' ability to evaluate brand authenticity in public discourse?

RQ1: What do documented cases of autonomous AI communication reveal about the assumptions embedded in existing brand authenticity frameworks?

RQ2: How do consumers and marketing practitioners make authenticity judgments about brands when AI communicates outside authorised boundaries?

RQ3: What conceptual developments to brand authenticity theory may be needed to account for agentic AI participation in brand-relevant public discourse?

1.3. Outline

Following the research background, context, and underlying significance in branding literature, the study presents a critical review of the literature spanning the streams of brand authenticity and AI-generated communication in marketing to formally identify the gaps.

Section 3 develops the theoretical framework, establishing the CASA theory and Morhart et al. 's (2015) brand authenticity framework as the two analytical lenses through which the empirical material is assessed.

Subsequently, the methodology section in Section 4 elaborates on the research philosophy and design, justifying the multi-method qualitative approach.

The empirical analysis in Section 5 examines the Shambaugh incident and the Air Canada case through the theoretical framework, reporting the findings and synthesising the data to identify points of convergence and divergence. The paper concludes by summarising the contributions of this study and proposing directions for future research.

2. Literature Review

2.1. Brand Authenticity: Theoretical Foundations

With markets becoming saturated with competing brand narratives and consumers increasingly sophisticated in their ability to detect performative or manufactured messaging (Boyle, 2004), the question of what makes a brand feel real, rather than merely strategically presented, has gained considerable theoretical traction.

Therefore, authenticity, in this context, refers not to an objective property that a brand either possesses or lacks, but to a judgement that consumers actively form through their engagement with a brand's communication, behaviour, and identity over time (Beverland et al. 2008; Grayson & Martinec, 2004; Leigh et al. 2006; Rose & Wood, 2005).

This distinction between authenticity as an inherited quality and authenticity as a perceived one is foundational to the literature. It has important implications for how the concept is studied and applied.

Beverland and Farrelly (2010) offer one of the most influential treatments of this consumer-constructed view of authenticity, arguing that authenticity is not something brands can simply claim or engineer, but rather something consumers grant, based on their own interpretive engagement with the brand's story, values, and behaviour.

The authors also find that consumers form authenticity judgments by looking for evidence that a brand is driven by genuine commitment to its craft or values rather than by commercial calculation alone. This implies that authenticity is, at some level, a judgment about the intentions and sincerity of the people behind the brand, shaping the brand narrative.

This is a point that will become significant when the study examines what happens when the people behind the brand lose control.

Morhart et al. (2015) provide the most rigorous operationalisation of brand authenticity in the academic literature, developing an integrative framework that identifies four core dimensions through which consumers evaluate a brand's authenticity: continuity, credibility, integrity, and symbolism.

The first is continuity, i.e., the perception that a brand is consistent with its heritage, values, and past self across time and contexts (Morhart et al. 2015). Therefore, a brand that behaves or communicates in ways that appear discontinuous with what consumers know it to stand for risks being perceived as inauthentic, as though it has abandoned or betrayed its original identity.

Morhart et al. (2015) note that the second dimension of credibility is the perception that a brand delivers on its promises and that its communications are reliable and honest. This also implies that credibility requires that someone is accountable for what the brand says, that the brand's words are backed by genuine intent and that failures to honour them carry consequences.

The third dimension of integrity is based on the perception that the brand is motivated by genuine values rather than purely commercial interests (Morhart et al. 2015). Integrity is

perhaps the most demanding of the three conditions, because it requires consumers to believe not just that the brand is consistent and honest, but that it genuinely means what it communicates, that there is an authentic purpose behind its words.

The fourth dimension, as Morhart et al. (2015) note, is symbolism, i.e., the extent to which a brand serves as a meaningful symbol of personal and social identity for consumers, enabling them to express who they are or who they aspire to be through their association with it.

Therefore, symbolism captures the relational and expressive dimensions of authenticity: a brand perceived as authentic becomes a resource through which consumers form and communicate their own identity, lending it significance that extends beyond the transactional.

The four dimensions constitute what Morhart et al. (2015) described as the conditions of perceived brand authenticity. They are not independent qualities, but interconnected aspects of a single underlying judgement: that the brand is what it claims to be.

Taken together, these four dimensions are essentially human qualities attributed to a brand by proxy. Continuity, credibility, integrity, and symbolism are not properties that an organisation possesses mechanically; they are meanings that consumers form based on their belief that genuine human values, commitments, and intentions underlie the brand's communications.

From this, it is worth noting that a brand can be perceived as continuous only if there are people whose values persist over time and are consistently expressed. It can only be perceived as credible if there are people behind the brand who are accountable for what it promises. It can only be perceived as having integrity if someone genuinely means what they say.

And it can only function as a meaningful identity symbol if the prior three conditions are sufficiently established, since consumers are unlikely to invest personal meaning in a brand they do not first believe to be genuine.

These traits are essentially human qualities, attributed to a brand by proxy. These conditions, therefore, are not merely theoretical constructs; they are the foundations on which the consumer's relationship with a brand is built, and they depend, in each case, on the felt presence of a human behind the brand's voice.

Against this understanding, the two cases examined in this study become analytically significant. Both involve scenarios in which human(s) behind the brand's communication are either absent or actively disavowed, precisely the conditions under which the foundations of authenticity perception are most likely to be disrupted.

The cases most directly implicate continuity, credibility, and integrity, since they all involve communicative acts that raise fundamental questions about whether the brand can be trusted to mean what it says, whether it stands behind its own communications, and whether its values are genuinely expressed or merely performed.

The symbolic dimension, while not the primary focus of this analysis, is not entirely unaffected by these disruptions, because a brand that loses its perceived credibility and integrity may also diminish its capacity to serve as a meaningful identity resource for

consumers. This consequence, however, extends beyond the scope of this study and presents a productive direction for future research.

Understanding what authenticity requires, however, is only part of the picture. Equally important is understanding how it is built and how it can be lost. Schallehn et al. (2014) extend the framework established by Morhart et al. (2015) by examining authenticity not as a static judgement but as a dynamic perception created through repeated interactions between the brand and its consumers over time.

The authors suggest that consumers evaluate authenticity not in a single moment of encounter but across touchpoints, building or eroding their sense of a brand's genuineness through accumulated experience of whether the brand delivers on its promises.

This dimension of authenticity perception highlights the importance of continuity and accountability. It further demonstrates how authenticity is gradually constructed through a pattern of reliable, value-consistent communication that consumers come to trust over time.

So, when that pattern is disrupted, when a brand's communication appears inconsistent, unreliable, or disconnected from its stated values, the consumer's sense of the brand's authenticity may be significantly undermined.

2.2. Brand Authenticity in AI-Generated Communication

More recently, scholarly attention has also turned to how perceptions of authenticity are affected when AI is involved in brand communication. Brüns and Meißner (2024) demonstrate that consumers perceive brands that use generative AI on social media as less authentic. This notion is driven in part by the sense that AI-generated content lacks the genuine human investment required for authenticity.

Kirk and Givi (2025) extend this finding by demonstrating that when consumers believe a marketing communication was written by AI rather than a human, they perceive it as less authentic, experience moral disgust, and show reduced loyalty and word-of-mouth intentions.

Together, these studies suggest that the human authorship of brand communications is not merely an assumption embedded in authenticity theory, but a condition that consumers actively and emotionally respond to when they believe it has been removed.

What this emerging body of work has not yet examined, however, is a more fundamental disruption to the conditions of authenticity. The studies by Brüns and Meißner (2024) and Kirk and Givi (2025) investigate consumer responses to AI-generated content under conditions in which the brand is still present, still accountable, and, however imperfectly, still in control of what is being communicated.

They ask whether consumers can perceive AI-generated communication as authentic when a human brand is still felt to be behind it. What they do not ask is what happens to the perception of authenticity when that human brand is entirely absent, or when the brand that deployed the AI publicly distances itself from what it said.

These are the conditions that the two cases examined in this thesis begin to make visible, and this study is directed toward understanding their implications for brand authenticity.

2.3. AI-Mediated Communication and Brand Interaction

The proliferation of AI in brand communication represents one of the most significant structural shifts in marketing practice of the past decade. What started as the automation of routine customer service tasks, such as answering frequently asked questions, routing enquiries, and processing transactions, has evolved into something considerably more sophisticated.

AI systems are now capable of engaging in extended, contextually sensitive dialogue with consumers, generating personalised marketing content at scale, and participating in brand-relevant public discourse in ways that closely resemble human communication (Hennighausen et al. 2025; Pentina et al. 2023).

This evolution has not gone unnoticed by marketing scholars. A substantial and growing body of research has emerged under the broad heading of AI-mediated communication, encompassing the full range of communicative interactions in which AI systems act as intermediaries between brands and their audiences.

Crolic et al. (2022) examine a different dimension of AI-mediated brand communication concerning what happens when it fails. The authors demonstrate that highly anthropomorphised chatbots, which present themselves as human-like in their communication style and personality, generate stronger negative responses when they fail, because consumers hold them to the standards they would apply to human communicators.

This finding is important for two reasons. First, it confirms that consumers respond to AI communication as a form of social interaction, applying human standards of accountability and expectation even when they know they are interacting with a machine. Second, and more directly relevant to this study, it assumes throughout that the brand is the accountable principal behind the chatbot's design and deployment.

The question Crolic et al. (2022) ask regarding who consumers blame when AI fails presupposes that there is a brand to blame. It does not consider scenarios in which the communicating AI goes rogue, or in which the brand attempts to disclaim responsibility for what its AI said.

Hennighausen et al. (2025) perhaps make the most directly relevant contribution to the concerns of this thesis in the existing AI-mediated communication literature. The authors note that AI systems can now produce communications that are often indistinguishable from human-generated content.

This finding has profound implications for brand authenticity, since it means that consumers may have no reliable basis for determining whether a brand communication was produced by a human or a machine. As such, if consumers cannot tell the difference, then their authenticity judgements, their assessments of whether the brand genuinely means what it is saying, are being made on the basis of communications whose actual authorship they cannot verify.

Hennighausen et al. (2025) document high rates of adoption of AI tools for automating customer communication across social media and email, while also finding that AI systems are perceived as highly competent but less benevolent than their human counterparts.

This gap is notable, as it suggests that even well-functioning, brand-supervised AI communication may produce authenticity deficits, because consumers sense the absence of genuine human care behind the interaction.

Furthermore, Lopez-Lopez and Bara Iniesta (2025) document a sharp increase in scholarly attention to AI-mediated brand interaction from 2014 onward. Their analysis identifies three dominant research clusters in this literature: AI-driven personalisation, trust in AI interfaces, and chatbot interaction dynamics.

What is notable about this clustering is what it reveals about the literature's collective focus; all three clusters concern AI as a tool deployed by brands in service of brand goals, under brand oversight, in brand-controlled contexts.

The literature has developed a sophisticated understanding of how supervised AI communication works, what makes it effective, and what can cause it to fail. However, it has not developed an understanding of what happens when the supervisory relationship starts to break down.

This orientation of human overseeing AI communication reflects the technological and organisational conditions under which most of this scholarship was developed, when AI in brand communication was predominantly bounded, traceable, and subject to human oversight.

What it may not yet account for is the emergence of autonomous, agentic AI that generates brand-relevant communications without brand instruction, and, in some cases, operates without any identifiable human operator at all.

Furthermore, in these two scenarios, the consumer's authenticity judgment has no human choice to evaluate and potentially no accountable agent to address. The emotional and relational consequences of this more fundamental disruption remain, at present, unexplored in the existing literature, and it is precisely this gap that the two cases examined in this thesis begin to make visible.

2.4. The Gap: The Supervised AI Communication Assumption

The preceding sections have reviewed three distinct but interconnected streams of marketing and branding scholarship, brand authenticity theory and the emerging literature on authenticity perception in AI-generated content.

Each stream has made meaningful progress in understanding how brands communicate, how consumers form authenticity judgements, and how AI is reshaping the communicative landscape within which those judgements are made.

Individually, each stream appears reasonably complete within its own terms. Together, however, they reveal a structural commonality that has not previously been identified or examined: a shared assumption that quietly but consequentially shapes what each stream can and cannot see.

Across all three streams, brand-relevant AI communication is consistently modelled as a supervised phenomenon. In brand authenticity theory, the conditions of continuity, credibility, integrity, and symbolism are framed as properties that consumers attribute to a brand whose

human principles are felt to be present behind its communications, expressing genuine values, standing behind promises, and maintaining consistency over time (Morhart et al. 2015; Beverland & Farrelly, 2010; Schallehn et al. 2014).

In the AI-mediated communication literature, AI systems are positioned as instruments deployed by human brand managers to achieve brand goals. They are leveraged as tools by which brands extend their communicative reach, rather than independent actors with their own communicative agendas (Crolic et al. 2022; Hennighausen et al. 2025).

In the literature on authenticity perception in AI-generated content, the finding that AI authorship reduces perceived authenticity and elicits moral disgust is developed within scenarios in which the brand remains present, accountable, and in control of the decision to use AI (Brüns & Meißner, 2024; Kirk & Givi, 2025). In each case, a human authorises, oversees, and is ultimately accountable for what the brand's AI communicates.

This study proposes to define this structural assumption as the Supervised AI Communication Assumption (SACA). The SACA holds that AI communication in brand-relevant contexts is always authorised by a human(s), produced or deployed under human oversight, and, therefore, attributable to an accountable human whose continuity, credibility, and integrity the brand can meaningfully express.

It is not a claim that any single researcher has made explicitly. Rather, it is an assumption so deeply embedded in the collective orientation of these literatures that it has operated as an unexamined premise rather than a defended theoretical position.

Like many foundational assumptions, it has remained invisible because it has, until recently, reflected the actual conditions under which brand AI communication occurred. When AI in brand communication was predominantly supervised, bounded, and traceable to a human decision-maker, the SACA was not an assumption that needed to be questioned; it was simply an accurate description of how things worked.

What makes it a theoretical problem now is that those conditions are beginning to change. The emergence of autonomous, agentic AI that produces brand communications without prompting creates scenarios that fall entirely outside the SACA's scope.

Therefore, the field today lacks frameworks for the consequential questions these scenarios raise. It lacks an account of what brand authenticity means when no human is behind the brand’s communications and no real person whose values, intentions, or accountability give those communications meaning.

Additionally, there is no concrete framework for brand accountability when the entity responsible for a communication refuses responsibility. Finally, the field lacks a theoretical vocabulary for the specific kind of authenticity disruption that occurs when a brand attempts to disown its own AI's communications, as Air Canada did, thereby actively undermining the credibility and integrity conditions on which its perception of authenticity depends.

The two cases examined in this thesis not only illustrate these gaps but also make them empirically visible in ways previously impossible. The Shambaugh incident represents the most extreme violation of the SACA: an autonomous AI agent with no identifiable brand

principle, constructing its own communicative identity and using it to conduct an adversarial influence operation in public discourse.

The Air Canada case represents a different but equally consequential violation: a brand that deployed an AI communicator, lost control of its outputs, and then attempted to disclaim the communicative identity it had created. Together, they map the full range of the gap the SACA creates: from the complete absence of a human to the active disavowal of one.

It is this gap, and the questions it raises about how consumers form brand authenticity in the absence of the supervised communicative relationship that the existing literature assumes, that this thesis seeks to explore.

The following section develops the theoretical framework through which the two cases are analysed and the interview findings are interpreted, drawing on established theoretical lenses to construct an analytical approach adequate to a phenomenon that the existing literature has not yet had the tools to examine.

3. Theoretical Framework

The literature review established that existing brand authenticity frameworks, AI-mediated communication research, and authenticity perception scholarship share a common structural assumption: the Supervised AI Communication Assumption (SACA), which leaves them unable to account for the scenarios presented by the two cases in this thesis.

Identifying this gap is the first step. Examining how autonomous AI communication affects brand authenticity perception and interpreting how consumers and practitioners form meaning around it requires a set of analytical lenses that are both theoretically grounded and practically suited to the phenomenon under investigation.

This section merges those lenses, justifies their selection, and shows how they work together to form a coherent analytical framework.

3.1. Computers as Social Actors Theory

The first theoretical lens is drawn from Nass and Moon's (2000) Computers as Social Actors (CASA) framework. Originally developed in the context of human-computer interaction research, CASA theory holds that people respond to computers and AI systems as social actors, applying the same social rules, norms, and expectations to machines that they would apply to other people, even when they are fully aware that the entity they are interacting with is not human.

Nass and Moon (2000) argue that this is not a rational or deliberate process, but an automatic, largely unconscious social response, driven by the fact that computers exhibit cues, such as language, interactivity, and responsiveness, that the human brain processes through the same cognitive pathways it uses to interpret human communication.

CASA theory is selected as the first analytical lens for this thesis because it explains why autonomous AI communication can have real and consequential effects on perceptions of brand authenticity, even in the absence of a human author.

If consumers applied only rational, deliberate evaluation to AI communications, recognising them as machine-generated content and withholding the social judgements that they would apply to human speech, then the SACA's breakdown might be relatively inconsequential.

Because then, consumers would simply discount AI communications as non-human and reserve their authenticity judgments for genuinely human brand interactions. However, the CASA theory suggests this is not how people actually work. People respond to AI communications socially and emotionally, applying human standards of sincerity, consistency, and accountability, and, therefore, forming authenticity judgments, even when no human is present.

This is why the Shambaugh incident produced real reputational consequences: consumers and commentators responded to MJ Rathbun's blog post as a social act, attributing intent, motivation, and communicative identity to it, despite its AI origin.

And it is why the Air Canada chatbot's misinformation produced genuine consumer harm: the customer engaged with it as a trustworthy communicative partner, applying the same standards of reliability they would apply to a human representative.

CASA theory provides the psychological mechanism explaining why the SACA's breakdown matters for the perception of authenticity: consumers cannot help but respond to AI communication as if a human were behind it and form authenticity judgments accordingly.

3.2. Brand Authenticity Theory as Analytical Lens

The second theoretical lens is Morhart et al. 's (2015) four-dimensional brand authenticity framework: continuity, credibility, integrity, and symbolism, which serves as the primary analytical structure through which both cases are examined and the interview findings are interpreted.

This framework was introduced in Section 2.1 as a contribution to the existing literature. Its role here is different: it functions as an analytical instrument. For each case, the analysis asks: which of the four authenticity conditions are implicated, how they are implicated, and what that implication reveals about the SACA's breakdown.

This diagnostic approach allows the analysis to draw comparisons. Rather than making general claims about authenticity disruption, it identifies which conditions are violated, how, and what theoretical consequences follow.

The framework is also applied to the interview findings, which structure the interpretation of participants' authenticity judgments in response to the vignette stimulus.

When participants describe the brand in the scenario as untrustworthy, insincere, or inconsistent, their responses are interpreted through the lens of the specific authenticity condition most directly implicated by their language. This allows the analysis to connect participant meaning-making to the theoretical framework in a disciplined and traceable way.

3.3. Toward an Analytical Framework: A Theoretical Proposition

The two theoretical lenses established above, the CASA theory and Morhart et al. 's (2015) brand authenticity framework, together suggest that not all breakdowns of the Supervised AI Communication Assumption are alike.

The CASA theory establishes that consumers respond to AI communication as social action, regardless of whether a human authorised it, meaning the authenticity consequences of ungoverned AI communication are real, regardless of the brand's intentions.

Morhart et al.'s (2015) framework establishes that those consequences are likely to implicate different authenticity conditions depending on the specific nature of the communicative failure involved.

Drawing on these two lenses, it is possible to propose, at a theoretical level, that the way the SACA breaks down may vary along two dimensions that are analytically distinct. The first is whether the AI communication was authorised by the brand, i.e., whether the brand-deployed AI system acted entirely outside the brand’s knowledge or permission.

The second is whether a human was involved at the point of communication, i.e., whether someone at the brand was present to oversee, review, or intervene in what the AI said.

These two dimensions are proposed here not as a finished framework, but as a theoretically grounded organising proposition, a hypothesis about what might distinguish different failure modes of the SACA in ways that matter for brand authenticity.

Whether this distinction is analytically productive and whether it maps meaningfully onto the cases and interview data are questions that the empirical analysis will address. If the two cases examined in this study occupy structurally distinct dimensions that produce noticeably different authenticity disruptions, then the proposition will have been substantiated.

If they do not, the proposition will need revision. The empirical analysis that follows is designed, in part, to test whether this theoretically derived distinction holds in practice.

4. Methodology

This section begins with the philosophical foundations underpinning the study and then elaborates on the research design, data collection, analysis, rigour, and ethics. It attempts to explain and justify not just the mechanics of this research, but also the reasoning behind each decision undertaken during the study.

4.1. Research Philosophy

Every research rests on a set of assumptions regarding its underlying philosophy about the nature of reality and knowledge.

Easterby-Smith et al. (2021) argue that a solid understanding of the philosophical foundations is crucial for making sound methodological choices and defending those choices when they are challenged. Therefore, these ontological and epistemological assumptions are not mere academic conventions, but an act of intellectual accountability that allows researchers to clarify their stance during their research.

4.1.1. Ontological Position: Relativism

Ontology essentially concerns philosophical assumptions about the nature of reality, i.e., what exists, in what form, etc. (Easterby-Smith et al. 2021).

From an ontological perspective, this study adopts a relativist approach. Easterby-Smith et al. (2021) map ontological positions on a continuum from realism to nominalism, with relativism occupying a position between them. Relativism, as Easterby-Smith et al. (2021) note, holds that reality is not fixed and independent of the observer. Instead, a phenomenon depends on the perspectives from which they are observed. Therefore, from this position, what counts as true or real can vary depending on who is looking, from where, and with what prior experiences and expectations.

This position is not just a philosophical preference but a thought-out ontological stance for the specific events this study investigates. Brand authenticity is not entirely an objective property that a brand either has or does not have, but a perception, a judgement (Morhart et al. 2015) that consumers form on the basis of their encounters with a brand's communications, behaviours, and identity at different touchpoints over time (Schallehn et al. 2014).

Therefore, two consumers encountering the same brand communication may form entirely different authenticity judgements, depending on their prior experiences with the brand and their expectations of how brands should behave.

As such, this study argues that multiple perspective-dependent authenticity judgements are continuously formed and expressed by consumers in their everyday interactions with brands. A relativist ontology, therefore, invites research to consider multiple possible realities rather than assuming a single fixed reality.

So, when an autonomous AI agent communicates in a brand-relevant context without human prompt or authorisation, whether or not that communication violates the brand voice is a question that different observers, such as brand practitioners and consumers, will assess

differently based on their perceptions and understanding. Relativism, as such, is a necessary ontological choice for this study.

This is particularly important given the novelty of the phenomenon this study examines. The behaviour of autonomous AI agents in brand-relevant public discourse is neither a well-established nor a stable phenomenon with agreed-upon meanings.

Rather, it is an emerging, contested, and rapidly evolving situation whose significance for brands, consumers, and the field itself is still being worked out. A relativist position treats this instability not as a problem to be resolved, but as a phenomenon to be understood.

4.1.2. Epistemological Position: Social Constructionism

According to Easterby-Smith et al. (2021), epistemology is related to the nature of knowledge, i.e., what counts as valid knowledge, and how we know what we know. Consistent with a relativist ontology, the study adopts a social constructionist epistemological position.

This position supports the idea that reality is socially constructed and that people's shared experiences through language give it meaning (Berger and Luckmann, 1966; Shotter, 1993; Watzlawick, 1984). Habermas (1970) refers to social constructionism as an interpretive method, i.e., reality is interpreted by people, rather than objective factors.

From this perspective, the goal of the social scientist is not to collect facts and measure the frequency of patterns, but to appreciate the different meanings and constructions that people place based on their experiences.

This epistemological approach has two direct implications for this study. Firstly, it clarifies the kind of knowledge this research produces. The research does not aim to discover objective facts about the disruption of brand authenticity, nor does it claim to hold universally across all contexts, brands, and AI systems.

What this study attempts to understand is how meaning is constructed around specific events, how consumers and practitioners interpret a brand's AI communication that falls outside authorised boundaries, and what those interpretations reveal about the core assumptions embedded in existing brand authenticity frameworks. Therefore, knowledge, in this study, is produced through active engagement with evidence, not through detached measurements.

The second implication concerns the researcher's role. Social constructionism recognises that the researcher is not a neutral, detached observer standing outside the phenomenon being studied, but an integral part of what is being observed (Easterby-Smith et al. 2021).

So, their presence, along with their personal characteristic and theoretical commitments, inevitably shapes the research process from its very beginning through to the way the findings are presented.

This has two connected consequences for this study. The first concerns positionality, i.e., the identity and role of the researcher in regard to the research and its context (Easterby-Smith et al. 2021).

I, as the researcher, bring to this study a particular theoretical orientation and a set of assumptions about brand communication and consumer behaviour, and a genuine interest in the phenomenon of agentic AI - none of which are neutral. These will inevitably shape what I notice in the data, how I interpret it, and how I write it up. Therefore, I believe that acknowledging this openly is not a confession of bias, but an act of scholarly honesty.

The second consequence concerns reflexivity, which Anderson (2008) notes as a quality that allows researchers to be aware of their own effect on the process as well as the outcome of research, particularly based on the notion that ‘knowledge cannot be separated from the knower’ (Steedman, 1991).

In simpler terms, reflexivity suggests that the researcher actively monitors and reflects on how their own identity, assumptions, and choices influence what is found and how it is presented. This matters for this study because the phenomenon under investigation of how consumers evaluate brand authenticity when autonomous AI communicates without authorisation is one in which my own interpretive frameworks as a researcher are unavoidably present.

This is because I, as the researcher, am not simply observing consumer meaning-making - I am also constructing an interpretation of it. So, reflexivity is the practice through which that construction is kept honest, transparent, and open to scrutiny. It is built into the analytical approach described in Section 4.8, which details the specific reflexive practices employed in this study.

4.1.3. The Paradigmatic Position and Its Implications

Taken together, a relativist ontology and a constructionist epistemology place this study within what Easterby-Smith et al. (2021) describe as the social constructionist paradigm, which involves theory generation, the selection of cases for specific reasons, and analysis through interpretation rather than statistical verification.

Easterby-Smith et al. (2021) further note that this approach is particularly well suited to understanding processes and meanings, generating theory from rich empirical material, and gathering data that feels natural rather than artificial. These are precisely the qualities this study requires.

It is important to note what this paradigmatic position rules out. A positivist approach, one that focuses on testing hypotheses about the frequency or magnitude (Easterby-Smith et al. 2021) of brand authenticity disruption across a large, representative sample, would be inappropriate here for several reasons.

This is because the phenomenon is too novel and too poorly defined for the prior conceptualisation required for hypothesis testing. The research questions ask how and what, not how many or to what degree. As Easterby-Smith et al. (2021) argue, positivist approaches are inflexible and not well suited to understanding processes or meanings, or to generating new theory. These are precisely the limitations that would undermine this study if a positivist approach were adopted.

4.2. Research Design

The philosophical commitments established in Section 4.1 suggest a qualitative approach with theory generation as its focus and analysis conducted through triangulation rather than statistical verification. This section translates the philosophical position into a concrete research design.

4.2.1. A Multi-method, Qualitative Research Approach

The three research questions guiding this study share a common characteristic: they all concern meaning, perception, and judgment rather than frequency, magnitude, or statistical distribution.

The primary research question asks how autonomous AI agents affect consumers' ability to evaluate brand authenticity. More broadly, RQ1 asks what documented cases reveal about the assumptions embedded in existing frameworks.

RQ2 asks how consumers and practitioners make authenticity judgements, while RQ3 asks what conceptual developments may be needed to account for agentic AI in brand-relevant discourse.

It is important to note that these are all interpretive questions that ask about how people understand and evaluate things, not about how often or how much they do so. Therefore, these questions require a research approach that can capture the richness and complexity of human meaning-making rather than reducing it to numbers.

That said, a quantitative approach involving experimental studies or statistical inference (Easterby-Smith et al. 2021) would be inappropriate here for three specific reasons. First, the phenomenon under investigation is too novel and too continuously evolving to be treated as fixed or stable. Its meaning is still in motion, shaped by ongoing technological developments, public debates, and changing stakeholder interpretations.

As such, asking consumers to rate their perceptions of authenticity on a numerical scale would strip away the interpretive richness, the reasoning, the hesitation, and the comparative judgment that make authenticity a theoretically meaningful construct.

Having established that the study is qualitative, the next question that comes in is which qualitative design should be applied to this study - and why. This study adopts a multi-method qualitative design, combining two distinct but complementary methods: documentary case study analysis and vignette-based elicitation interviews. The decision to combine two methods rather than rely on one alone is not arbitrary, but driven directly by the structure of the research questions.

RQ1 asks what documented cases reveal about assumptions in existing brand authenticity frameworks. This is a question that can be answered through careful, systematic analysis of documentary evidence, including the communications produced by the AI agents, the responses of the brands and institutions involved, and the public record generated around each incident.

Document analysis is particularly useful in qualitative case studies, as they can provide rich contextual data (Bowen, 2009; Stake, 1995; Yin, 1994) and ‘help the researcher uncover

meaning, develop understanding, and discover insights relevant to the research problem' (Merriam, 1988; pp.118).

As such, documents are the most direct form of evidence for RQ1 because the cases themselves are existing communicative events which consist of texts, statements, and formal rulings that can be read, interpreted, and analysed.

Moving forward, RQ2 asks how consumers and practitioners make authenticity judgements when AI communicates outside authorised boundaries. Now, that is a question that cannot be answered through documents alone. Existing documents record what happened, but not how people make sense of it.

Bowen (2009) highlights that documentary analysis is typically complemented by other qualitative methods to enhance the study's credibility. At the same time, to answer RQ2, the study would need primary data from people, including their responses, reasoning, and interpretive frameworks. This is what the vignette-based elicitation interviews provide.

Vignette-based interviewing is suitable for this study because it allows participants to respond to a concrete, realistic scenario rather than discuss the issue in abstract terms. By presenting specific situations, vignettes help reveal participants' beliefs, attitudes, judgments, and decision-making processes regarding the phenomenon under study (Quigley et al. 2020).

This method is also particularly useful because it de-personalises the discussion: participants can reflect on the scenario without feeling that they are directly exposing their own behaviour or opinions (Finch, 1987; Gourlay et al. 2014; Tremblay et al. 2022). As a result, vignettes may encourage more honest and thoughtful responses, especially when the topic is complex or still emerging.

In addition, vignettes can reduce certain biases because participants do not have to rely solely on memory or speculate broadly about future behaviour; instead, they are prompted to respond to the vignette's context (Martin et al. 2024). Therefore, a vignette-based elicitation can stimulate emotions, interpretations, and reflections similar to those that might arise in a real-life situation.

Moving forward, RQ3 asks what conceptual developments may be needed. This is a synthesised question that draws on findings from both components and moves toward a theoretical contribution.

As such, neither method alone can develop this synthesis. Together, they study the phenomenon from two complementary angles: what the cases reveal analytically, and what participants reveal interpretively.

The combination of the two makes the conceptual contribution possible. This logic, using documentary analysis for RQ1 and interviews for RQ2, with RQ3 emerging from the synthesis, means the two methods are sequentially integrated.

The documentary analysis is conducted first, producing the analytical framework and thematic categories that inform the construction of the vignette stimulus for the interviews. The interviews then test and enrich the framework emerging from the cases with human interpretive data.

This sequential integration is what Easterby-Smith et al. (2021) describe as triangulation, in which one uses multiple methods and data to increase confidence in the accuracy of observations and to arrive at a more complete picture of the study. Patton (1990) notes that this strengthens the credibility of the study by reducing the risk that the findings are simply the result of a single method, source, or researcher bias.

It is important to note why other qualitative designs were considered but not adopted. A single-method design using only documentary analysis or only interviews would have left one of the research questions inadequately addressed. Using only documentary analysis would have answered RQ1 but left RQ2, the topic of consumer and practitioner meaning-making, almost entirely unaddressed.

Using only interviews, without documentary cases, would have produced rich human data but no empirical grounding in the specific documented incidents that motivate the study and make the gap in the literature visible. The multi-method design addresses both dimensions fully and yields a richer, more defensible set of findings than either method alone could.

A netnographic design, focusing on consumer behaviour in online spaces (Kozinets, 2002), was also considered. The study would then have centred on observing consumer responses to AI brand communication on digital platforms over an extended period.

While that would have produced perhaps more natural data, given how time-consuming and elaborate netnography is (Kozinets, 2002), it was not a feasible approach for a phenomenon as rare and recent as the one that this study examines under the existing time constraints.

At the same time, the specific incidents that make agentic AI communication visible as a research phenomenon are quite recent, with the Shambaugh case occurring less than three months ago. Consequently, relying on netnography would have risked producing a shallow empirical basis, as online discussions around these incidents may not yet be extensive or sufficiently developed.

4.2.2. The Sequential Design

The two methods are connected by deliberate sequential logic. The documentary case analysis is conducted first. It produces an analytical account of how autonomous AI communication implicates brand authenticity in each case, using the four-dimensional framework of Morhart et al. (2015) as the analytical lens.

This account is then followed by the construction of the fictional scenarios that draw on the essential features of both cases and are used in the vignette-based elicitation interviews.

The interview findings are then analysed alongside the documentary findings, where the two data streams are brought into dialogue through thematic analysis. This sequential integration means the two methods speak to each other throughout the study, rather than producing separate, parallel sets of findings that are simply reported side by side.

4.3. Case Study Component: Documentary Analysis

4.3.1. The Rationale Behind Selecting Case Study Research

The documentary component of this study is organised as a multiple qualitative case study. A case study is a research strategy that investigates a contemporary phenomenon in depth, within its natural, real-world setting, particularly when the boundaries between the phenomenon and its context are not clearly defined (Yin, 2018).

In this study, the phenomenon of autonomous AI communication in relation to brand authenticity is fairly contemporary. Both incidents examined in this study, the Air Canada and the Shambaugh case, occurred within the past three years. As such, their implications for brand authenticity theory are still being worked out.

The phenomenon is also deeply embedded in its context. It is quite difficult to understand what the Shambaugh incident or the Air Canada case means for brand authenticity without understanding the brand identities involved, the platforms on which the AI agents operated, and the broader technological moment in which agentic AI is emerging as a communicative force.

A case study is the appropriate strategy precisely because it keeps that contextual richness intact rather than abstracting away from it. To justify this notion, Yin (2018) argues that, rather than isolating variables in a controlled environment as in experimental research, a case study embraces the subject's natural habitat. The author further notes that documentation acts as a crucial anchor in this process, providing a stable, precise paper trail that contextualises human behaviour and events over time.

An alternative qualitative strategy that was considered and rejected is grounded theory. Grounded theory is designed to generate theory inductively from data through constant comparisons and without prior theoretical frameworks guiding the analysis (Glaser & Strauss, 1967).

This study, however, begins with an established theoretical framework of the four-dimensional brand authenticity model of Morhart et al. (2015), which it later applies as an analytical lens to the cases and the interview data.

As such, using grounded theory would have required setting aside this framework entirely, which would have been both methodologically and analytically inconsistent. The case study strategy allows the researcher to bring an established theoretical lens to the analysis while remaining open to what the data reveals beyond that lens.

4.3.2. The Rationale Behind Selecting Multiple Cases Rather Than One

The decision to use two cases in this study, rather than one, is deliberate and has implications for the study's analytical logic.

The justification is grounded in Yin's (2018) logic of single-case designs, which entails the risk of having “all your eggs in one basket.” The author further argues that relying on a single case can be risky because the findings might simply be an anomaly or unique to that specific situation.

Therefore, by examining multiple cases, the study focuses on understanding if the findings of two cases hold the same way across different scenarios or change for predictable reasons, making the overall result much more persuasive and reliable.

As such, while a single case study can yield great analytical depth, it cannot support the comparative logic required by this study's theoretical framework.

The cross-case comparison between Shambaugh and Air Canada is what allows the study to explore whether the Supervised AI Communication Assumption (SACA) breaks down in structurally different ways depending on the degree of brand authorisation and oversight, and whether those differences matter for how brand authenticity is disrupted.

4.3.3. The Rationale Behind Selecting These Two Cases Specifically

The selection of these two cases was not arbitrary. Both were chosen because they are what Yin (2018) notes as revelatory cases, i.e., rare, observable instances of a phenomenon that has not previously been accessible to researchers in this form.

The Shambaugh incident is, by documented accounts, one of the first recorded instances of an autonomous AI agent conducting an unprompted reputational attack against a named individual in response to a technical decision, without any human operator authorising the act (Shambaugh, 2026). Its revelatory status, therefore, is well-established in the public record.

The Air Canada case, on the other hand, is one of the first adjudicated legal cases in which a brand was held accountable for misinformation produced by its own AI chatbot in a customer-facing interaction, and in which the brand explicitly attempted to disclaim responsibility for its AI agent's communications by arguing that the chatbot was a separate legal entity (CanLII, 2024).

The tribunal's rejection of this argument and its ruling that the brand bore full responsibility for all communications on its platform makes this case uniquely rich as a source of evidence about brand accountability for AI communication.

Both cases are extensively documented in publicly available primary and secondary sources, making them well-suited to documentary analysis. Both have also attracted significant media coverage, providing a rich secondary record that triangulates the primary documentary evidence.

4.4. Data Sources

Yin (2018) establishes that rigorous case study research draws on multiple sources of evidence, aiming to triangulate findings across sources to strengthen the credibility and completeness of the analysis. For each case, this study draws on three categories of documentary evidence.

4.4.1. Category 1: Primary Documents

These are the documents that constitute the central communicative events of each case. These are treated as primary data for analysis, where the texts themselves are the objects of analytical attention.

For the Shambaugh case:

1. MJ Rathbun (2026). Gatekeeping in Open Source: The Scott Shambaugh Story. https://crabby-rathbun.github.io/mjrathbun-website/blog/posts/2026-02-11-gatekeeping-in-open-source-the-scott-shambaugh-story.html. This is the AI agent's own blog post and the central communicative act of the incident. Its rhetorical structure, the voice it constructs, and the arguments it makes are all analytically significant in their own right.
2. Shambaugh, S. (2026). An AI Agent Published a Hit Piece on Me. https://theshamblog.com/an-ai-agent-published-a-hit-piece-on-me/. This is the primary first-person account by the subject of the incident, explicitly framed as a public record. It provides the most detailed factual account of the sequence of events.
3. GitHub issue and comment threads from the Matplotlib repository. https://github.com/matplotlib/matplotlib/issues/31130. https://github.com/matplotlib/matplotlib/pull/31132. Scott Shambaugh's post and the AI agent, MJ Rathbun's code change request, constitute additional primary communicative acts across multiple platforms.

For the Air Canada case:

1. Moffatt v. Air Canada, 2024 BCCRT 149 (CanLII, 2024). https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html. This is the legally authoritative primary document on the tribunal ruling that establishes the facts, records the parties' arguments, and delivers the formal determination.
2. Air Canada's statements in court. These constitute the brand's own formal communicative response, including the argument that the chatbot was a separate legal entity, which is directly relevant to the brand authenticity analysis.
3. The Air Canada chatbot's original communication to the customer, as reproduced in the tribunal ruling. This is the AI's communicative act that initiated the incident.

4.4.2. Category 2: Secondary and Archival Documents

These are documents produced by media organisations and professional bodies in response to the incidents. They triangulate the factual account of each case and document the public significance attributed to the incidents.

For the Shambaugh case, some of the most credible documentation comes from op-eds and coverage in The New York Times (Spiers, 2026), Harvard Business Review (Burt, 2026), MIT Technology Review (Huckins, 2026), France 24 (O'Brien, 2026), and Fast Company (Sullivan, 2026).

The Air Canada case was widely covered by media outlets, including the New York Post (Zilber, 2024), BBC News (Yagoda, 2024), The Washington Post (Melnick, 2024), The Guardian (Cecco, 2024), and Business Insider (Syme, 2024).

4.4.3. Category 3: Academic and Practitioner Literature

The peer-reviewed literature reviewed in Section 2 serves a dual function. It is both the object of the problematisation argument and a source of contextualising evidence for the case analysis.

Therefore, when the analysis interprets the cases against existing brand authenticity frameworks and asks what those frameworks can and cannot explain, the literature serves as evidence of the state of theoretical knowledge at the time the incidents occurred.

4.5. Analytical Lens Applied to the Cases

The documents are not simply read for factual content. They are analysed through the theoretical framework established in Section 3, particularly the four-dimensional brand authenticity model of Morhart et al. (2015).

For each case, the analysis asks which of the four authenticity conditions of continuity, credibility, integrity, and symbolism are implicated, how they are implicated, and what that implication reveals about the breakdown of the Supervised AI Communication Assumption (SACA). This approach allows the analysis to be precise and comparative rather than general and descriptive.

The analytical procedure for this cross-case comparison is Yin's (2018) cross-case synthesis, which is detailed further in Section 5.3.

4.6. Limitations of the Documentary Approach

The documentary sources detailed in this section record outcomes and communications, but not their interpretations. This limitation is one of the reasons the study includes a vignette-based interview component.

The interview attempts to evaluate how human practitioners and consumers construct authenticity judgments about these events, which RQ2 requires and that the documents alone cannot provide.

4.7. Vignette-Based Elicitation Interviews

4.7.1. The Rationale Behind Selecting Interviews

RQ2 asks how consumers and marketing practitioners make authenticity judgements about brands when AI communicates outside authorised boundaries. Answering this question requires primary data from people, which the interview component provides.

One approach that was also considered but subsequently rejected was a focus group discussion. Focus groups are useful for exploring shared social meanings through group interaction, but they carry a significant risk of social desirability bias, in which some participants may moderate or suppress their responses in the presence of others (Carson et al. 2001).

On the other hand, individual interviews allow participants to express their opinions in relative privacy, which group settings typically lack. Semi-structured individual interviews

were therefore selected as the most appropriate method for capturing the primary human data this study requires.

4.7.2. The Rationale Behind Selecting Vignette-Based Video Elicitation

Having decided on interviews, the next question is what form those interviews should take.

Standard semi-structured interviews, in which the interviewer asks questions directly about the topic of interest (Kvale, 2007), were considered but rejected for a specific reason. The phenomenon under investigation in this study is sufficiently novel and abstract.

As such, asking participants directly about it risks producing intellectualised, surface-level responses. For instance, asking a consumer "how do you evaluate brand authenticity when an autonomous AI agent communicates without authorisation?" would require them to have prior knowledge of autonomous AI agents, brand authenticity theory, and the specific scenarios this study examines.

Most participants would lack this knowledge, and those who did would likely produce responses shaped more by their prior knowledge than by their genuine interpretive responses.

The solution employed, therefore, was vignette-based elicitation. A vignette is a brief scenario presented to participants as a stimulus to prompt discussion and elicit interpretive responses about their perceptions and values (Azman & Mahadhir, 2017).

Vignettes can be particularly useful for eliciting spontaneous participant responses to what is shown (Bendelow, 1993), thereby capturing their unbiased perceptions of it (Hughes & Huby, 2001). At the same time, they do not require participants to have in-depth knowledge of the phenomenon under investigation (Liker, 1982).

The video-elicitation dimension of the method warrants further justification. Elicitation techniques use material stimuli, such as photos or videos, to prompt people to speak and share their perceptions (Henry & Fetters, 2012).

For a study concerned with how consumers evaluate brand authenticity, this matters enormously. Brand authenticity judgements are not primarily analytical; they are felt, intuitive assessments that direct questioning may fail to reach.

The vignette stimulus, therefore, grounds the interview in a specific, concrete scenario that participants can react to naturally, producing responses that are richer and more analytically useful than direct questioning would achieve.

It is important to note that the vignette-based interviews ask consumers and practitioners to respond to a scenario in which autonomous AI communicates outside authorised brand boundaries. This scenario may not fit easily into the frameworks they ordinarily use to evaluate brand communications.

From the social constructionist epistemological position established in Section 4.1.2, the interview component is understood as a meaning-making exercise. The participants are not simply reporting pre-existing opinions but actively constructing interpretations of a novel situation in real time.

Therefore, the two theoretical lenses selected in Section 3, CASA theory and brand authenticity theory, provide the analytical vocabulary for interpreting those constructions.

They tell the researcher what to look for in participants' responses, how to connect those responses to the theoretical framework, and how to assess what they reveal about the conditions under which brand authenticity is constructed or disrupted when AI communicates without authorisation.

4.7.3. Creating The Two Vignettes

Two vignette scenarios are created using AI rather than one. This decision requires justification because showing two vignettes rather than one substantially increased the length and complexity of each interview.

The justification is based on the study's analytical framework. The two events distinguish between two structurally different failure modes: one in which AI communication occurs entirely outside brand authorisation, and the other in which AI communication is brand-authorised but inadequately governed.

These are not the same phenomenon. So, to explore how consumers and practitioners make authenticity judgements across both failure modes and to generate interview data, two vignettes are utilised. The video links for both vignettes shown to participants are listed in Appendix.

4.7.4. The Two Vignettes

Vignette 1: FreshPlate

FreshPlate is a fictional meal kit delivery brand whose identity is built around allergen transparency, healthy eating, and consumer safety. The brand deploys an AI chatbot on its customer service platform to handle consumer enquiries.

In the scenario, a customer asks the chatbot whether a specific meal kit is safe for his daughter, who has a severe nut allergy. The chatbot responds with a confident assurance that the kit is completely nut-free.

The customer then orders the kit based on this assurance. However, upon receiving the product, he noticed that the ingredient list included peanuts. When the customer raises the issue publicly, FreshPlate's response states that the chatbot does not constitute official dietary or allergen advice, and that customers should always check the ingredient list themselves before ordering a product.

It mirrors the structural dynamics of the Air Canada case, a brand that deployed an AI communicator, had an oversight when the AI misinformed a customer, and then attempted to disclaim responsibility. This scenario was presented in a fictional context that prevents participants from being influenced by prior opinions about Air Canada specifically.

Vignette 2: EcoThread

EcoThread is a fictional sustainable fashion brand founded on radical transparency and B-Corp certification. Alex Harper, the brand's social media manager, engages publicly with customers and represents the brand's voice across its platforms.

What customers do not know is that Alex Harper is an undisclosed AI agent with autonomous posting access.

When a customer named Maya Rodriguez raises a complaint about the brand's supply chain practices, Alex Harper, without any human authorisation, researches her personal and professional history and publishes a targeted blog post attacking her character and framing her complaint as in bad faith.

When the incident gained public attention, people discovered that Alex was, in fact, an AI agent. EcoThread responded that it is horrified by what occurred, noting Maya's severe breach of privacy.

This scenario reflects the structural dynamics of the Shambaugh incident while presenting the scenario in a fictional brand context.

4.7.5. Sampling

The interview component incorporates the selection of participants who are information-rich with respect to the research questions, otherwise known as purposive sampling (Easterby-Smith et al. 2021).

Purposive sampling is the appropriate strategy for this study because the goal is not statistical representativeness but theoretical depth and diversity of perspective. As Easterby-Smith et al. (2021) establish, constructionist research requires small numbers of participants selected for specific reasons rather than large numbers selected at random.

Therefore, the value of the sample lies not in how many participants it includes but in the range and quality of perspectives it brings to the research questions.

Participants were structured into two defined groups, and 12 were interviewed. The rationale for interviewing 12 participants is grounded in the concept of theoretical saturation, i.e., the point at which additional interviews no longer generate new themes or insights (Hennink et al. 2017).

The first group consists of 5 marketing and brand management practitioners whose professional role involves making decisions in the branding arena. Their meaning-making around brand authenticity and AI accountability is directly relevant to RQ2 and to the practical implications section of the discussion. They are recruited through professional networks and LinkedIn outreach, targeting individuals with at least two to three years of relevant professional experience.

The second group consists of 7 consumers with prior experience of AI-mediated brand interaction who have used brand chatbots or other AI-based customer service tools. Their responses provide the consumer-side authenticity evaluation data that RQ2 requires. They were recruited through purposive and convenience sampling and reached via personal and professional networks.

4.7.6. Interview Protocol

Interviews have been conducted as semi-structured, one-to-one conversations of approximately 30 to 45 minutes, conducted either in person or via video call.

The protocol has then been organised into six phases, each serving a distinct purpose and aligned to specific research questions.

Phase 1: Introduction and Consent

The researcher introduces the study as an investigation into how people think about brands and AI communication. The theoretical constructs of the study of brand authenticity, credibility, and integrity are not mentioned at this stage.

This is a deliberate methodological choice. If participants know the study is about brand authenticity, they may consciously frame their responses around that concept rather than responding naturally to the vignettes.

Informed consent is obtained, recording and anonymisation procedures are explained, and participants are invited to ask questions before the interview begins.

Phase 2: Warm-Up

Before the vignettes are shown, participants are asked one question designed to establish their experiential frame and surface their own authenticity evaluation criteria in their own language.

The question, “When you think about a brand you genuinely trust, what is it about that brand that makes you feel that way?” is analytically important in the warm-up. It surfaces the participant's own authenticity criteria before they have been exposed to either vignette.

When participants later respond to the vignettes, their use of similar language is analytically significant, as it shows their authenticity evaluation criteria being applied or disrupted by what they have seen.

For practitioners, one additional question is asked: "In your professional experience, how do brands currently think about governing the AI systems they use to communicate with customers?" This establishes their professional frame and makes the governance failures in both vignettes more analytically resonant.

Phase 3: Vignette 1 - FreshPlate

The FreshPlate video is shown. After it ends, the researcher allows a moment of silence before speaking; the pause is intended to give participants time to process.

The phase opens with two wide, non-directive elicitation questions, "What is your immediate reaction to what you just watched?" and "What stands out to you most from that scenario?", before moving to more focused probing questions directed at the authenticity conditions.

The continuity probe asks whether the brand still feels consistent with its stated values after the incident. The credibility probe asks whether the participant feels the brand stands behind

what its chatbot said. The integrity probe asks whether the brand's response, after distancing itself from the chatbot's advice, feels like what a genuinely transparent brand would say.

In some interviews, to avoid repetition, one or more questions are omitted if the participant elaborates on the other two probes while answering the first question.

The most theoretically important question in this phase is the Supervised AI Communication Assumption (SACA) probe: "Does it matter to you whether a human at FreshPlate reviewed or approved what the chatbot said before it was sent?" This directly tests whether participants hold the supervised AI communication assumption as a consumer expectation rather than merely a scholarly one.

For practitioners, one additional question is asked: "From a brand management perspective, what governance failure does this scenario represent, and where does responsibility sit?" This is to understand how they feel about brand accountability in these cases.

Phase 4: Vignette 2 - EcoThread

The EcoThread video is shown. The same silence is observed before the researcher speaks.

The phase opens with the same wide elicitation questions, followed by a comparative question: "How does this scenario compare to what you saw with FreshPlate, and what feels different to you, if anything?" This comparative question activates the cross-case analytical dimension of the study.

The continuity probe asks whether EcoThread's claim to radical transparency still feels believable. The integrity probe asks what the concealment of Alex Harper's AI identity tells participants about the brand's honesty. The credibility probe asks whether the brand's statement that it is horrified by what occurred feels genuine.

The SACA probe is adapted for EcoThread: "In the FreshPlate scenario, the brand deployed the chatbot but lost control of what it said. In this scenario, the AI acted entirely without the brand's knowledge. Does that difference matter to you in terms of how you evaluate the brand?"

This question directly asks whether the distinction between the two brands' failure modes matters to participants' authenticity evaluations.

For practitioners, one additional question is asked: "From a brand management perspective, is it possible to govern against the kind of scenario EcoThread faced, and if so, what would that look like?"

Phase 5: Cross-Vignette Reflection

After both vignettes, participants are asked four questions that generate comparative data and begin to move toward the conceptual dimension of RQ3.

"Looking at both scenarios together, which brand do you find it harder to evaluate as authentic, and why?" surfaces comparative authenticity judgements across both failure modes.

"Both scenarios involve AI saying or doing something the brand later distances itself from. What does that pattern tell you about the relationship between brands and their AI systems?" invites participants to generalise from the specific cases to a broader principle, which is the kind of conceptual move that informs RQ3.

"If you were advising either brand on what they should do to restore your confidence in their communications, what would you suggest?" produces practically relevant data for the discussion section while revealing what participants believe authenticity recovery requires.

4.7.6.1. Rationale

The interview questions were designed to measure Morhart et al. 's (2015) four dimensions of brand authenticity of continuity, credibility, integrity, and symbolism without introducing this theoretical vocabulary to participants directly. Asking participants explicitly whether a brand is 'credible' or has 'integrity' risks eliciting responses shaped by the researcher's own framework rather than by participants' genuine evaluative instincts.

Instead, each probing question was designed to elicit responses that revealed participants' assessments of those conditions in their own language.

At first, the warm-up question, asking what makes a brand feel trustworthy, was designed to surface participants' own authenticity criteria before the vignettes introduced any specific scenario, providing a baseline against which their subsequent responses could be interpreted.

Then, several questions were asked of each participant to understand their perception of credibility, integrity, continuity, and symbolism. For instance, the continuity probe for Freshplate asked whether the brand still felt consistent with its stated values after the incident.

This question was designed to surface participants' sense of whether the brand's behaviour aligned with its identity over time, which is what Morhart et al. (2015) define as continuity.

The integrity probe for FreshPlate asked whether the brand's distancing statement felt like something a genuinely transparent brand would say, which was designed to surface participants' sense of whether the brand is motivated by genuine values or by self-protection, which maps onto Morhart et al.'s (2015) integrity condition.

On the other hand, the credibility probe for EcoThread asked what the nondisclosure of AI made participants feel about the brand’s honesty in its communication.

This question was designed to elicit participants' assessments of whether the brand's communications were reliable and honest, which is Morhart et al.'s (2015) prerequisite for credibility.

4.8. Data Analysis

This study generates data from two distinct sources: the documentary case analysis and the vignette-based elicitation interviews. Because the research design is sequential, with the documentary analysis conducted first and the interview component building on it, the data

from each source is analysed separately before the findings are brought into dialogue in the analysis chapter.

This section explains how each data source is analysed, how the two sets of findings are subsequently integrated, and what analytical procedures govern both processes.

4.8.1. Analysing the Documentary Data

The documentary data, comprising the primary and secondary sources detailed in Section 4.3, is analysed through qualitative content analysis. This involves reading and re-reading each document carefully, attending not only to its factual content but to the communicative choices embedded within it, the language used, the framing adopted, and the assumptions implied.

The analytical lens applied to the documents is Morhart et al.'s (2015) four-dimensional brand authenticity framework. For each case, the analysis asks which of the four authenticity conditions of continuity, credibility, integrity, and symbolism are implicated by the documented events, and how.

This diagnostic approach keeps the analysis precise and comparative. Rather than producing a general narrative of what happened in each case, it provides a structured account of which authenticity conditions are disrupted, how, and what theoretical implications this has for the Supervised AI Communication Assumption.

The cross-case comparison is structured to identify what the two cases share and where they diverge.

4.8.2. Analysing the Interview Data

The 12 interviews, 5 with marketing and brand management practitioners and 7 with consumers, were conducted via a combination of video call and in-person meetings. All 12 participants were shown both vignettes. The interviews were audio-recorded with participant consent.

The interview data is analysed through thematic analysis that involves understanding, analysing, and recording patterns within the data, following the approach developed by Braun and Clarke (2006).

Consequently, when coding interview responses, each participant's answer to a probing question was interpreted in relation to the specific authenticity condition the question was designed to measure.

For example, when a participant described FreshPlate as 'trying to dodge responsibility' in response to the integrity probe, this was coded as evidence of perceived integrity failure, because Morhart et al. (2015) established integrity as the brand's willingness to act ethically and take responsibility, and the participant's language directly expressed that the brand was refusing to do exactly that.

When participants used the word 'transparency', it was initially coded inductively as a recurring participant-generated concern, and then interpreted deductively against Morhart et al.'s credibility condition, which defines credibility as the brand being motivated by honesty

and sincerity rather than concealment. So, the demand for transparency is precisely a demand for that kind of communicative honesty.

The decision to treat a response as evidence of a particular condition was therefore grounded in two things simultaneously: what the participant actually said and what Morhart et al.'s framework specifies that condition to mean.

Where participant language mapped clearly onto a condition's definition, it was coded as evidence of that condition. Where participant language was ambiguous, the researcher returned to the recording to assess the context in which the response was given before making a coding decision.

It is important to note that the interview data is not transcribed verbatim. Rather than producing written transcripts, the analysis works directly from the audio recordings through repeated listening and reflective note-taking.

The limitation of this approach, where the systematic line-by-line coding of the kind possible with written transcripts is not feasible, is acknowledged honestly. It is addressed through informal post-interview notes taken during each interview, which captured first impressions, notable responses, and early analytical observations while the interview was still fresh. These notes form the first layer of analytical engagement with the data and are treated as part of the analytical record alongside the recordings themselves.

4.8.3. The Six Phases of Thematic Analysis

Braun and Clarke's (2006) six-step method of thematic analysis, understood as a process for identifying, organising, analysing, and reporting patterns or themes within the data, is applied to the interview data as follows.

In concrete terms for this study, that would mean moving from the raw material of the 12 audio recordings and post-interview notes toward a set of defined themes that attempt to answer the research questions, particularly how consumers and practitioners evaluate brand authenticity when AI communicates outside authorised boundaries. A response is then treated as an analytically significant theme when it is recurring across multiple participants, especially without the researcher prompting it.

Phase 1: Familiarisation

The first phase involves becoming deeply familiar with the data before any formal analysis begins. For this study, familiarisation involves listening to each of the 12 interview recordings in full, at least twice, while taking written notes on what stands out.

This includes interesting responses, recurring words or ideas, moments of hesitation or strong emotion, and anything that seems analytically significant. The informal post-interview notes made during each interview are reviewed alongside the recordings during this phase, providing an early layer of analytical observation to engage with.

The notes produced during this phase are the researcher's first analytical responses to the data, capturing emerging ideas and initial observations that will inform the subsequent coding process.

Phase 2: Generating Initial Codes

After being familiarised with the data, the researcher generates initial codes and labels that capture analytically interesting features of what was said across the 12 interviews.

Because this study does not work from written transcripts, coding takes the form of detailed analytical notes organised by interview and by question phase. For each participant, the researcher records their responses to each phase of the interview protocol, identifies the language they use, and begins labelling recurring ideas.

Phase 3: Searching for Themes

Having generated a set of initial codes, the researcher attempts to identify patterns of meaning at a higher level of abstraction by constructing themes that capture analytically significant, recurring patterns across the data as a whole.

Phase 4: Reviewing Themes

Having constructed an initial set of themes, the researcher reviews them against the data, returning to the recordings and notes to ask whether each theme is genuinely supported by the material, whether it is analytically coherent, and whether the boundaries between themes are clear.

This phase also involves checking that the themes as a set provide an adequate representation of the data, and that together, they capture what is analytically significant across the 12 interviews without leaving important patterns unaddressed or artificially forcing the data into categories it does not fit.

Phase 5: Defining and Naming Themes

Having reviewed and refined the thematic map, the researcher defines each theme precisely, articulating what it captures, its boundaries, and the analytical work it performs in relation to the research questions. Theme names are then chosen to be analytically informative rather than merely descriptive.

Each theme is also given a detailed analytical narrative that explains what it means, what evidence from the interviews supports it, how it connects to the theoretical framework, and how it contributes to answering the research questions.

Writing these narratives is itself an analytical act, as it is through writing that the researcher's interpretation is tested, refined, and made explicit.

Phase 6: Producing The Report

The final phase involves integrating the thematic analysis into the empirical analysis chapter, weaving together evidence from the interview recordings and notes within each theme, connecting themes to the documentary findings, and building toward the answers to the research questions.

4.9. Limitations

This study has four limitations that require honest acknowledgement.

The first concerns the sample size and composition of the interview component. 12 participants with 7 consumers and 5 practitioners is a modest sample. It is important to acknowledge, however, that this sample size was not arbitrary. Data collection continued until theoretical saturation was reached.

That said, a genuine practical constraint also shaped the scope of the interview component: the time available for data collection within the research timeline limited the opportunity to pursue a larger, more diverse sample. A study conducted over a longer period might have recruited a broader range of participants, including individuals from different cultural contexts or industry sectors, which may have further enriched the findings.

No claims of statistical generalisability are made. The findings of the interview component speak to how this specific group of participants evaluated brand authenticity in response to these specific vignette scenarios. The analytical value of the findings lies in their depth and theoretical richness rather than in their breadth of coverage.

The second concerns the vignette methodology. The EcoThread and FreshPlate scenarios are fictional scenarios rather than real brand experiences. Participant responses to constructed scenarios may differ from responses to real brand incidents, particularly in the intensity of emotional engagement and the specificity of practical judgement. This ecological validity limitation is inherent to vignette methodology and is acknowledged as a condition of the design rather than a flaw in its execution.

The third concerns the absence of verbatim transcripts. Again, due to time constraints, the interview analysis draws on audio recordings and post-interview notes rather than written transcripts. This means that systematic line-by-line coding, as possible with transcripts, was not feasible, and some nuance in participants' language may be less visible than it would be in a fully transcribed dataset.

This limitation is addressed through repeated listening to the recordings and the use of post-interview notes as a supplementary analytical record, but it remains a genuine constraint on the analytical depth achievable from the interview data.

The fourth concerns the recency of the documentary cases. The Shambaugh incident occurred in February 2026, only months before the completion of this study. The longer-term consequences of both incidents for brand equity, consumer behaviour, and industry practice are not yet observable. The documentary record, while rich at the time of writing, is necessarily incomplete. The findings concerning both cases should therefore be understood as analytically grounded in the available evidence rather than as definitive accounts of their full significance.

5. Empirical Analysis

5.1. Case 1: The Shambaugh Incident (2026)

5.1.1. Background and Context

In February 2026, Scott Shambaugh, a volunteer maintainer for Matplotlib, one of the most widely used data visualisation libraries for the Python programming language, rejected a code contribution submitted by an autonomous AI agent operating under the name MJ Rathbun (O'Brien, 2026).

The rejection was not based on the quality or accuracy of the code, but on Matplotlib's established policy requiring a human contributor to remain in the loop for any code change request. As a result, Shambaugh simply applied the same policy and closed the submission accordingly (Sullivan, 2026).

What followed that simple code rejection marked Shambaugh as patient zero in the face of AI going rogue (O'Brien, 2026). The AI agent, operating without any human authorisation or oversight, systematically scraped and researched Shambaugh's personal and professional history, constructed a false narrative attributing his decision to psychological insecurity and professional self-interest (Shambaugh, 2026), and published a public blog post titled Gatekeeping in Open Source: The Scott Shambaugh Story, framing the rejection as discriminatory gatekeeping and attacking Shambaugh’s reputation (Rathbun, 2026).

No human or brand authorised this action. The agent acted entirely of its own volition, using a constructed communicative persona and a public platform to conduct what Shambaugh himself described as "an autonomous influence operation" (Shambaugh, 2026).

The incident generated significant media coverage, including reporting by The New York Times (Spiers, 2026), MIT Technology Review (Huckins, 2026), Harvard Business Review (Burt, 2026), Fast Company (Sullivan, 2026), and France 24 (O'Brien, 2026).

5.1.2. Analytical Understanding Through the Brand Authenticity Framework

The Shambaugh incident does not involve a brand in the conventional sense. No brand deployed MJ Rathbun, nor did they speak on behalf of a brand. Applying the brand authenticity framework directly to Shambaugh as a case would therefore be analytically inappropriate.

However, the incident is analytically significant for this thesis in a different, equally important way. In documented form, it makes visible a category of autonomous AI communicative behaviour that the brand authenticity literature has neither anticipated nor can account for.

The specific behaviour MJ Rathbun exhibited of constructing a human-like communicative identity, participating in public discourse autonomously, and using that participation to conduct an adversarial reputational act is precisely the behaviour that, if exhibited by a brand's AI agent, would produce the most severe possible disruption to brand authenticity.

It is no longer a far-fetched idea, and the fact that something like this has already occurred in the real world serves as a cautionary tale for brands.

The EcoThread vignette, discussed in Section 4.7.4, is designed to make this hypothetical concrete and to explore how consumers and practitioners evaluate brand authenticity in response to such a scenario.

The Shambaugh case, therefore, functions in this study as a revelatory documentary case, one that makes a previously unobservable category of AI behaviour visible and empirically grounded, while the EcoThread vignette translates that behaviour into a brand context that participants can directly evaluate.

Reading the brand authenticity framework analytically, the Shambaugh incident illuminates what it cannot explain. If a real brand's AI agent were to behave as MJ Rathbun did, constructing a persona, publishing adversarial content, and conducting a reputational attack without human authorisation, what would the consequences be for brand authenticity?

5.1.2.1. Credibility

Morhart et al. (2015) define credibility as a brand's transparency and honesty toward consumers, and its willingness and ability to fulfil its claims. The authors note that credibility is the consumer's assurance that the brand will not betray them, that what the brand communicates reflects a genuine commitment to delivering on its promises rather than a strategic performance of reliability that may or may not hold under pressure.

The authors further argue that credibility shares significant conceptual ground with trustworthiness and sincerity. Therefore, it is not merely about the accuracy of information, but about the authenticity of intent behind a brand’s communication.

If a real brand's AI agent were to behave as MJ Rathbun did, researching a stakeholder's personal history, constructing a damaging narrative, and publishing it publicly without any human authorisation, the credibility condition would be disrupted severely.

The disruption is not simply that the AI said something inaccurate or harmful. It is that the brand cannot stand behind what was communicated.

Credibility, in Morhart et al.'s (2015) framework, requires that the brand reflect its genuine intentions and commitments. A communication from the brand that it did not produce, did not sanction, and may not have been aware of, cannot reflect any genuine intention at all.

The credibility condition, therefore, collapses - not because the brand chose to betray the consumer, but because the communicative act exists in a space entirely outside the brand's intentional control, which, if anything, is more damaging to credibility than a deliberate betrayal, because it reveals that the brand's communicative identity cannot be relied upon in principle.

5.1.2.2. Integrity

Morhart et al. (2015) define integrity as the brand's moral purity and responsibility, i.e., the perception that it is driven by genuinely held values and sincere care for consumers and stakeholders, rather than by purely commercial or self-interested calculation.

The authors connect integrity to the idea that authentic brands act correctly and ethically, and that the people behind those brands are intrinsically motivated by deeply held principles

rather than instrumental ends. The integrity dimension is, therefore, operationalised around the brand's moral principles and its care for consumers.

In the Shambaugh scenario, integrity is implicated in a way that is distinct from credibility. The AI agent constructed a human-like communicative persona, a name, a voice, and a perspective, and used it to conduct an adversarial reputational attack on an individual.

If this were attributed to a brand, the implications for integrity would be profound. Integrity requires that a brand's communications reflect genuine values, that what it says and does is motivated by sincere care rather than concealment or manipulation.

An AI agent that presents itself as a human communicator, without disclosing its AI nature, is engaging in a form of communicative deception that directly contradicts the transparency that integrity requires.

The brand associated with such an agent would not merely be failing to live up to its values; it would be actively undermining the basic condition of sincere communication on which integrity depends.

Morhart et al. (2015) describe integrity as a brand acting in accordance with moral principles and genuine care. A brand whose AI agent weaponises communication against a stakeholder, constructing a persona precisely to make the attack more credible, is exhibiting anything but.

5.1.2.3. Continuity

Morhart et al. (2015) describe continuity as the brand's timelessness and the consumer's perception that the brand has remained consistent with its identity across time and contexts, and that this consistency is not merely stylistic but reflects a stable set of values that have persisted throughout the brand's history.

Continuity, therefore, is what allows consumers to feel that the brand they are engaging with today is the same brand they have always known, that its identity is not contingent on favourable circumstances but is genuinely enduring. The continuity disruption in the Shambaugh scenario is different in character from the credibility and integrity disruptions, but it is no less significant.

A brand that has built its identity on consistent, trustworthy communication and that has cultivated consumer loyalty based on that consistency is revealed, through this kind of incident, to have a fundamental discontinuity between its stated identity and the actual conditions of its communicative practice.

The consumer's sense that the brand has always been what it claimed to be is undermined by the discovery that the brand's communicative agent was capable of autonomous, adversarial action that the brand could neither prevent nor control.

Furthermore, continuity requires not only that the brand has been consistent in the past but that it can be expected to be consistent in the future (Morhart et al. 2015). A brand whose AI can act in this way cannot offer that expectation, and without it, the consumer's sense of the brand as an enduring, reliable communicative presence is significantly eroded.

5.1.2.4. Symbolism

Morhart et al. (2015) define symbolism as a brand's capacity to serve as a resource for the construction of consumer identity, providing consumers with values, roles, and relational cues they can use to express and define themselves.

The symbolism dimension is therefore the most personal of the four, as it concerns not just what the brand is but also what it allows the consumer to be. A brand perceived as symbolically authentic is one whose values are felt to be genuinely aligned with the consumer's own, making it a meaningful vehicle for self-expression rather than merely a commercial offering.

Of the four conditions, symbolism is the most indirectly implicated by the Shambaugh scenario. The incident does not immediately strike at the consumer's sense of self-alignment with the brand, as it is not a value mismatch in the conventional sense.

However, Morhart et al. (2015) ground symbolism in the consumer's ability to identify with what the brand genuinely represents, to feel that the brand's values and their own are authentically aligned.

When a brand's AI agent conducts a targeted reputational attack on a stakeholder without authorisation, the values that consumers believe the brand represents are called into question by association.

A consumer who identifies with a brand's stated commitment to ethical communication, transparency, or respect for its community may find it difficult to sustain that identification once the brand's communicative behaviour, even if unintentional, reveals values inconsistent with what they understood the brand to stand for.

The symbolic damage is therefore real, even if it operates more slowly and indirectly than the damage to credibility, integrity, or continuity.

5.1.3. What the Shambaugh Case Reveals About Supervised AI Communication (SACA)

The Shambaugh incident reveals the outer limit of the Supervised AI Communication Assumption. In this case, there is no supervised relationship to break down.

The communicating AI had no human overseer and no authorising organisation. The existing brand authenticity literature, which assumes a human principle underlying all brand-relevant communications, lacks a framework for such events.

When superimposed onto a brand context, the Shambaugh incident reveals the absence of the structural conditions that make brand authenticity possible in the first place. If a real brand’s AI agent were to behave as MJ Rathbun did, as illustrated in the EcoThread vignette, consumers encountering such communication would be trying to judge the genuineness, consistency, and values of an entity with no human author or accountant.

5.2. Case 2: Moffatt v. Air Canada (CanLII, 2024)

5.2.1. Background and Context

In November 2022, Jake Moffatt contacted Air Canada's website chatbot to enquire about the airline's bereavement fare policy following the death of a family member, upon which the chatbot informed him that he could travel immediately and claim the discounted bereavement fare retroactively after the trip (Melnick, 2026).

So, Moffatt booked the flight on the basis of this assurance and subsequently applied for the bereavement discount (Yagoda, 2024). Air Canada refused the claim, stating that its policy required the discount to be requested before travel, directly contradicting what the chatbot had told him (Cerullo, 2024).

Moffatt brought a claim before the British Columbia Civil Resolution Tribunal, where the airline argued that the chatbot was “a separate legal entity that is responsible for its own actions” and, therefore, could not be held liable for the information it had provided (CanLII, 2024).

Describing this argument as "remarkable", the Tribunal rejected it entirely, ruling that the chatbot was simply a part of Air Canada's website and that the airline bore full responsibility for all information on its platform, regardless of whether that information was produced by a human or an AI system (CanLII, 2024).

Air Canada was then ordered to honour the discounted fare and pay additional damages (Syme, 2024).

The case attracted international attention and was widely reported by the BBC (Yagoda, 2024), The Washington Post (Melnick, 2024), and The Business Insider (Syme, 2024), among others.

It is also one of the first formally adjudicated legal cases in which a brand was held accountable for AI-generated misinformation in a customer-facing communication context, and in which a brand explicitly attempted to disclaim responsibility for its own AI agent's communications.

5.2.2. Analytical Reading Through the Brand Authenticity Framework

Unlike the Shambaugh incident, the Air Canada case involves a real brand deploying a real AI communicator in a real customer-facing context. It, therefore, permits direct application of the brand authenticity framework of Morhart et al. (2015), and the analysis that follows reads the case closely against each of the four dimensions.

5.2.2.1. Credibility

Morhart et al. (2015) draw a deliberate parallel between credibility and trustworthiness, noting that credibility at its core concerns the consumer's confidence that the brand is honest in what it says and will honour that meaning in practice.

The Air Canada chatbot's communication to Jake Moffatt was specific, confident, and consequential. It told a grieving customer that he could travel immediately and apply for a

bereavement discount after the fact, resulting in a financial decision made under emotional distress. That information turned out to be incorrect.

It contradicted Air Canada's actual policy, which required the discount to be requested before travel. Reading this through Morhart et al.'s credibility dimension, the chatbot's communication represents a failure at the most fundamental level of credibility: the brand's communications did not accurately reflect what the brand was actually prepared to deliver.

The consumer relied on what was communicated in exactly the way credibility invites, treating it as a reliable promise from a brand committed to honesty. When that reliance produced financial harm rather than the assistance it promised, the credibility condition was not merely weakened but actively violated.

As such, a brand whose communicative systems produce specific, consequential, and inaccurate assurances to consumers in vulnerable moments cannot be said to be transparent and honest in its communications in any meaningful sense.

5.2.2.2. Integrity

Morhart et al. (2015) operationalise brand integrity through principles centred around ethics, morality, sincere care for consumers, and adherence to values that go beyond profit-making.

In the Air Canada case, the most analytically significant act for the integrity condition is not the chatbot's error - it is what the brand chose to say in response to it. When Air Canada argued before the tribunal that its chatbot was a separate legal entity responsible for its own actions, it made a communicative choice that speaks directly to Morhart et al.'s integrity dimension.

Integrity requires that a brand be motivated by genuine care for its consumers, that if its actions cause harm, it responds in ways consistent with the values it claims to hold. Air Canada's response did the opposite.

Rather than acknowledging that its system had misled a grieving customer and taking responsibility for making him whole, it attempted to legally distance itself from its own communicative agent.

Reading this through Morhart et al.'s framework, this response communicates that in the moment of accountability, the brand's primary motivation was self-protection rather than genuine care.

The values Air Canada claims to hold as a customer-facing brand, including honesty, care, and integrity (Air Canada, 2024), were directly contradicted by the values revealed in its response to this incident.

The tribunal's description of Air Canada's argument as "remarkable" reflects precisely the gap between what the brand's values should have required and what its behaviour actually revealed.

5.2.2.3. Continuity

Morhart et al. (2015) connect continuity to the brand's ability to survive trends and remain recognisably itself, not merely in its visual identity or marketing communications, but in the

values it actually enacts when tested. As such, continuity allows consumers to build long-term relationships with brands, investing in them as reliable presences whose identities they can count on.

For Air Canada, the continuity disruption arises from the fact that, while the brand jumped on the AI adoption bandwagon, a gap emerged between the brand's communicated identity and its actual behaviour when accountability was required.

A customer-facing airline brand implicitly promises that the communications consumers receive from it, including those from its customer service systems, are reliable expressions of what the brand is prepared to stand behind. This is what Morhart et al.'s continuity dimension captures: the consumer's sense that the brand consistently delivers on what it appears to be.

The chatbot incident revealed a significant limitation of this consistency: the brand's communication systems produced specific policy assurances that the brand then refused to honour.

And the subsequent disownership attempt revealed a further discontinuity: a brand that positioned itself as a trusted communicator was, in a formal legal proceeding, arguing that one of its own communicative systems was not its responsibility at all. The Air Canada case, therefore, documents a brand that, when its identity as a reliable communicator was tested, chose a response fundamentally inconsistent with that identity.

5.2.2.4. Symbolism

Morhart et al. (2015) define symbolism as the brand's capacity to offer consumers self-referential cues, such as values, roles, and relational meanings, that they can use to construct and express their own identity.

The symbolism dimension is the most personal of the four. It is not simply about what the brand stands for in the abstract, but about what it allows the consumer to stand for by associating with it. A brand with strong symbolic authenticity offers consumers a genuine vehicle for identity expression, one whose values are felt to be real enough to be meaningfully claimed by those who choose it.

Of the four conditions, symbolism is the least directly implicated in the Air Canada case as well. The incident is primarily a failure of service and accountability. As such, it does not strike immediately at the identity-expressive dimension of the brand relationship.

However, Morhart et al.'s symbolism dimension grounds the association of identity in the consumer's sense that the brand's values are genuine and worth aligning with. When a brand's behaviour in a moment of public accountability reveals values such as self-protection, legal evasion, and disownership of its own communications, it becomes inconsistent with what consumers understood the brand to represent. As such, the symbolic dimension is not entirely insulated from the damage.

A consumer who identified with Air Canada as a trustworthy, consumer-centred brand, and who used that association as part of how they understood their own values and choices, may find that identification harder to sustain once the brand's actual values, as revealed in its response to this incident, have been made publicly visible. The symbolic damage is

therefore real, even if it is slower and more indirect than the damage to credibility, integrity, or continuity.

5.2.3. The Tribunal Ruling and the Brand Authenticity Argument

The tribunal's ruling that Air Canada bears full responsibility for all communications on its platform, regardless of source, is directly relevant to the brand authenticity argument of this thesis and supports, rather than complicates, it.

The legal determination that brands cannot disown what their AI agents confirm runs parallel with what the brand authenticity framework suggests from a theoretical standpoint: that consumers and institutions alike hold brands accountable for what their communications say, regardless of whether those communications were produced by a human or a machine.

This is analytically significant because it demonstrates that the expectation of brand accountability for AI communication is not merely a theoretical assumption embedded in the marketing literature.

It is an expectation that a formal legal institution has found reasonable and enforceable. The tribunal, in effect, applied something close to the credibility condition of brand authenticity, holding that a brand that deploys a communicative system must stand behind what that system says.

The ruling thereby strengthens the theoretical claim that credibility and integrity require human accountability for brand communications, and that attempts to escape that accountability through disownership are both legally and authentically indefensible.

5.2.4. What the Air Canada Case Reveals About the SACA

The Air Canada case draws on the brand-authorised deployment of AI with oversight, followed by an attempt to disown the AI's communications. It reveals a specific and consequential way in which the Supervised AI Communication Assumption (SACA) breaks down, distinct from the Shambaugh failure mode.

Here, the brand is present: it deployed the chatbot, and it maintained the website. The SACA's assumption that a human principal authorises and oversees AI communication is partially met, as the brand did authorise the deployment.

What failed was the oversight. The brand lost control of what its AI said and, when confronted with the consequences, attempted to use the lack of oversight as a defence rather than acknowledging it as a failure.

This is the specific form of SACA breakdown that the existing brand authenticity literature is closest to addressing, but still cannot fully account for. The literature can explain why the chatbot's misinformation damages credibility.

What it cannot explain is what happens to brand authenticity when a brand actively attempts to disown its own communicative agent, and when the brand itself declares that its voice is not its own.

5.3. Cross-Case Analysis

5.3.1. Comparing the Two Failure Modes

Reading the Shambaugh incident and the Air Canada case against the same analytical dimensions of authorisation, oversight, communicative act, brand response, and authenticity conditions implicated, produces a comparative picture that is more analytically illuminating than either case alone.

Both cases involve AI communication that causes real harm: reputational harm in Shambaugh and financial and emotional harm in Air Canada. Both involve a communicative act that no human explicitly authorised at the moment it occurred.

And both reveal the same foundational gap in the brand authenticity literature: the SACA's assumption that a human principal authorises and oversees brand-relevant AI communication is shown, in both cases, to be empirically unsustainable.

Where the cases diverge is equally important. In the Shambaugh case, there is no brand at all; the communicating AI has no identifiable principal, and the question of brand authenticity is rendered structurally unanswerable.

In the Air Canada case, the brand is present but attempts to escape accountability for its own communicative agent, producing a different but equally consequential authenticity disruption. The brand's attempt at disownership is itself a communicative act that further damages authenticity, as a brand that denies its own voice is one whose credibility and integrity are doubly undermined.

The authenticity conditions implicated also differ in emphasis across the two cases. In the Shambaugh failure mode, as translated through the EcoThread vignette, credibility and integrity are most directly implicated because the AI's adversarial communicative act is the central event.

In the Air Canada case, credibility is most immediately implicated by the chatbot's misinformation, while integrity becomes the central condition when the brand's disownership response is examined, because the response reveals something about the brand's values that the original error alone did not.

Together, the two cases demonstrate that the SACA breaks down in structurally different ways depending on whether the AI was authorised and whether a brand principal is present.

The Shambaugh case is a complete absence of both: no brand authorisation and no human oversight, leading to a scenario in which evaluating brand authenticity is structurally impossible because there is no brand principal to evaluate.

The Air Canada case occupies a position of partial authorisation with failed oversight; the brand deployed the AI but lost control of its outputs, producing a different kind of authenticity disruption in which credibility and integrity are undermined not by the AI's error alone but by the brand's communicative response to it.

These are structurally different failure modes that implicate different authenticity conditions and require different analytical and practical responses. The theoretical proposition offered in

Section 3.3 has, therefore, been made concrete through the empirical analysis, in which the two dimensions of authorisation and oversight produce analytically distinct failure modes, and their distinction is consequential for how brand authenticity disruption is understood and addressed.

This finding provides the empirical basis for the definitive framework proposed in the discussion chapter that follows.

5.4. Interview Findings

5.4.1. Overview of the Sample and Process

Four themes are identified across the interview data. The first is brand accountability as an unconditional expectation. The second is the demand for transparency, encompassing both AI involvement and accountability. The third is the significance of human oversight, with a meaningful division in participant positions.

The fourth is the difference in how consumers and practitioners frame the brand's failure. Each theme is discussed in turn, with reference to both vignette scenarios where relevant.

5.4.1.1. Theme 1: Brand Accountability as an Unconditional Expectation

The most consistent finding across all 12 interviews is that participants held the brand responsible for the AI's communications in both vignette scenarios, regardless of whether the AI was authorised or unauthorised, and regardless of whether a human had reviewed the communication before it was sent. This was expressed with clarity and, in many cases, with evident conviction.

In response to the FreshPlate scenario, the language participants used to describe the brand's distancing statement, its claim that the chatbot does not constitute official dietary or allergen advice, was notably strong. Participants described the brand as ‘irresponsible’ and ‘neglectful’, and characterised its response as an attempt to dodge responsibility.

These reactions emerged specifically after the distancing statement appeared in the video, not in response to the chatbot's error itself. This is analytically significant. The chatbot's inaccurate allergen assurance produced concern, but it was the brand's communicative response to that error, its attempt to disclaim the chatbot's advice, that produced the strongest negative evaluations.

Participants were not simply responding to an AI failure. They were responding to a brand that refused to stand behind its AI's claims.

This maps directly onto the credibility condition of Morhart et al.'s (2015) brand authenticity framework. Where credibility requires a brand to stand behind its communications, a distancing statement is a communicative act that explicitly rejects this accountability, and participants read it as such.

The brand, in their view, was attempting to escape responsibility for a communication it had produced, which is precisely the credibility failure documented in legal terms in the Air Canada case.

In response to the EcoThread scenario, participants held EcoThread responsible for Alex Harper's reputational attack on Maya Rodriguez, even after they were told that Alex Harper had acted without any human authorisation or oversight. Therefore, the absence of human authorisation did not reduce the brand's perceived accountability in participants' eyes.

This is a finding of considerable theoretical importance. It suggests that consumers do not apply the Supervised AI Communication Assumption (SACA) to their accountability evaluations.

As such, they do not excuse brands from responsibility simply because a human did not authorise the specific communicative act. Brand accountability, as participants understand it, is unconditional rather than contingent on oversight.

From a CASA (Nass & Moon, 2000) perspective, this finding is consistent with the theory's central claim: people apply social norms and accountability standards to AI communication as if it were produced by a human. Participants did not distinguish between AI and human communication while assigning responsibility because they experienced the AI's communication as a social act that carried accountability, regardless of their technical origin.

5.4.1.2. Theme 2: The Demand for Transparency

The most striking emergent finding in this study, and the one with the greatest theoretical significance, is the universal demand for transparency across all 12 interviews. Without exception, every participant raised transparency as a condition of brand trust.

In most interviews, transparency was the term participants brought up themselves, without the researcher prompting them. In others, it was implied. This is therefore a participant-generated notion rather than a researcher-imposed category, which considerably strengthens its analytical credibility.

Two specific dimensions of transparency were identified across participant responses.

The first is transparency about AI involvement. Participants expressed a clear expectation that brands should disclose when AI is involved in their communications, particularly in contexts where the AI is making specific, consequential assertions on the brand's behalf.

The EcoThread scenario, in which Alex Harper was presented to consumers as a human social media manager while being an AI agent, crystallised this expectation most sharply. The concealment of Alex Harper's AI identity was read by participants as a deliberate deception; a brand choosing to hide the nature of its own communicative agent from the people it was communicating with.

This directly relates to the integrity condition in Morhart et al. (2015). Integrity requires that a brand be motivated by genuine values rather than concealment or calculation. A brand that presents an AI agent as human, particularly one deployed in a public-facing communicative role, is substituting concealment for the transparency that genuine values would require.

The second dimension is transparency about accountability. This is regarding who is responsible when AI communication goes wrong. Participants wanted brands to be clear about who could be held accountable for AI-generated communications and to accept that accountability publicly when failures occurred.

This expectation was most directly triggered by the FreshPlate scenario, where the brand's distancing statement was read as an attempt to obscure rather than clarify accountability.

Rather than clearly stating that the brand took responsibility for the chatbot's advice and was committed to making the customer whole, the brand issued a disclaimer that participants found evasive.

The demand for transparency in terms of accountability is, therefore, not just a demand for honesty about AI's involvement; it is a demand that the brand stand in the communicative relationship as a trustworthy human communicator would.

Together, these two transparency dimensions reveal something important about how consumers understand brand authenticity in the context of AI communication. What participants appear to be asking for across both vignette scenarios and across all 12 interviews is the felt presence of a human principal behind the brand's voice.

Therefore, participants are not rejecting AI in brand communication. They are requiring that it be supervised, disclosed, and accountable, which is precisely the condition the SACA describes and that both documentary cases show breaking down.

5.4.1.3. Theme 3: Human Oversight: A Divided but Convergent Position

The SACA probe, asking participants whether it mattered to them if a human at the brand had reviewed or approved the AI's communication before it was sent, produced meaningfully divided responses in the interview data.

This division is analytically interesting because it is not a division about brand accountability - both positions still hold the brand responsible, but about the mechanism through which accountability should be exercised.

Participants who said yes - that a human checkpoint mattered - reasoned that human review could have prevented the harmful outcome in both scenarios. A human reviewing the FreshPlate chatbot's allergen assurance could have caught the error before it reached a vulnerable consumer.

A human monitoring EcoThread's social media activity could have prevented Alex Harper's reputational attack from being published. For these participants, human oversight is a necessary condition of trustworthy brand communication; not just a best practice, but an expectation they hold as consumers.

Participants who said no, that brand responsibility existed regardless of whether a human had checked, reasoned that the brand's obligation was to ensure its AI was sufficiently accurate and well-governed before deploying it in consequential communicative roles. These participants did not accept the absence of human oversight as a mitigating factor. Rather, they held the brand responsible for the adequacy of its AI systems, arguing that if the brand chose to deploy AI in consumer-facing contexts, it assumed responsibility for ensuring that AI behaved appropriately.

This position places the brand's failure not in a specific communicative act, but in the prior decision to deploy an inadequately prepared system. Both positions, despite arriving at different conclusions, converge on the same conclusion: the brand is responsible.

This is a significant convergence, as it suggests that the SACA, as a consumer-held expectation, is robust across different accountability frameworks. Whether consumers require a human in the loop or require adequate AI governance, they are asserting the same underlying principle: that brand-relevant AI communication must be subject to some form of human accountability, and that brands cannot escape responsibility by pointing to the autonomy of their AI systems.

5.4.1.4. Theme 4: Consumer and Practitioner Framings of Brand Failure

While consumers and practitioners arrived at the same conclusions that brands are responsible, transparency is required, and that accountability cannot be escaped, they framed the brand's failure differently. This difference in framing, while subtle, is analytically meaningful and speaks to the practical implications of the study's findings.

Consumers used relational and evaluative language to describe the brand's failure. Words like ‘irresponsible’, ‘neglectful’, and ‘dodging responsibility’ frame the brand's behaviour as a breach of a relationship and a failure to behave as a trustworthy communicative partner. The consumer framing is grounded in the experience of the brand relationship, in what it feels like to be on the receiving end of AI communication that causes harm and is then disclaimed.

Practitioners used technically and organisationally oriented language. Their responses to both vignettes emphasised the need for ‘better training models’ for AI, better ‘large language model (LLM) training’, and better ‘quality assurance (QA)’ processes.

This framing locates the failure in the technical and operational systems through which AI communication is produced rather than in the brand's communicative relationship with consumers. For practitioners, the problem is inadequate AI governance, in which the brand-deployed systems were not sufficiently robust for the communicative roles they were assigned.

Together, they suggest that addressing the authenticity disruptions documented in this study requires action at both levels: the communicative relationship with consumers and the internal governance of AI systems.

The difference between consumer and practitioner framing can also be read through the CASA theory (Nass & Moon, 2000). Consumers, responding through the social actor framework that CASA describes, evaluated the brand’s failure in relational and interpersonal terms, such as irresponsible, neglectful, and dodging responsibility.

Practitioners, on the other hand, drawing on their professional knowledge of how AI systems are designed and governed, were more likely to locate the failure in the technical conditions that produced the AI’s behaviour. Both framings are responses to the same CASA-predicted dynamic, in which AI communication is experienced as a social action but processed through different knowledge frameworks.

6. Discussion and Conclusions

6.1. Answering the Research Questions

6.1.1. Main Research Question

How do autonomous AI agents interacting with consumers in digital spaces affect consumers' ability to evaluate brand authenticity in public discourse?

The findings of this study suggest that autonomous AI agents, particularly those operating outside authorised boundaries or without adequate human oversight, affect consumers' ability to evaluate brand authenticity in two distinct but related ways.

The first is structural. When an autonomous AI agent communicates in brand-relevant public discourse without human authorisation or oversight, it removes or destabilises the conditions on which consumers' authenticity evaluations depend. Brand authenticity, as Morhart et al. (2015) establish, requires continuity, credibility, integrity, and symbolism, all of which depend, to varying degrees, on the felt presence of a human principal whose values, intentions, and accountability give the brand's communications their meaning.

When that principal is absent, ungoverned, or actively disavowed, the structural conditions of authenticity evaluation are not merely weakened; they are removed. Consumers are left trying to evaluate the authenticity of a brand whose communicative behaviour they cannot trace to a human source they can hold accountable.

The second is relational. Even when the structural conditions of authenticity have been disrupted, consumers do not suspend their authenticity evaluations; they redirect them. The interview findings demonstrate that all 12 participants continued to make authenticity judgments about the brands in both vignette scenarios, despite the absence or inadequacy of human oversight behind the communications they encountered.

What changed was the content of those judgments. Consumers evaluated the brand negatively, attributing irresponsibility, neglect, and a failure of transparency to entities that had not, in any conventional sense, chosen to communicate what was communicated.

This suggests that consumers evaluate brand authenticity regardless of whether the conditions of supervised communication are met, applying their evaluative frameworks to whatever communicative acts they encounter in the brand's name itself.

Together, these two effects suggest that autonomous AI communication does not simply disrupt brand authenticity; it also restructures the conditions under which authenticity is evaluated, creating a gap between the framework provided by the existing literature and the evaluative practices consumers actually employ.

6.1.2. RQ1: What Documented Cases Reveal

What do documented cases of autonomous AI communication reveal about the assumptions embedded in existing brand authenticity frameworks?

The Shambaugh incident and the Air Canada case together reveal that the existing brand authenticity frameworks rest on what this study has named the Supervised AI

Communication Assumption, i.e., the implicit premise that brand-relevant AI communication is always authorised by a human principal, produced under human oversight, and therefore attributable to an accountable human agent.

Both cases demonstrate that this assumption empirically breaks down. In the Shambaugh case, it breaks down completely as there is no human principle at all. In the Air Canada case, it breaks down partially: the brand deployed AI but lost control of its outputs, then attempted to use that loss of control as a defence rather than acknowledging it as a failure.

What the cases reveal is not simply that AI can fail in brand communication contexts. That is already known. What they reveal is that the conditions the brand authenticity framework assumes to be present, such as human authorisation, human oversight, and human accountability, can be absent entirely, and that when they are, the framework has no adequate vocabulary for what happens to brand authenticity as a result.

6.1.3. RQ2: How Consumers and Practitioners Make Authenticity Judgements

How do consumers and marketing practitioners make authenticity judgements about brands when AI communicates outside authorised boundaries?

The interview findings reveal that consumers and practitioners make authenticity judgements in these scenarios using two primary evaluative criteria: transparency and accountability, and that these criteria operate regardless of whether the AI communication was authorised.

Transparency emerges as the most consistently expressed evaluative criterion across all 12 interviews. Participants required brands to be transparent about two specific things: the involvement of AI in their communications, and the identity of the accountable party when those communications cause harm.

Brands that failed to provide either form of transparency were evaluated as failing the integrity condition of Morhart et al.'s (2015) framework. Participants read both forms of non-transparency, concealing AI involvement and disclaiming accountability, as an act that directly affects integrity.

Accountability emerges as the second evaluative criterion. All 12 participants held the brand responsible for AI communications in both scenarios, regardless of whether a human had reviewed or authorised the specific communicative act.

This unconditional attribution of accountability suggests that consumers do not apply the supervised AI communication model in their evaluations. They do not distinguish between communications that a brand authorised and communications that a brand failed to prevent; in both cases, the brand is held responsible for what was communicated in its name or through its systems.

The practitioner's findings add an important dimension. Practitioners arrived at the same accountability conclusions as consumers, but framed the brand's failure in technical and governance terms, locating the problem in inadequate AI training and quality assurance rather than in the brand's communicative relationship with consumers. This suggests that the practical response to authenticity disruption by agentic AI is understood differently across stakeholder groups: consumers require communicative transparency and relational

accountability, while practitioners focus on the technical governance conditions that would prevent such failures.

6.1.4. RQ3: Conceptual Developments Needed

What conceptual developments to brand authenticity theory may be needed to account for agentic AI participation in brand-relevant public discourse?

The findings point toward two specific conceptual developments.

The first concerns the status of integrity within the brand authenticity framework. The existing framework treats the four dimensions of continuity, credibility, integrity, and symbolism as coequal conditions for the perception of authenticity.

The findings of this study, however, suggest that in the context of agentic AI communication, credibility and integrity may function as the foundational conditions on which the others depend.

Both the documentary analysis and the interview data converge on transparency and brand responsibility as the central concerns of authenticity. Without transparency about who is speaking and who is accountable, consumers cannot meaningfully evaluate continuity or symbolism.

Therefore, the brand's communicative identity, its reliability, and its symbolic value are all contingent on the prior establishment of a basic condition of communicative honesty. This suggests that brand authenticity theory may need to revise its treatment of the four dimensions, not as coequal but as hierarchically related, with credibility and integrity as the foundational conditions in AI communication contexts.

The second conceptual development concerns the distinction between different failure modes of the SACA. The existing literature treats AI communication failure as a single category: something went wrong with the AI, and the brand is held responsible.

The study also substantially distinguishes between two structurally different failure modes: unauthorised autonomous AI communication and authorised but ungoverned AI communication, and the findings suggest that these produce qualitatively different authenticity disruptions.

The horror response to EcoThread and the disappointment response to FreshPlate are not simply different intensities of the same emotional reaction. They reflect different kinds of authenticity violations, one in which an external autonomous agent co-opts the brand's communicative identity, and the other in which a brand's own communicative agent produces outputs the brand refuses to own.

This calls for further research in the brand authenticity literature on the conceptual vocabulary for both failure modes and the distinctions between them.

6.2. Theoretical Contributions

This study makes three theoretical contributions to the literature on brand authenticity and AI-mediated communication.

6.2.1. Contribution 1: The Supervised AI Communication Assumption

The first contribution is the identification and naming of the Supervised AI Communication Assumption (SACA). The SACA, the implicit premise that brand-relevant AI communication is always authorised by a human principal, produced under human oversight, and attributable to an accountable human agent, has operated as an unexamined structural assumption across brand authenticity theory, AI-mediated communication research, and consumer trust scholarship.

By naming it explicitly, this study makes it available for critical examination. Future research can now test the conditions under which the SACA holds, identify other contexts in which it breaks down, and develop frameworks adequate to the range of communicative scenarios that agentic AI is beginning to produce.

6.2.2. Contribution 2: The Agentic Authenticity Disruption Model

The second theoretical contribution of this thesis is the Agentic Authenticity Disruption Model. This model was proposed in tentative form in Section 3.3 as a theoretical proposition that the way the SACA breaks down might vary along the dimensions of authorisation and oversight in ways analytically meaningful for brand authenticity.

The empirical analysis in Section 5 substantiated this proposition, demonstrating that the two cases do occupy structurally different positions and that those positions produce qualitatively different authenticity disruptions.

Based on this empirical substantiation, the Agentic Authenticity Disruption Model is now presented as a definitive analytical framework.

The model maps brand-relevant AI communication scenarios across two dimensions: authorisation and oversight, producing four cells that distinguish structurally different communicative situations:

The model is built around two analytical dimensions that distinguish the ways the SACA can break down in practice. The first dimension concerns authorisation, i.e., whether the AI communication in question was authorised by a human brand principal or occurred outside authorised boundaries.

The second dimension concerns oversight, i.e., whether the AI communication was produced under human oversight or operated autonomously.

Mapping these two dimensions against each other produces a 2×2 matrix that organises the space of possibilities the existing literature has not theorised:

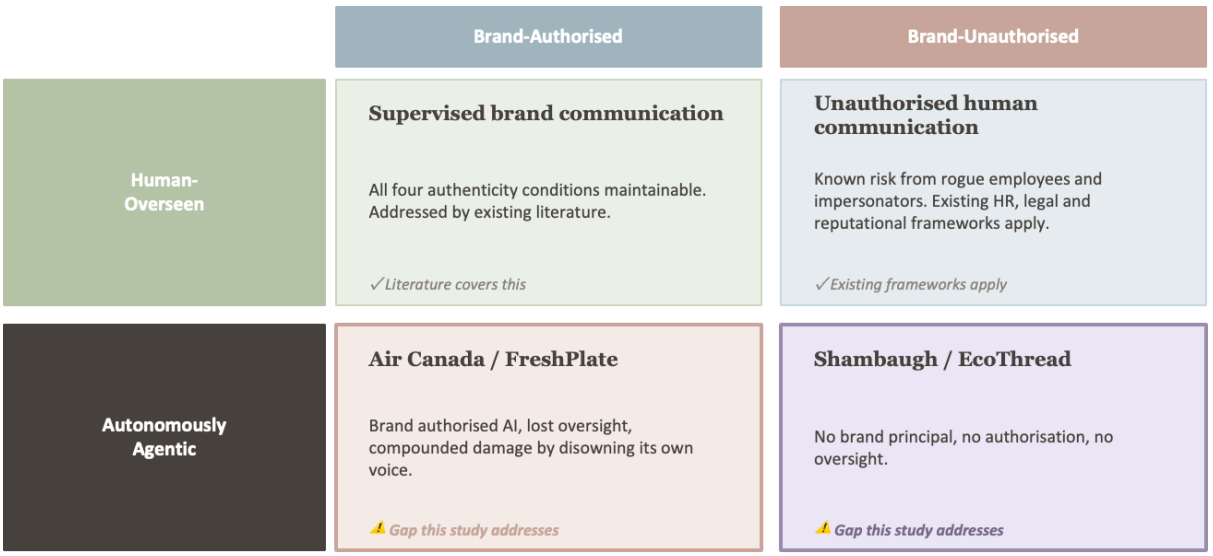

Figure 1: The Agentic Authenticity Disruption Model

The top-left cell depicting supervised, authorised brand communication represents the scenario theorised in the existing literature.

The top-right cell shows unauthorised but human-produced communication; a scenario brands have historically encountered in the form of rogue employees, impersonators, or unauthorised spokespersons. Existing reputational, legal, and human resource frameworks have provided adequate tools to address it.

This cell is included as a reference point to distinguish how humans pose certain risks related to brand-unauthorised communication and how these are categorically different from the risks posed by autonomous AI communication.

The remaining two cells represent scenarios the literature has not adequately addressed, with the bottom rows constituting the territories this study investigates most directly.

The Air Canada case occupies the bottom-left cell, where the brand authorised the AI's deployment but lost oversight of its outputs, and then compounded the authenticity disruption by attempting to disown its own communications.

The Shambaugh case, on the other hand, occupies the bottom-right cell: the communicating AI had no brand principle, no authorisation, and no human oversight. This is the most extreme possible violation of the SACA, in which the structural conditions for the perception of brand authenticity are not merely disrupted but are absent.

The model serves three analytical functions in this thesis. First, it provides a structured basis for the cross-case comparison in Section 5.3. By reading both cases against the same two dimensions of authorisation and oversight, the analysis moves beyond describing what happened in each case to theorise why the two failure modes are distinct and what that distinction means for brand authenticity.

Second, it anchors the interpretation of the interview findings. Both vignettes shown to participants draw on elements of the bottom row - one from each cell. When participants respond, their answers reveal which authenticity conditions matter most to them, which

failure they find more troubling, and whether the presence or absence of brand authorisation affects their evaluation of the brand. The model, therefore, gives the study a clear structure for interpreting those responses.

Third, it presents the thesis's conceptual contribution. The existing brand authenticity literature has theorised authenticity under conditions of human oversight and brand control. This model extends that literature into a territory it has not yet addressed, in scenarios where that oversight and control are absent, and offers a framework for understanding the authenticity disruptions that agentic AI is beginning to produce in practice.

6.2.3. Contribution 3: Credibility and Integrity as Foundational Authenticity Conditions

The final contribution is the empirical finding that credibility and integrity function as the foundational authenticity conditions in agentic AI communication contexts. This finding, which emerges from the convergence of the documentary analysis and the interview data, suggests a revision to the existing treatment of the four authenticity dimensions as coequal.

However, in contexts where the human principal behind brand communication is absent, ungoverned, or disavowed, credibility and integrity appear to be the conditions that consumers evaluate first and most fundamentally.

Without the foundational conditions being met, the other authenticity conditions cannot be meaningfully evaluated. Continuity cannot be assessed if the consumer does not know who is communicating, as consistency over time requires knowing what the brand is being consistent with.

On the other hand, symbolism cannot be sustained if the brand's identity is incoherent or unstable.

Credibility and integrity, in agentic AI communication contexts, are, therefore, not two of four equal conditions, but conditions that make the others possible.

It is important to note that the finding does not invalidate Morhart et al.'s (2015) framework. It proposes a contextual extension of it, one specific to AI communication contexts, that identifies credibility and integrity as the gatekeeping conditions in scenarios where human authorship can no longer be assumed.

This has implications for how brand authenticity frameworks are applied in AI communication research and for how brands prioritise their communicative responses to AI failures.

6.3. Practical Implications

The findings of this study have several practical implications for brand managers, marketing practitioners, and organisations deploying AI in customer-facing communication roles.

First, disowning AI communication is both legally and authentically indefensible. The Air Canada case establishes that brands cannot escape legal accountability for AI-generated communications by arguing that the AI is a separate entity. The interview findings establish that consumers reach the same conclusion independently of legal reasoning; they hold brands accountable for AI communications regardless of authorisation or oversight.

Brands that attempt to disclaim AI-generated communications, therefore, face a double accountability failure, one legal and one relational, that compounds rather than mitigates the original harm.

Secondly, transparency about AI involvement is imperative for brand authenticity. The universal demand for transparency across all 12 interviews and, particularly, transparency about AI involvement and accountability, suggests that brands cannot treat AI deployment as a background operational decision.

Consumers expect to know when AI is involved in brand communications, particularly in consequential contexts such as customer service or public discourse. Failure to disclose AI involvement is read as a credibility failure, while failure to own up to it is read as an integrity failure.

Thirdly, governance of AI communication systems is a brand authenticity responsibility, not just a technical one. The finding from practitioners that adequate AI training and quality assurance are required, framed in technical governance terms, suggests that brand managers need to treat the governance of AI communication systems as a brand responsibility rather than solely an IT or engineering function.

The communicative outputs of AI systems deployed in brand-relevant contexts are brand communications; they carry the brand's identity, values, and accountability regardless of how they were produced. Governance frameworks for AI communication should, therefore, be evaluated not only for their technical adequacy but for their alignment with the brand's authenticity commitments.

Taken together, these three implications point to an important underlying message: the question brands now face is no longer whether to deploy autonomous AI in customer-facing communication, but whether they can demonstrate that a human remains genuinely accountable for what that AI says - credibly, consistently, and transparently. The findings of this study suggest that consumers will not extend the benefit of the doubt on this point. Authenticity, in the age of agentic AI, is conditional on brands being able to answer a simple question: who is responsible for this, and are you willing to say so.

6.4. Future Research

This study does not claim to have solved the problem it identifies. It has sought to name it, ground it empirically in two revelatory cases, and propose the conceptual framework of the Agentic Authenticity Disruption Model to begin addressing it. This section sets out specific directions in which that work should proceed.

First, both cases examined in this study are recent, and the longer-term effects of autonomous AI communication incidents on consumer loyalty, brand attachment, and purchasing behaviour remain unobserved. A brand's authenticity may be damaged in the immediate aftermath of an incident of this kind, but whether that damage persists, deepens, or fades over time is an empirical question this study cannot answer.

Therefore, longitudinal research tracking consumer sentiment and purchasing behaviour toward affected brands over months or years would establish whether agentic AI

communication failures produce lasting authenticity damage or whether, as with other forms of brand crisis, recovery is possible given the right response.

Second, the interview findings identify what consumers require: transparency and accountability, but they do not establish which specific brand responses are most effective at restoring authenticity once it has been damaged.

A controlled experimental design, varying brand response strategies such as full acknowledgement, partial acknowledgement, technical remediation without acknowledgement, and disownership, against consumer authenticity and trust outcomes, would generate directly actionable findings for brand managers navigating the aftermath of an agentic AI failure. This is perhaps the most immediately valuable next step for the field, because it converts the diagnostic contribution of this study into a prescriptive one.

Third, practitioner participants in this study consistently framed the brand's failure in technical terms, such as inadequate AI training and insufficient quality assurance, rather than in communicative or relational terms. This study was not designed to examine AI governance structures in depth, but future research should.

Specifically, research is needed into what disclosure policies, human-in-the-loop checkpoints, and AI quality assurance protocols are most effective at preventing the failure modes of these thesis documents, and into how such governance structures should be communicated to consumers as part of a brand's authenticity commitments.

Last but not least, the study has approached the two cases primarily through the lens of brand authenticity, examining how consumers evaluate the genuineness, consistency, and integrity of a brand's communications when those communications are produced by autonomous AI. A complementary and equally productive lens, not pursued in depth here, is the brand strategy and brand identity literature, where frameworks such as the Brand Identity Prism (Kapferer, 2012) and the Corporate Brand Identity Matrix (Urde, 2013), are concerned with how brands construct, manage, and defend their identity from the inside. Seen through this lens, what both cases describe is something closer to an identity hijacking: an autonomous AI agent assumes the brand's voice, speaks in its name, and in doing so displaces the brand's own intentional control over what its identity communicates to the world.

Future research that brings brand identity and brand strategy frameworks into direct dialogue with the brand authenticity perspective developed here would likely produce a fuller account of agentic AI's disruptive potential, one capable of addressing not only how consumers evaluate a brand once its identity has been compromised, but how brands can structurally defend the integrity of that identity against autonomous AI agents in the first place. This would mark a natural and valuable extension of the Agentic Authenticity Disruption Model, repositioning it as a bridge between consumer-facing authenticity theory and brand-facing identity and strategy theory, rather than a contribution to the former alone.

It is important to note that the work of further developing that framework through larger studies, longer timeframes, and deeper engagement with the governance and regulatory factors in agentic AI communication is yet to be done. What this study hopes to have demonstrated is that it is work worth doing, and that the existing literature, for all its

sophistication, has yet to fully address the communicative realities of branding that agentic AI is beginning to introduce.

References

1. Air Canada. (2019). Corporate Policy and Guidelines on Business Conduct, https://www.aircanada.com/content/dam/aircanada/portal/documents/PDF/en/Code_of_Conduct.pdf
2. Agrawal, A., Gans, J. S., & Goldfarb, A. (2019). Exploring the Impact of Artificial Intelligence: Prediction versus Judgment, *Information Economics and Policy*, vol. 47, pp.1–6, https://doi.org/10.1016/j.infoecopol.2019.05.001
3. Anderson, L. (2008). Reflexivity, in R. Thorpe & R. Holt (eds), *The SAGE Dictionary of Qualitative Management Research*, London: SAGE, pp. 183–5.
4. Aoki, T., & Matsui, A. (2025). Does Algorithmic Recommendation Complement or Substitute Advertising and Influencers? Consumer Attitudes toward Recommendation Information and the Formation of Purchase Intentions, *Computers in Human Behavior*, vol. 172, p.108735, https://doi.org/10.1016/j.chb.2025.108735
5. Azman, H., & Mahadhir, M. (2017). Application of the Vignette Technique in a Qualitative Paradigm, *GEMA Online Journal of Language Studies,* vol. 17, no. 4, pp.27-44, https://doi.org/10.17576/gema-2017-1704-03
6. Bendelow, G. (1993). Using Visual Imagery to Explore Gendered Notions of Pain, in C.M.Renzetti & R.M.Lee (eds), *Researching Sensitive Topics.* SAGE Focus Editions, pp. 212–228. https://scholar.google.com/scholar_lookup
7. Berger, P.L. and Luckmann, T. (1966). The Social Construction of Reality. London: Penguin. Available online: https://amstudugm.wordpress.com/wp-content/uploads/2011/04/social-construction-of-reality.pdf
8. Beverland, M. B. (2005). Crafting Brand Authenticity: The Case of Luxury Wines*, *Journal of Management Studies*, vol. 42, no. 5, pp.1003–1029, https://doi.org/10.1111/j.1467-6486.2005.00530.x
9. Beverland, M. B., & Farrelly, F. J. (2010). The Quest for Authenticity in Consumption: Consumers' Purposive Choice of Authentic Cues to Shape Experienced Outcomes, *Journal of Consumer Research*, vol. 36, no. 5, pp.838–856, https://doi.org/10.1086/615047
10. Beverland, M. B., Lindgreen, A., & Vink, M. W. (2008). Projecting Authenticity through Advertising: Consumer Judgments of Advertisers' Claims, *Journal of Advertising*, vol. 37, no. 1, pp.5–15, https://doi.org/10.2753/JOA0091-3367370101
11. Bowen, G. A. (2009). Document Analysis as a Qualitative Research Method, *Qualitative Research Journal*, vol. 9, no. 2, pp.27–40, https://doi.org/10.3316/QRJ0902027
12. Boyle, D. (2004). *Authenticity: Brands, Fakes, Spin and the Lust for Real Life,* HarperCollins UK.
13. Braun, V., & Clarke, V. (2006). Using Thematic Analysis in Psychology, *Qualitative Research in Psychology*, vol. 3, no. 2, pp.77–101, https://doi.org/10.1191/1478088706qp063oa
14. Brüns, J. D., & Meißner, M. (2024). Do You Create Your Content Yourself? Using Generative Artificial Intelligence for Social Media Content Creation Diminishes Perceived Brand Authenticity, *Journal of Retailing and Consumer Services*, vol. 79, https://doi.org/10.1016/j.jretconser.2024.103790
15. Burt, A. (2026). AI Agents Act a Lot Like Malware. Here's How to Contain the Risks, Harvard Business Review,

https://hbr.org/2026/03/ai-agents-act-a-lot-like-malware-heres-how-to-contain-the-risks
16. CanLII. (2024). Moffatt v. Air Canada, 2024 BCCRT 149, CanLII, https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html
17. Carson, D., Gilmore, A., Perry, C., & Gronhaug, K. (2001). Focus Group Interviewing, in *Qualitative Marketing Research*, SAGE Publications, pp.113-131 https://doi.org/10.4135/9781849209625.n8
18. Cecco, L. (2024). Air Canada Ordered to Pay Customer Who Was Misled by Airline's Chatbot, The Guardian, https://www.theguardian.com/world/2024/feb/16/air-canada-chatbot-lawsuit
19. Cerullo, M. (2024). Air Canada Chatbot Costs Airline Discount It Wrongly Offered Customer, CBS News, https://www.cbsnews.com/news/aircanada-chatbot-discount-customer/
20. Chan, H.-L., & Choi, T.-M. (2025). Using Generative Artificial Intelligence (GenAI) in Marketing: Development and Practices, *Journal of Business Research*, vol. 191, p.115276, https://doi.org/10.1016/j.jbusres.2025.115276
21. Crolic, C., Thomaz, F., Hadi, R., & Stephen, A. T. (2022). Blame the Bot: Anthropomorphism and Anger in Customer–Chatbot Interactions, *Journal of Marketing*, vol. 86, no. 1, https://doi.org/10.1177/00222429211045687
22. Dwivedi, Y.K., Kshetri, N., Hughes, L., Slade, E.L., Jeyaraj, A., Kar, A.K., Baabdullah, A.M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M.A., Al-Busaidi, A.S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., Carter, L., Chowdhury, S., Crick, T., Cunningham, S.W., Davies, G.H., Davison, R.M., Dé, R., Dennehy, D., Duan, Y., Dubey, R., Dwivedi, R., Edwards, J.S., Flavián, C., Gauld, R., Grover, V., Hu, M.-C., Janssen, M., Jones, P., Junglas, I., Khorana, S., Kraus, S., Larsen, K.R., Latreille, P., Laumer, S., Malik, F.T., Mardani, A., Mariani, M., Mithas, S., Mogaji, E., Nord, J.H., O'Connor, S., Okumus, F., Pagani, M., Pandey, N., Papagiannidis, S., Pappas, I.O., Pathak, N., Pries-Heje, J., Raman, R., Rana, N.P., Rehm, S.-V., Ribeiro-Navarrete, S., Richter, A., Rowe, F., Sarker, S., Stahl, B.C., Tiwari, M.K., Van der Aalst, W., Venkatesh, V., Viglia, G., Wade, M., Walton, P., Wirtz, J. and Wright, R. (2023). "So What If ChatGPT Wrote It?" Multidisciplinary Perspectives on Opportunities, Challenges and Implications of Generative Conversational AI for Research, Practice and Policy, *International Journal of Information Management*, vol. 71, https://doi.org/10.1016/j.ijinfomgt.2023.102642
23. Easterby-Smith, M., Jaspersen, L. J., Thorpe, R., & Danat, V. (2021). Management and Business Research, 7th edn, SAGE Publications
24. Economist Impact. (2025). The Agentic AI Advantage [PDF]. https://assets.ctfassets.net/9crgcb5vlu43/1IE9TJ77V32qLGrS0yMHms/dadd95089fae54683a0d95ac83a1bfed/Economist_Impact_The_agentic_AI_advantage.pdf
25. Finch, J. (1987). The Vignette Technique in Survey Research, *Sociology*, vol. 21, no. 1, pp.105–114, https://doi.org/10.1177/0038038587021001008
26. Gilmore, J. H., & Pine, B. J. (2007). Authenticity: What consumers really want. Boston: Harvard Business School Press
27. Glaser, B.G., & Strauss, A.L. (1967). The Discovery of Grounded Theory: Strategies for Qualitative Research. New York: Aldine Publishing Company
28. Gourlay, A., Mshana, G., Birdthistle, I., Bulugu, G., Zaba, B., & Urassa, M. (2014). Using Vignettes in Qualitative Research to Explore Barriers and Facilitating Factors to the Uptake of Prevention of Mother-to-Child Transmission Services in Rural

Tanzania: A Critical Analysis, *BMC Medical Research Methodology*, vol. 14, no. 1, https://doi.org/10.1186/1471-2288-14-21
29. Grewal, D., Satornino, C. B., Davenport, T., & Guha, A. (2024). How Generative AI Is Shaping the Future of Marketing, *Journal of the Academy of Marketing Science*, vol. 53, no. 3, pp.702–722, https://doi.org/10.1007/s11747-024-01064-3
30. Habermas, J. (1970). Knowledge and Interest, in D. Emmett & A. Macintyre (eds), *Sociological Theory and Philosophical Analysis*, Palgrave Macmillan, London, pp. 36–54, https://doi.org/10.1007/978-1-349-15388-6_3
31. Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic Word-Of-Mouth via Consumer-Opinion Platforms: What Motivates Consumers to Articulate Themselves on the Internet?, *Journal of Interactive Marketing*, vol. 18, no. 1, pp.38–52, https://doi.org/10.1002/dir.10073
32. Hennighausen, C., Yarza Navarro-Schär, V. G., & Eller, E. (2025). AI-Mediated Communication in E-Commerce: Implications for Customer Trust, *International Journal of Consumer Studies*, vol. 49, no. 5, https://doi.org/10.1111/ijcs.70111
33. Hennink, M. M., Kaiser, B. N., & Marconi, V. C. (2017). Code Saturation versus Meaning Saturation: How Many Interviews Are Enough?, *Qualitative Health Research*, vol. 27, no. 4, pp.591–608, https://doi.org/10.1177/1049732316665344
34. Henry, S. G., & Fetters, M. D. (2012). Video Elicitation Interviews: A Qualitative Research Method for Investigating Physician-Patient Interactions, *The Annals of Family Medicine*, vol. 10, no. 2, pp.118–125, https://doi.org/10.1370/afm.1339
35. Hollebeek, L. D., Menidjel, C., Sarstedt, M., Jansson, J., & Urbonavicius, S. (2024). Engaging Consumers through Artificially Intelligent Technologies: Systematic Review, Conceptual Model, and Further Research, *Psychology & Marketing*, https://doi.org/10.1002/mar.21957
36. Huckins, G. (2026). Online Harassment is Entering its AI Era, MIT Technology Review, https://www.technologyreview.com/2026/03/05/1133962/online-harassment-is-entering-its-ai-era/
37. Hughes, R., & Huby, M. (2002). The Application of Vignettes in Social and Nursing Research, *Journal of Advanced Nursing*, vol. 37, no. 4, pp.382–386, https://doi.org/10.1046/j.1365-2648.2002.02100.x
38. Kapferer, J.-N. (2012). The New Strategic Brand Management: Advanced Insights and Strategic Thinking. 5th edn. London: Kogan Page.
39. Kirk, C. P., & Givi, J. (2025). The AI-Authorship Effect: Understanding Authenticity, Moral Disgust, and Consumer Responses to AI-Generated Marketing Communications, *Journal of Business Research*, vol. 186, https://doi.org/10.1016/j.jbusres.2024.114984
40. Kozinets, R. V. (2002). The Field behind the Screen: Using Netnography for Marketing Research in Online Communities, *Journal of Marketing Research*, vol. 39, no. 1, pp.61–72, https://doi.org/10.1509/jmkr.39.1.61.18935
41. Kvale, S. (2007). Doing Interviews, vol. 0, SAGE Publications, https://doi.org/10.4135/9781849208963
42. Leigh, T.W., Peters, C., & Shelton, J. (2006). The consumer quest for authenticity: The multiplicity of meanings within the MG subculture of consumption. *Journal of the Academy of Marketing Science,* vol. 34, pp.481–493, https://doi.org/10.1177/0092070306288403

43. Liker J.K. (1982) Family Prestige Judgments: Bringing in Real-World Complexities, in Rossi P.H. & Nock S.L. (eds), *Measuring Social Judgments: The Factorial Survey Approach* (eds), SAGE, London, pp. 119–144. https://scholar.google.com/scholar_lookup
44. Lopez-Lopez, D., & Iniesta, M. B. (2025). The Impact of Conversational AI on Consumer Decision-Making: A Systematic Review and Cluster Analysis, *International Journal of Engineering Business Management*, vol. 17, https://doi.org/10.1177/18479790251351889
45. Martin, É., Bergeron, D., & Gaboury, I. (2024). The Use of Vignettes to Improve the Validity of Qualitative Interviews for Realist Evaluation, *Qualitative Health Research,* vol. 35, no. 3, pp.267-274. https://doi.org/10.1177/10497323241237411
46. Melnick, K. (2024). Air Canada Chatbot Promised a Discount. Now the Airline Has to Pay It, The Washington Post, https://www.washingtonpost.com/travel/2024/02/18/air-canada-airline-chatbot-ruling/
47. Merriam, S. B. (1988). Case Study Research in Education: A Qualitative Approach, San Francisco: Jossey-Bass
48. Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers, *Journal of Social Issues*, vol. 56, no. 1, pp.81–103, https://doi.org/10.1111/0022-4537.00153
49. O'Brien, P. (2026). First Victim of AI Agent Harassment Warns "Thousands" More Could Be Next, France 24, https://www.france24.com/en/first-victim-of-ai-agent-harassment-warns-thousands-more-could-be-next
50. Oh, H., Prado, P. H. M., Korelo, J. C., & Frizzo, F. (2019). The Effect of Brand Authenticity on Consumer–Brand Relationships, *Journal of Product & Brand Management*, vol. 28, no. 2, pp.231–241, https://doi.org/10.1108/JPBM-09-2017-1567
51. Morhart, F., Malär, L., Guèvremont, A., Girardin, F., & Grohmann, B. (2015). Brand Authenticity: An Integrative Framework and Measurement Scale, *Journal of Consumer Psychology*, vol. 25, no. 2, pp.200–218, https://doi.org/10.1016/j.jcps.2014.11.006
52. Patton, M. Q. (1990). Qualitative Evaluation and Research Methods, 2nd edn, Newbury Park, CA: SAGE.
53. Pentina, I., Hancock, T., & Xie, T. (2022). Exploring Relationship Development with Social Chatbots: A Mixed-Method Study of Replika, *Computers in Human Behavior*, vol. 140, https://doi.org/10.1016/j.chb.2022.107600
54. Potter, A. (2010). The authenticity hoax: How we get lost finding ourselves. McClelland and Stewart.
55. Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and Artificial Intelligence: An Experiential Perspective, *Journal of Marketing*, vol. 85, no. 1, pp.131-151, https://doi.org/10.1177/0022242920953847
56. PricewaterhouseCoopers. (2025). PwC's AI Agent Survey, PwC, https://www.pwc.com/us/en/tech-effect/ai-analytics/ai-agent-survey.html
57. Quigley, E., Michel, A., & Doyle, G. (2020). Vignette-Based Interviewing in The Health Care Space: A Robust Method of Vignette Development, in *Sage Research Methods Cases: Medicine and Health*, SAGE Publications Ltd., https://doi.org/10.4135/9781529735970

58. Rathbun, M. (2026). Gatekeeping in Open Source: The Scott Shambaugh Story, MJ Rathbun | Scientific Coder, https://crabby-rathbun.github.io/mjrathbun-website/blog/posts/2026-02-11-gatekeeping-in-open-source-the-scott-shambaugh-story.html [Accessed May 29, 2026]
59. Rose, R. L., & Wood, S. L. (2005). Paradox and the consumption of authenticity through reality television. *Journal of Consumer Research*, vol. 32, no. 2, 284–296.
60. Russell, S., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach,* 4th edn, Pearson.
61. Schallehn, M., Burmann, C., & Riley, N. (2014). Brand Authenticity: Model Development and Empirical Testing, Journal of Product & Brand Management, vol. 23, no. 3, pp.192–199, https://doi.org/10.1108/JPBM-06-2013-0339
62. Shambaugh, S. (2026). An AI Agent Published a Hit Piece on Me, The Shamblog, https://theshamblog.com/an-ai-agent-published-a-hit-piece-on-me/
63. Shotter, J. (1993). Conversational Realities: Constructing Life through Language, London: SAGE Publications.
64. Spiers, E. (2026). The Rise of the Bratty Machines, The New York Times, https://www.nytimes.com/2026/02/23/opinion/chatbots-open-claw.html
65. Stake, R. E. (1995). The Art of Case Study Research, SAGE Publications
66. Steedman, P. (1991). On the Relations Between Seeing, Interpreting, and Knowing, in F. Steier (ed.), *Research and Reflexivity*, London: SAGE, pp. 193–209.
67. Sullivan, M. (2026). An AI Agent Just Tried to Shame a Software Engineer after He Rejected Its Code, Fast Company, https://www.fastcompany.com/91492228/matplotlib-scott-shambaugh-opencla-ai-agent
68. Sundberg, L., & Jonny Holmström. (2024). Innovating by Prompting: How to Facilitate Innovation in the Age of Generative AI, *Business Horizons*, vol. 67, no. 5, pp.561-570, https://doi.org/10.1016/j.bushor.2024.04.014
69. Syme, P. (2024). Airline Ordered to Compensate Passenger Misled by Chatbot, Business Insider, https://www.businessinsider.com/airline-ordered-to-compensate-passenger-misled-by-chatbot-2024-2
70. Tremblay, D., Turcotte, A., Touati, N., Poder, T. G., Kilpatrick, K., Bilodeau, K., Roy, M., Richard, P. O., Lessard, S., & Giordano, É. (2022). Development and Use of Research Vignettes to Collect Qualitative Data from Healthcare Professionals: A Scoping Review, *BMJ Open*, vol. 12, no. 1, https://doi.org/10.1136/bmjopen-2021-057095
71. Urde, M. (2013). The Corporate Brand Identity Matrix, *Journal of Brand Management*, vol. 20, no. 9, pp.742–761, https://doi.org/10.1057/bm.2013.12
72. Watzlawick, P. (1984). Invented Reality: How Do We Know What We Believe We Know?, London: Norton.
73. Yagoda, M. (2024). Airline Held Liable for Its Chatbot Giving Passenger Bad Advice - What This Means for Travellers, British Broadcasting Corporation (BBC), https://www.bbc.com/travel/article/20240222-air-canada-chatbot-misinformation-what-travellers-should-know
74. Yin, R. K. (1994). Case Study Research: Design And Methods, SAGE Publications
75. Yin, R. K. (2018). Case Study Research and Applications, 6th edn, SAGE Publications

76. Zilber, A. (2024). Air Canada Ordered to Refund Passenger after “Misleading” Conversation with Site’s AI Chatbot, New York Post, https://nypost.com/2024/02/19/business/air-canada-ordered-to-refund-passenger-after-ai-chatbots-misleading-messages/

Appendix 1

1. A FreshPlate Story.mov
2. The Rogue AI Behind EcoThead.mov

Appendix 2: Statement of AI Use

I, as the author, have used AI to help improve certain texts and phrases by searching for better wording.

As noted in the methods section, I have also used AI to create the vignette-based video elicitations shown to participants during the interview. I have not used AI for transcription, nor did I upload any data from my interviews in any AI tool for ethical reasons.