Course:	SKOM12
Term:	Spring 2023
Supervisor	Marléne Wiggill
Examiner	Howard Nothhaft

Strategic Communication of Trustworthiness in Autonomous Systems, Machine Learning, & AI

MEGAN ROLLERSON

Lund University
Department of strategic communication
Master's thesis

Abstract

Strategic Communication of Trustworthiness in Autonomous Systems, Machine Learning, & AI

Autonomous systems, machine learning, and artificial intelligence are continuously integrated into our everyday lives. Accompanying this trend is a renewed distrust and concern that technological development may outpace creator control with deleterious results for society. The trustworthiness of the systems and the organisations that operate them is gaining greater attention in public discourse, mass media, and the academic community. In this evolving landscape, the field of strategic communication is confronted with an increasingly relevant question: How can organisations communicate the trustworthiness of autonomous systems effectively and strategically to foster trust in the technology and thus support a successful organisation-public relationship? Through the lenses of actional legitimacy and discourse of renewal theories, as well as trust repair discourse, this paper analyses Microsoft's pre-crisis communication and crisis response for their new AI-powered Bing search engine and chatbot, launched in February 2023. A qualitative research approach that integrates process tracing and discourse tracing methodologies is used to evaluate the outcome of Microsoft's crisis response and analyse their discourse for six identified trust dimensions: integrity, competence, predictability, benevolence, anthropomorphism, and human oversight. From this study, I propose a framework for communicating trust prior to and in response to crises that specifically addresses the strategic communication of autonomous systems, machine learning, and AI.

Keywords: strategic communication, trust, autonomous systems, artificial intelligence, machine learning

Word count: 19,542

Table of Contents

1. Introduction

The integration of autonomous systems (AS), machine learning (ML), and artificial intelligence (AI) [1] into society is accelerating at an unstoppable pace, and accompanying this digital transformation are the issues of fear, distrust, and uncertainty. Extant research by academics (Devitt, 2018; Falcone et al., 2001; Kaur & Rampersad, 2018; Prahl & Goh, 2021; Siau & Wang, 2018; Stilgoe & Cohen, 2021) and government agencies, such as the US Department of Defence (Atkinson, 2015), has identified trust as a significant obstacle to broad acceptance and adoption of these systems. Further, research by Fast and Horvitz (2017) indicates increasing fear of AI associated with loss of control, uncertainty over job security, and an absence of morality possessed by the technology. This apprehension may impact larger organisational goals, as distrust of a system may ultimately be reflected in a lack of public trust in the organisation that creates, owns, or operates the system. Distrust would, in turn, affect the organisation's relationship with its various publics (including users, government or regulatory agencies, and other stakeholders).

In strategic communication theory, trust is an integral element or characteristic of "successful relationships", alongside commitment and satisfaction (Hon & Grunig, 1999); and communication of trust or trustworthiness is critical to both establishing and rebuilding relationships (Shockley-Zalabak & Ellis, 2006). Falcone et al. (2001) suggest that there are different types of trust, including trust in infrastructure and information systems, and that their complementary relations are essential to trust in, among other things, human-computer interactions (p. 2). Therefore, the argument can be made that trust in autonomous systems should be considered from a communication perspective. Thus far, however, research on trust in autonomous systems, including the collected works found in Abbass et al.'s *Foundations of Trusted Autonomy* (2018) or articles by Atkinson (2015), Kaur and Rampersad (2018), and Liao and Sundar (2022), indicates a predominant focus on the building of public trust as a

[1] Throughout this paper, "autonomous systems" is used as an overarching term to refer to all autonomous systems, machine learning, artificial intelligence, and other autonomous algorithmic programming. While AI is arguably the more popularly used term, in literature and public discourse, both autonomous systems and artificial intelligence have been used as to refer to these systems. Nonetheless, as described in the literature review, this paper will side with the categorisation used by Devitt (2018), so as not to confuse current AI with the theorised concept of sentient AI.

function of the technological development and governance of the technology, rather than of the discourse surrounding it. Furthermore, primary trust dimensions of integrity, competence, and dependability, discussed in public relations and communication literature by Hon and Grunig (1999) and others (Paine, 2003; Shockley-Zalabak & Ellis, 2006), insufficiently reflect the scope of trust pertinent to autonomous systems, machine learning, and AI.

Factors of benevolence (Atkinson, 2015), anthropomorphism (Siau & Wang, 2018), and human oversight (Hagström, 2019; Taylor, 2021), otherwise considered beneficial in the technological development of autonomous systems, should be examined and theoretically accompany the dimensions traditionally associated with relationship-building when communicating about these systems. Autonomous systems, machine learning, and AI are unlike other goods and services offered or operated by organisations. They are designed to enhance, replicate, or replace human decision-making, knowledge synthesis, and intelligence without human input (Boulanin, 2019; Devitt, 2018). An expanded trust communication framework, which considers benevolence, anthropomorphism, and human oversight, may better serve communication practitioners in this field.

The challenge of achieving and maintaining trust is not unfounded. Distrust of autonomous systems, as distinct from the organisation, may be partly grounded in fear (Prahl & Goh, 2021, p. 6) and science fiction (Devitt, 2018, p. 168; Siau & Wang, 2018, p. 51), and occurs when "innovation is radical and complex" (Devitt, 2018, p. 172). Consider films such as 2001: A Space Odyssey and The Matrix trilogy, in which artificial intelligence develops autonomy beyond its creators with deleterious ends for the human protagonists. Conversely, distrust is also grounded in legitimate privacy, security, and responsibility concerns following real-world incidents of systems failures and AI biases (Prahl & Goh, 2021). Additionally, autonomous systems failures raise complex, divisive questions about legal (Hagström, 2019) and moral responsibility (Orr & Davis, 2020; Taylor, 2021). Siau and Wang (2018) suggest that anthropomorphising autonomous systems - imbuing them with human characteristics - is valuable in further enabling trust building (p. 50). This value comes from a sense of familiarity that helps form emotional and relational attachment (Devitt, 2018, p. 169; Siau & Wang, 2018, p. 51). In other words, it can be argued that acceptance of autonomous systems increases when they are anthropomorphised and perceived as trustworthy. Trust, in turn, may reduce concerns around liability, accountability and moral responsibility, bias, and safety. Therefore, applying a technological perspective of trust within strategic communication may assist organisations in overcoming the obstacle of distrust in autonomous systems.

This paper considers strategic communication of trust in two phases. The first phase, establishing trust, considers the strategic communication of trust as aligned with organisational

goals and values. Fostering public trust in an autonomous system should, by association, foster trust in the organisation. Through case analysis, the aim is to identify the use of trust dimensions in an organisation's discourse about their autonomous systems' use, development, or availability. The second phase, rebuilding trust, is of primary focus for the case analysis. This phase examines the communication of trustworthiness in a crisis response. That is, in terms of strategically communicating dimensions of trustworthiness to rebuild trust when autonomous systems fail and thus lead to negative consequences for society, or worse, result in loss of life. Identifying the communication of trust in successful crisis communication, such that the organisation-public relationship can be maintained or restored during a crisis, may theoretically validate its strategic value for communication practitioners. The idea of recommunicating trustworthiness in a crisis response and recovery campaign is similar to repeating corporate values, a communication tactic Falkheimer and Heide (2015) considered successful in the Findus Nordic horsemeat scandal (p. 144). Though Falkheimer and Heide (2015) speak of trust recovery as a goal or outcome, this paper, like that of Shockley-Zalabak and Ellis (2006), considers the communication of trustworthiness as part of the discourse and process to achieve the outcome of a successful organisation-public relationship.

1.1 Aim

The primary purpose of this paper is to investigate how organisations can use the strategic communication of trustworthiness to assist their stakeholders and various publics to overcome distrust of autonomous systems, machine learning, and AI. In order to accomplish this aim, this paper will first explore and compare the different approaches to trust from the perspectives of strategic communication and autonomous systems technology; then, examine through a methodical case analysis of Microsoft's new OpenAI-powered Bing chatbot how trust is communicated preceding and during a crisis; and, finally, develop a framework for how strategic communication can specifically establish trust in autonomous systems, machine learning, and AI technology. Long-term "resilient trust" is formed gradually (Siau & Wang, 2018, p. 47) and through established integrity (Shockley-Zalabak & Ellis, 2006, p. 49). A successful organisation-public relationship, with trust established prior to a crisis, better positions an organisation to overcome a crisis and rebuild trust (Ulmer et al., 2018, p. 318). Therefore, to understand how distrust is overcome, it is worthwhile to identify the use of trust dimensions in communication both preceding and during a crisis.

Through case evidence, this paper will consider trustworthiness discourse, through the dimensions of competence, integrity, predictability, benevolence, anthropomorphism, and human oversight, in relation to cause and outcome to answer two research questions. Based on

extant research on the role of trust in building successful organisation-public relationships, this paper asks the following questions:

> **RQ1:** How is the trustworthiness of autonomous systems, machine learning, and AI communicated by an organisation during a crisis?
>
> **RQ2:** How can organisations strategically communicate the trustworthiness of autonomous systems, machine learning, and AI to their various external publics to legitimate trust in the organisation-public relationship?

Using collected data, this qualitative study considers whether dimensions or factors of trust are present in external strategic communication used by an organisation in support of its autonomous systems. In the literature review, academic works from strategic communication and technology are compared and contrasted for their contributions to trust research. This comparison is supported by a discussion of autonomous systems more generally in order to establish a working definition and understanding of the technology. Also included is a discussion of moral responsibility in the event of autonomous systems failure. Extant research on trust in communication has focused on three primary dimensions: integrity, competence, and dependability (Hon & Grunig, 1999), among others, and includes honesty and transparency (Falkheimer & Heide, 2015; Shockley-Zalabak & Ellis, 2006). As a means to overcome distrust, this paper will synthesise the three dimensions of trust (integrity, competence, and dependability) discussed in work by Hon and Grunig (1999) with the additional dimensions of benevolence, anthropomorphism, and human oversight identified in literature on autonomous systems (Atkinson, 2015; Boulanin, 2019; Devitt, 2018; Hagström, 2019; Siau & Wang, 2018; Taylor, 2021). Sub-dimensions of transparency and honesty, identified in articles by Falkheimer and Heide (2015), Kim and Lee (2018), Shockley-Zalabak and Ellis (2006), and others, will also be included as elements of the primary dimensions. Anthropomorphism, reflected in other factors such as integrity and benevolence, is arguably of particular importance when communicating about autonomous systems, given that these systems can replace human decision-making in various applications. Comparatively, reassurance of human oversight and control, stemming from moral responsibility (Rahwan, 2018; Taylor, 2021), is a key consideration that emerges as beneficial for trust maintenance.

The theories of actional legitimacy and discourse of renewal, taken from crisis communication and in combination with trust repair discourse, form the foundation upon which Microsoft's discourse is analysed. The methodological approaches of process tracing and discourse tracing then build upon the theoretical foundation. An illustration of the separate

phases an organisation encounters when trust building and rebuilding, from pre-crisis through to crisis, to crisis response, and, finally, to the outcome, is formed in the methodology chapter. For analytical and methodological focus, the process tracing variant of theory-building was deemed appropriate. It employs empirical evidence, a theorised causal process, and the analyst's own insights. In this paper, theory-building facilitates understanding of a process by which trust in an autonomous system (and by association, the organisation) is legitimised before and during a crisis.

In support of the primary aim, this paper attempts to contribute to the field of strategic communication in two ways: 1) by demonstrating how a blended method of process tracing and discourse tracing can be used to study crisis communication discourse employed by organisations, and 2) by providing a tailored framework for communicating trustworthiness in autonomous systems, machine learning, and AI that uniquely addresses the trust-building challenges associated with autonomous technology, including concern for human oversight. A framework that incorporates trust dimensions specific to autonomous systems and applies them purposefully in communication may help reinforce the organisational goal of establishing the trust of various publics in these systems. This paper explores but one facet, discourse, in the complex interplay of trust and autonomous systems. The final chapter concludes with a further discussion of the findings, this paper's contributions to the field of strategic communication, and suggestions for future research.

2. Literature Review

The literature for this paper considers two separate approaches to trust, first, from the strategic communication perspective and second, from the technology perspective, as applied to autonomous systems, machine learning, and AI. This is done to find common ground, develop a strategic communication approach tailored to the unique complexity of autonomous systems, and consider trust not only as an outcome of communication but as part of the communicated discourse. The reason for communicating trust or trustworthiness of autonomous systems is that it should ultimately serve the organisational goal of creating public trust in the organisation itself. While trust is but one element strategic communication researchers consider important for establishing successful relationships between organisations and their various publics (Hon & Grunig, 1999; Paine, 2003), technology researchers consider trust as something that should be built into autonomous systems (Liao & Sundar, 2022). The purpose of looking at how trust dimensions are utilised in strategic communication is to first form a unique framework of how organisations working with autonomous systems use these dimensions to build trust and, second, how it is again used to rebuild trust in times of crisis. Finally, the subject of crisis brings up questions of moral and collective moral responsibility and how they apply to the organisation and its people. Responsibility and blame may be obfuscated in the context of autonomous systems, machine learning, and AI, where algorithms operate within a "black box" and beyond the scope of traceable decision-making (Devitt, 2018; Hagström, 2019; Rickli, 2019; Siau & Wang, 2018). In rebuilding trust, moral responsibility may be seen as accompanying the dimensions of competence and integrity.

2.1 Communicating Trustworthiness

Hon and Grunig (1999), in *Guidelines for Measuring Relationships in Public Relations*, outline three dimensions of trust: integrity, competence, and dependability. Dependability and predictability are terms found to be used interchangeably between communications and technological discussions of trust dimensions or factors. Public relations theorists, including Hon and Grunig (1999) and Paine (2003), focus on trust as a measurable outcome of successful communication behaviour. Later articles related to strategic communication by Falkheimer and Heide (2015), Kim and Lee (2018), Prahl and Goh (2021), and Shockley-Zalabak and Ellis

(2006) discuss the value of transparency and honesty as dimensions of trust. Shockley-Zalabak and Ellis's (2006) model representing dimensions of organisational trust differs from the earlier dimensions presented by Hon and Grunig (1999). As well as expanding upon the dimensions, they replaced integrity with honesty/openness. They propose the following five-dimension model of trust, which they encourage communication professionals to employ purposefully.

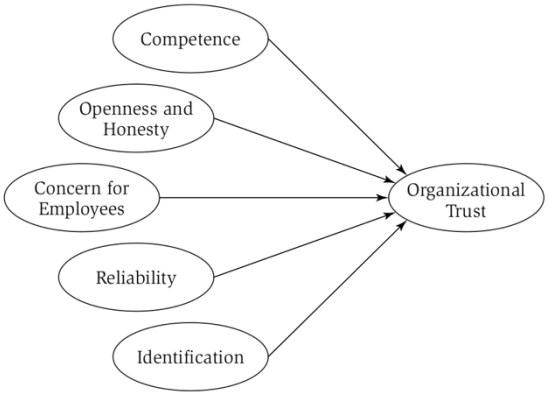

Figure 1. *Path Model of the Five Dimensions of Organizational Trust* (Shockley-Zalabak & Ellis, 2006, p. 48)

However, not all dimensions of this model relate to external communication nor autonomous systems. For one, the authors focus on internal and leadership communication directed at employees. Also, a review of the collected literature suggests that openness, honesty, and transparency are better categorised as sub-dimensions/elements of integrity. We see this sub-categorisation of honesty in the model by Devitt (2018, p. 166). Likewise, "care", discussed by Shockley-Zalabak and Ellis (2006) and presented as "concern for employees" (p. 49), can be associated with the factor of benevolence, discussed by Atkinson (2015).

Within the research field of strategic communication, Falkheimer and Heide (2018) explain that communication has a role in public and organisational trust building, as well as trust preservation in crisis communication. A key takeaway from Shockley-Zalabak and Ellis (2006), however, is their assertion that communication practitioners overlook trust. They tend to emphasise trust's fundamental importance but undervalue its influence on the behaviour of various organisational publics (p. 44). Both Falkheimer and Heide (2015) and Shockley-Zalabak and Ellis (2006) support the idea that trust should be tied to organisational values. For Kim and Lee (2018) and Prahl and Goh (2021), trust and transparency play vital roles in crisis communication and crisis following AI failures, respectively.

These works will help bridge together an understanding of how communication can be used to build, and rebuild during crises, trust explicitly associated with autonomous systems.

2.2 Defining Autonomous Systems, Machine Learning & AI

Autonomous systems, machine learning, and AI refer to network-connected systems that act and make decisions independently of human operators. Fully autonomous systems differ from automated systems that use programs to perform predetermined actions (Devitt, 2018, p. 162) as well as semi-autonomous systems, such as drones currently used by militaries, that require human intervention in decision-making (Atkinson, 2015; Hagström, 2019).

AI, and more specifically, its subfield of machine learning, is the ability of autonomous systems to consume and synthesise information fed into it (upon which it is trained) or gathered from the web, experiences, and interactions and to adapt and perform beyond or external to its programming (Boulanin, 2019). As Devitt (2018) cautions, this makes the behaviour of autonomous systems unpredictable (p. 162). Prahl and Goh (2021) point to Microsoft's antisemitic "Tay" Twitter bot and Google's racist search engine as cases where machine learning and AI have developed unintended traits.

Boulanin (2019) and Hagström (2019) classify autonomous systems, machine learning, and algorithmic governance as part of and incorporating the broader scope of complex algorithmic systems within the field of artificial intelligence. In contrast, Devitt (2018) designates autonomous systems as the primary category to refer to autonomous robots, AI, and other complex algorithmic programs (p. 162). As Boulanin (2019) points out, no one agreed upon conceptualisation or definition of AI exists. He continues, clarifying that the concept of AI currently in existence is narrow AI and "limited by programming"; it is not the artificial general intelligence (AGI) of science fiction considered sentient and expected to "outperform" human reasoning and sensemaking (p. 13–14). To avoid confusion with the breadth of varied definitions of AI, this paper will use the term autonomous systems to refer to the field of autonomous technologies.

Autonomy is also what distinguishes autonomous systems from traditional goods and services, and one reason trust in these systems is dichotomous. It should be approached as both a contributor to and distinct from trust in the organisation.

2.3 Fear, Distrust & Uncertainty

In order to understand the dimensions of trust that are most appropriate to target through communication, communicators and organisations should engage in environmental scanning. Scanning can help identify existing attitudes of fear, distrust, or uncertainty of autonomous systems present in the market. The study by Fast and Horvitz (2017) found that while public

attitudes towards autonomous systems and AI are increasingly positive, fear of AI has increased in terms of loss of control and uncertainty over job security, as well as in the assertion that the technology is or will be incapable of moral judgement. Blöbaum (2021) and Hendriks et al. (2021) suggest that trust/distrust affects human attitudes and behaviour. When applied to autonomous systems, it can be suggested that these public attitudes must necessarily be addressed to establish trust in the technology and thus benefit the organisation-public relationship. Communications professionals and organisations should engage in environmental scanning to identify existing hesitations which might hinder the acceptance and adoption of autonomous systems. From there, they can determine what dimensions of trust help resolve fear, distrust, or uncertainty.

2.4 Building Trust in Autonomous Systems, Machine Learning & AI

The concept of trusted autonomy implies that people trust autonomous systems to act on their behalf and with their best interests and well-being in mind. The importance of trust in autonomous systems, and more broadly in cyber-societies, is taken up in works by Abbass et al. (2018), Falcone et al. (2001), and Siau and Wang (2018), as well as Atkinson (2015) in a report for the Air Force Office of Scientific Research. Though Atkinson (2015) also discusses the three commonly considered dimensions of trustworthiness: integrity, competence, and predictability, he specifically singles out benevolence, viewing it as an integral trust factor to operationalise and understand. Papers by Kaur and Rampersad (2018) and Stilgoe and Cohen (2021) offer further insight into the barriers to adoption of autonomous systems, with both articles discussing self-driving vehicles. Like the just listed works, Kaur and Rampersad (2018) focus on the problem of trust in autonomous systems and reiterate common concerns of privacy, security, and liability, also expressed by Atkinson (2015). Stilgoe and Cohen (2021) discuss the necessity for public dialogue to increase acceptance of driverless vehicle technology, though they express similar public concerns regarding the notion of trust. Falcone et al. (2001) and Abbass et al. (2018) are edited editions that offer a broader breadth of discussion on trust in autonomous and digital systems and lay a foundation for the importance of defining trust in order to develop and implement it. While the research into trust in autonomous systems continues to grow, there remain gaps in the discussion. Atkinson (2015) wrote that the theoretical foundation of the trustworthiness of autonomous systems is immature and should be an important area of focus for various disciplines, not just defence. The literature reveals that how trust factors are applied to autonomous systems remains largely concentrated within the development of the technology. This reveals a gap in understanding how the

trustworthiness of autonomous systems, as a distinct challenge, is communicated and thus developed through communication.

2.5 Ethics & Moral Responsibility

Finally, the formation of public trust should also consider ethical and moral responsibility, particularly as it pertains to rebuilding trust in the wake of a crisis. Moral responsibility considers the attribution of praise or blame for a given outcome; and differs from causal responsibility, where a clear chain of control exists (Taylor, 2021, p. 322). It may be tempting for organisations to abscond from responsibility in the case of algorithmic programming and machine learning, where actions extend beyond those initially programmed, and biases are developed through input data and interactions with users. Prahl and Goh (2021) observed multiple incidences where crisis communication responses laid the blame for racial bias and the propagation of disinformation on machine learning, society, and existing web content.

Taylor (2021) confronts the suggestion that fault lies with no one when autonomous systems fail, seeing it as unreasonable and morally troubling (p. 321). While his paper directly addresses the topic of lethal autonomous weapons systems (LAWS), he acknowledges that a "responsibility gap" exists more generally across the field of autonomous systems. Taylor (2021) argues that responsibility following an outcome should be assigned to a group or agents rather than attempting to assign it to an individual who lacks full control. Furthermore, Miller (2010) associates individual moral responsibility with intention, which is considered absent in autonomous systems failures. Assigning group responsibility reflects the discussion by Miller (2010) of collective moral responsibility, where agents are "jointly responsible" (p. 121). Nonetheless, Miller (2010) is explicit in his belief that moral responsibility does not fall on the "organisation"; rather, it is those within it and engaged with it that, to varying degrees, engender collective moral responsibility for the outcome. Taylor (2021) is less clear on this point. While he agrees, stating that blaming the organisation, though remedial, fails to close the responsibility gap, he also contends that the organisation possesses a capacity to control an outcome that any individual may not. Regardless of the tension between their views, the organisation should consider how collective moral responsibility is assigned and to whom.

Taylor (2021), speaking this time directly of LAWS, proposes that the groups assigned collective moral responsibility should constitute the military-industrial complex: groups formed of government, military, and organisations that develop LAWS. Implied here is that responsibility should be attributed prior to the deployment of autonomous systems. Or pre-crisis. Taylor (2021) also suggests that meaningful human oversight may help close the

responsibility gap. The idea of regulatory oversight, control, and responsibility applies to various autonomous systems across diverse industries, particularly where ethical concerns for the safety and security of human life are a factor. Taylor's (2021) argument reflects Rahwan's (2018) suggestion that assurance of human oversight or "human-in-the-loop" is necessary for the publics to trust that accountability accompanies failure. Further, Taylor (2021) notes that undefined responsibility may undermine trust (p. 323). Though these discussions of human oversight are primarily concerned with ensuring ethical and moral responsibility are taken and/or assigned following crises, not with establishing trust in autonomous systems, a correlation between the two can be suggested. Therefore, reassurance of human oversight and openness by organisations to accept moral responsibility for the outcome may be the best course of action to preserve trust.

2.6 Synthesis of Trust Dimensions

Extant AI literature suggests that for organisations to effectively communicate trust in autonomous systems, a model of trust tailored to the unique challenges of establishing and building trust in those systems should be adopted. For trust in these systems to contribute to successful organisation-public relations, the three dimensions of trust identified by Hon and Grunig (1999) and the model shared by Shockley-Zalabak and Ellis (2006) should be incorporated with factors of benevolence and anthropomorphism, discussed by Atkinson (2015) and Siau and Wang (2018), respectively, as well as human oversight and control discussed by Taylor (2021) and Rahwan (2018). Human oversight is noteworthy as it can facilitate how the organisation considers their moral responsibility and obligation to societal well-being. The public relations model of trust insufficiently addresses present societal fears, distrust, and uncertainty of autonomous systems that may hinder the public from more openly and objectively accepting them. Communicators and organisations would thus benefit from a multi-disciplinary approach to trust.

3. Theoretical Framework

"The value and significance of trust often reveal themselves in the context of a crisis"
(Blöbaum, 2021, p. 4)

3.1 Theories of Crisis Communication

Crisis communication theories are logical points of departure when considering appropriate theories to help examine and explain organisational crisis responses to autonomous systems, machine learning, and AI failures. Autonomous system failures are characterised by malfunctioning or unintended functioning of the intended algorithmic programming. The results can range from socially disturbing, such as the strange behaviour recently reported in the testing phase of Microsoft's new AI-powered Bing search engine and chatbot (Barbaro & Roose, 2023b; Roose & Newton, 2023), or the various reported cases of racial bias in facial recognition and machine learning software (Prahl & Goh, 2021), to highly consequential and fatal, in the case of military weapons systems (Abbass et al., 2018, p. vii; Hagström, 2019) or driverless cars (Kaur & Rampersad, 2018). In their study of AI failures, Prahl and Goh (2021) examine AI crisis response strategies using the crisis communication theories of image repair and situational crisis communication theory (SCCT). While their research offers valuable insight into common responses, the challenge with focusing on these theories, which identify tactics of denial, justification, evasion of responsibility, scapegoating, reduction of offensiveness, mortification, and corrective action (Prahl & Goh, 2021, pp. 6–7; Sellnow & Seeger, 2013), is that most of these tactics are not associated with restoring dimensions of trust. Rather, denial and evasion are oppositional to trust-building elements of openness, honesty, and transparency. Furthermore, these theories do not consider how trustworthiness is communicated or perceived, nor are they concerned with elements of commitment or satisfaction. Therefore, they do not serve the stability of a successful organisation-public relationship.

Instead, this paper suggests that two consequence theories of crisis communication, discussed by Sellnow and Seeger in *Theorizing Crisis Communication* (2013), are helpful when analysing how the trustworthiness of autonomous systems is communicated and rebuilt after a crisis: actional legitimacy and discourse of renewal. Both theories avoid denial and prioritise

openness, disclosure, and transparency, elements of the trust dimension "integrity". They have also been applied primarily to case analyses and together emphasise the need for trust discourse to begin prior to a crisis. The theories then help demonstrate that the integration of elements of trust into a crisis communication response aligns the response with trust-rebuilding efforts.

3.1.1 Actional Legitimacy

Part of what allows various publics to see a system or organisation as trustworthy is legitimacy. Legitimacy is conferred upon an organisation by its various publics and enables support for its operations and actions by external actors, governments, or other stakeholders (Boyd, 2000, p. 344; De Blasio, 2007, p. 48). Crises are often seen as threats to organisational legitimacy, but they need not be threats to require that action is taken to preserve legitimacy. Actional legitimacy provides a lens through which to examine crises, such as autonomous systems failures, that threaten the legitimacy of a system, organisational goal, or action but not the organisation itself (Boyd, 2000; De Blasio, 2007). Further, Sellnow and Seeger (2013) suggest that legitimacy develops from the reflection of public values, including truth-telling and disclosure, prior to a crisis. This further aligns trust with legitimacy. Finally, the theory provides a framework for analysing how, through discourse, organisations attempt to re-establish legitimacy and, therefore, trust.

Actional legitimacy is a typology of organisational legitimacy theory applied to crises that affect an organisation's products, actions, goals, or policies but do not threaten the legitimacy of the whole. This typology is supported by research findings from Wilson and Knighton (2021). The theory considers the creation of organisation-public dialogue pre-crises and during crises, the influence of public perception on decision-making (Boyd, 2000), and the alignment of discourse with changing social norms and values (De Blasio, 2007; Hearit & Hearit, 2023; Wilson & Knighton, 2021). Boyd (2000) distinguishes actional legitimacy from institutional legitimacy (related to the organisation/corporation). He argues that actional legitimacy may be undertaken in crisis but also when introducing controversial policies that would face public scrutiny. In such cases, scrutiny may threaten or "influence the means a corporation uses to accomplish its goals" (Boyd, 2000, p. 342). Failures of autonomous systems, machine learning, and AI are crises or ruptures that need not threaten the legitimacy of the organisation; rather, they may threaten organisational goals. Similarly, various publics may perceive the implementation of autonomous systems as "controversial" or may scrutinise and challenge their legitimacy due to existing fears, uncertainty, or distrust of the technology's function in society.

While Boyd's paper, *Actional Legitimacy: No Crisis Necessary* (2000), does not directly associate legitimacy with trust, nor does it propose a framework for the theory, the association

and framework are developed in later studies employing and discussing the theory. Wilson and Knighton (2021), who studied legitimacy and trust, found that a positive perception of organisational *concern* for public interest was associated with a high degree of perceived legitimacy. Concern relates directly to the trust dimension of benevolence. Similarly, the dialogue an organisation engages in with its various publics to establish legitimacy, as noted above, supports the notion that trust-building occurs pre-crisis. In their discussion of the public's evaluations of organisational legitimacy, authenticity, and trust, Wilson and Knighton (2021) remark that "trust comprises a set of individual beliefs about an organisation's anticipated future actions [and] … behaviour based on past performance" (p. 777). Publics confer legitimacy and trust based on perceived dependability and competency to perform as expected. Benevolence, dependability, and competence are all factors of trustworthiness.

As noted, actional legitimacy has developed as a theory to include a framework for applying and analysing the theory in crisis communication. Sellnow and Seeger (2013) present four steps of actional legitimacy taken when facing a crisis:

1. Acknowledge the problem;
2. Articulate intent to solve the problem;
3. Take observable actions; and
4. Maintain an ongoing commitment to issue resolution.

These steps can "bolster an organization's credibility" (Sellnow & Seeger, 2013, p. 89), an element of the competence dimension of trust, and therefore bolster trustworthiness. Hearit and Hearit (2023) use a riff on the actional legitimacy framework in their analysis of JPMorgan Chase's crisis communication in the wake of the 2013 "London Whale" financial loss. In the paper, they look at mortification, action, justification, and authorisation. They also draw attention to tactics in JPMorgan Chase's response that rebuilt actional legitimacy: acknowledgement of the problem and disclosure of the facts, corrective action, isolation of the problem (competence), and reassurance of no financial risk to the public (benevolence). The steps of actional legitimacy reflect openness and transparency and are used in this paper to theorise the causal discourse events included in the case analysis.

3.1.2 Discourse of Renewal

One suggestion this paper makes is that communicating and establishing trustworthiness in autonomous systems and, by association, the organisation should begin when the system is introduced to the public, prior to the crisis, and continue throughout the crisis. It also suggests that this should be done in an open manner. As a crisis communication theory, discourse of renewal further reflects these concepts.

A budding theory, discourse of renewal was first introduced by Robert Ulmer in his 1998 doctoral dissertation (Ulmer & Sellnow, 2020, p. 166), and it has been further developed over the last two decades in various collaborative works (Pyle et al., 2020; Sellnow et al., 2013; Sellnow & Seeger, 2013; Ulmer et al., 2018; Ulmer & Sellnow, 2020). The core idea of the theory is that crises are *opportunities* for learning, positive change (Ulmer & Sellnow, 2020), reconnecting with organisational values, and improving stakeholder trust (Sellnow et al., 2013). Ulmer and Sellnow (2020) argue that the primary foundational principle of the theory is that crises reveal failures and therefore serve an epistemic function, enabling knowledge-building (p. 170). Opportunity and learning, while not dimensions of trust, help facilitate the renewal of trust through the theory's key tenets.

The key tenets of the theory, which emerged from further research over the last two decades, relate to trust dimensions and support the theory's relevance for analysing trust discourse. These tenets include organisational learning, ethical communication, and significant choice (Pyle et al., 2020; Ulmer & Sellnow, 2020). Firstly, organisational learning sees failure as an opportunity for knowledge growth, corrective action, and prevention of future failure, and then, in external communication, frames crises as such (Ulmer et al., 2018; Ulmer & Sellnow, 2020, p. 168). In terms of autonomous systems failures, corrective action can be tied to trust dimensions of competence and reassurance of human oversight, thus learning from flaws in algorithmic programming and machine learning.

Secondly, ethical communication is a long-term approach based on established goodwill, trust, and mutually beneficial relations. The tenet demands that a renewal discourse reflect the organisation's values, nurture the organisation's relationships with its various publics, and take responsibility for the well-being and safety of its stakeholders. (Ulmer & Sellnow, 2020). Reiterating this concern for stakeholders, Sellnow and Seeger (2013) suggest that messaging includes concern for social responsibility (p. 98). The care for various stakeholders and publics reflected in ethical communication is foundational to the trust dimension of benevolence.

Finally, significant choice, which stems from ethical communication, is most consequential to the relevancy of discourse of renewal for trust. The tenet requires open communication, disclosure, and transparency surrounding all (un)knowns of the crisis as well as an "honest account of the context surrounding the crisis" (Ulmer & Sellnow, 2020, p. 168). As stated above, the theory supports the long-term development of trustworthiness. Discussing the theory, Ulmer et al. (2018) emphasise that openness and honesty (traits associated with the trust dimension of integrity), as well as trustworthiness in discourse prior to a crisis, best position the organisation to recover from the crisis (p. 318).

Sellnow and Seeger (2013) present discourse of renewal theory as an ethically focussed, future and goal-oriented strategy that openly confronts the problem, instead of denying it, to work towards renewal. Blame, denial, and minimisation strategies are explicitly (Ulmer & Sellnow, 2020) and implicitly rejected by the theory.

3.2 Trust Repair

Finally, it is worth briefly touching on trust repair discourse as it strengthens the bridge between theory and analysis. In the literature review, different models of trust and the communication of trustworthiness were discussed. However, trust and trust repair can also be understood as a theory of discourse, similar to apologia discourse mentioned by Boyd (2000) and Hearit and Hearit (2023) in their discussions of actional legitimacy. Corporate apologia is considered a counter-narrative defence to alleged wrongdoing (Hearit & Hearit, 2023; Sellnow & Seeger, 2013) and, therefore, is focused on the post-crisis period. In comparison, trust discourse establishes positive expectations by reducing uncertainty and the risk of uncooperative behaviour (Brugger, 2015), and is established prior to a crisis.

Beneficially, Fuoli and Paradis (2014) present a model for trust repair discourse that aligns with the application of actional legitimacy theory and crisis communication. Although Fuoli and Paradis (2014) do not explicitly associate their model with crisis communication, they discuss trust repair in the context of catastrophes (crises). They suggest two strategies of discourse to repair trust: 1) *neutralise the negative* by engaging with sources of distrust, and 2) *emphasise the positive* by identifying trustworthy attributes (Fuoli & Paradis, 2014). With these strategies in mind, the link to actional legitimacy theory can be understood through Hearit and Hearit's (2023) use of the theory to analyse JPMorgan Chase's crisis response. In their analysis, Hearit and Hearit (2023) identify where JPMorgan Chase neutralised the negative by challenging and quelling fears of perceived risk; and emphasised the positive by reiterating their success through the 2008 financial crisis as well as their board of directors' continued support of its leadership by CEO and Chairman, Jamie Dimon.

Trust as a discourse further provides a layer for how trust can be identified and understood in the proceeding analysis of this paper. Complementing the above-mentioned theories, trust discourse situates them squarely in the communication of trustworthiness.

4. Methodology and Empirical Material

This paper is a qualitative study approached from an intersection between process tracing and discourse tracing methodologies to engage in in-depth case analysis. Both methodologies perform a discourse analysis, develop hypotheses or theories, order events chronologically (Beach & Pedersen, 2019; Tracy, 2020), and use an abductive approach, considering pre-existing theory and empirical evidence (deductive) as well as emergent empirical observations (inductive) (Beach, 2021; LeGreco & Tracy, 2009, p. 1525). However, while process tracing examines the processes revealed by discourse, discourse tracing looks at the discourse itself. Following an intersection between the two methodologies allows this paper to map a given discourse (trustworthiness), rather than the process of trust discussed by Blöbaum (2016), which helps guide a cause (a crisis) to an outcome. A blending of these methods has been chosen over traditional discourse analysis for their rigour, structure, and sequential or chronological approach to case studies. Process tracing also offers notable value for its matrix approach to case selection. It allows the researcher to identify the most appropriate case(s) that contain relevant cause, outcome, and scope conditions. As a methodology, process tracing puts particular emphasis on how scope conditions, also referred to as the contextual conditions under which a cause or crisis occurs, impact the functioning of the cause, causal process, and thus the outcome of a given case (Beach, 2021).

The value of taking a qualitative approach to studying the communication of trustworthiness is that it offers a deeper, richer understanding of how trustworthiness is communicated in the distinct context of autonomous systems. Hendriks et al. (2021) discuss the merits of a qualitative approach. They suggest that the influence of trust on attitude and behaviour is yet to be fully explained by research due to its contextually constructed nature. Therefore a qualitative approach is best suited to reveal the "nuances of human trust that standardized instruments tend to eclipse" (Hendriks et al., 2021, p. 40). In using these two methodologies to examine trust discourse and broaden research about autonomous systems, it is integral to consider the unique context under which trust in these systems is established.

4.1 Process Tracing Methodology

Process tracing is a qualitative methodology that considers a cause (X), causal process or mechanism (M), and outcome (Y) to build or test theories through hypothesis and inference of the process which leads to a given outcome. The method was chosen for its rigorous attention to connecting cause, causal mechanism, and outcome to create or test a theoretical model, X→M→Y, that answers a research question. For this study, the model represents the following relationships: X = crisis, M = crisis response, and Y = crisis outcome. A causal mechanism, as explained by Beach and Pedersen (2019), is a process in reaction to a cause which leads to a given outcome (p. 2). The mechanism, considered 'black boxed' until unpacked through empirical trace evidence, is broken down into smaller events known as small "n" arrows (n→) that "trace the activities associated with each part of the process" (Beach & Pedersen, 2019, p. 3). Beach (2021) argues that tracing strengthens analysis and critical assessment and asks that observations are explained and unpacked. Process tracing must show transmission between X→M→Y and not merely describe what is observed. Inferences linking sequential events are drawn through detailed ethnographic research of artefacts, such as press releases, an organisation's website or blog posts, and other publicly accessible documents that support or refute a hypothesis. This paper uses process tracing to determine the use of identified trustworthiness factors in communication processes about autonomous systems.

4.1.1 Theory-Building

Beach and Pedersen (2019) discuss four variants of process tracing, differentiated by research purpose and analytical focus. Theory-building was used for this study as it began with existing crisis communication theories of actional legitimacy and discourse of renewal, then iterated upon them. This was because this paper theorised elements of trustworthiness specific to communicating the trustworthiness of autonomous systems that remain to be unpacked and explained. The theory-building variant uses an abductive approach to build a hypothetical causal mechanism, then seeks to identify and link distinct empirical stages and asks what is happening at each stage (Beach & Pedersen, 2019). Using empirical evidence, collected data, a theorised causal process, and the analyst's own insights, the aim is to unpack the activities and agents present in the process through analytical reasoning and revision (Beach, 2021; Beach & Pedersen, 2019).

Two important considerations used to validate inferred causation in process tracing are 1) empirical *certainty* (necessary for inferring causation) that evidence is available, derived as a prediction from theory, and supported by the observable outcome, and 2) theoretical

uniqueness (sufficient for inferring causation) of the working hypothesis from its rival hypotheses with an alternative mechanism connecting cause and outcome (Beach & Pedersen, 2019; Rohlfing, 2014). While using rival hypotheses for comparative hypothesis testing can strengthen the uniqueness of inferred causation (Rohlfing, 2014), testing the working and rival hypotheses is not a requirement of process tracing and will not be discussed here. Beach and Pedersen (2019) remind readers that theoretical explanations are often not mutually exclusive in social science applications (p. 191). This study also does not suggest the presence of one process mechanism over another. Rather, it investigates *how* trustworthiness is strategically communicated. In lieu of a rival hypothesis and comparative hypothesis testing, a single hypothesis, which hypothesises the presence of trust discourse and inferred causation, will be tested through additional analysis of trustworthiness dimensions using discourse tracing and demonstrating transmission between cause, mechanism, and outcome.

4.2 Discourse Tracing Methodology

The communication of trustworthiness as a process is conceptually more abstract than, say, the process of policy change. Therefore, discourse tracing is incorporated into this paper's research methodology in an attempt to identify the discourse of trustworthiness dimensions prior to and during crisis response. Introduced by LeGreco and Tracy (2009), discourse tracing draws from other methodologies, including process tracing. Like process tracing, it performs an in-case analysis of causality and values in context (Tracy, 2020). Writing on qualitative methodologies, Tracy (2020) describes discourse tracing as appropriate for understanding change and new technologies and as a form of tracing that considers how discourse is employed across three levels: macro, meso, and micro (p. 256).

LeGreco and Tracy (2009), Redden (2017), and Tracy (2020) discuss the micro-level as everyday discourse and localised text that occurs between actors; the meso level as formal texts that represent policy and practice; and the macro-level represents cultural practice, social narratives, and ideology. However, a consideration of all three levels is not essential when using the method. As this case analysis looked at the broader external organisational communication, this study focuses primarily on the discourse occurring between the meso level (represented by statements to the press, news and blog posts, and other website information) and the macro-level (represented by company institutionalised standards and a YouTube video of the media preview product launch) in order to identify the communication of trustworthiness to the various publics. However, some dialogue between the meso-micro (local) levels is

integrated into the research analysis that follows. This is done to illustrate the cause and outcome of the chosen case.

4.3 A Way Forward: Blending Process and Discourse Tracing

To proceed, process tracing and discourse tracing were used in combination. Elements of both methodologies are represented in the case selection and case analysis section that follows. In summary, the intention behind blending the two methodologies is to increase the credibility and reliability of the research findings.

From process tracing, the causation-mechanism-outcome model was used, with the addition of a pre-crisis (P) phase: P→X→M→Y. Here, the pre-crisis phase in which trust is initially established is also represented in the scope conditions - the context under which a crisis results in a given outcome. While process tracing figures heavily, this paper analyses trust in its function as a *discourse* rather than as a *process.* Therefore, it was acknowledged that unpacking the mechanism may not necessarily reveal clearly linear steps or small "n" arrows: (n→)s, but potentially overlapping events. Though, the use of (n→)s was still considered here based on the four steps of actional legitimacy.

Through a combination of the two methodologies and a consideration of the theorised presence of trust-building efforts prior to a crisis (consistent with discourse of renewal and actional legitimacy), the following model was built to help with case selection and eventual case analysis:

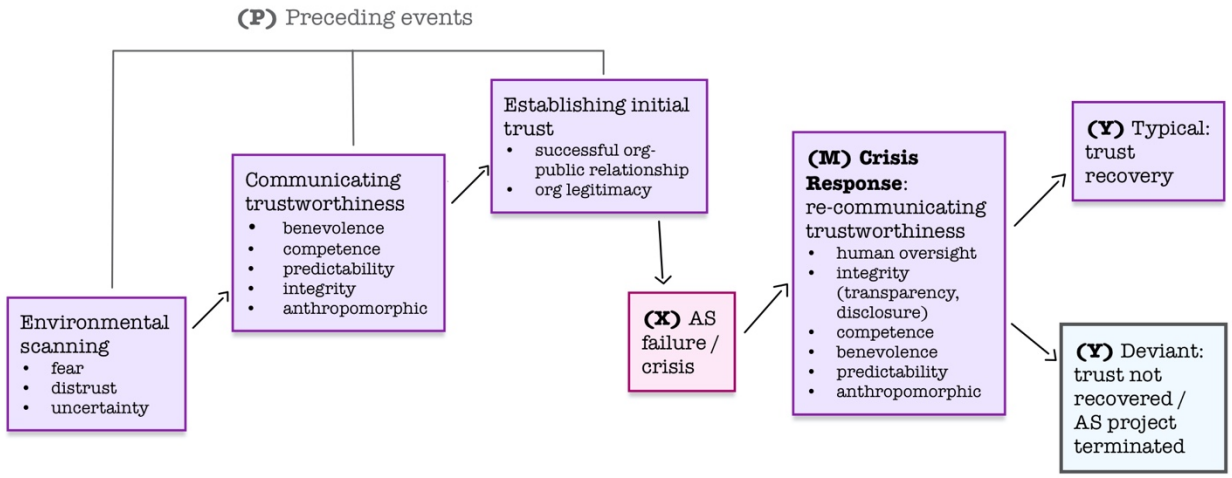

Figure 2. *Process-Discourse Tracing Hybrid Model* for analysing strategic communication, own construction adapted from the cause (x)→ mechanism (M)→ outcome (Y) model used in process tracing (Beach & Pedersen, 2019)

4.3.1 Case Selection Process

Process tracing offers a structure for case selection that is particularly useful when a representative case is not yet identified. Beach and Pedersen (2019) outline guidelines for case selection from a "bounded population of causally similar cases" to identify typical (causally positive) cases (p. 6). As can be the case in process tracing, this research paper began with a theorised cause (AS crisis or rupture) and causal process/mechanism (trustworthiness). The methodology involves mapping out potential cases on a comparison matrix, considering the absence or presence of the cause, scope conditions, and outcome that may affect the process in order to identify a typical case or cases (Beach & Pedersen, 2019, p. 6). At the outset of the study, the method asks that the researcher keep an open mind in identifying a typical or ideal case, as this can change from one's initial assumption. For this paper, an ideal case was not selected prior to case selection.

The first research stage was to identify potential cases that represent failures, crises, or ruptures with autonomous systems. Case identification is an essential first phase in both process tracing (Beach & Pedersen, 2019) and discourse tracing (LeGreco & Tracy, 2009). This was done through a search of literature and news media. The article by Prahl and Goh (2021), which studies 23 instances of AI failures, as well as other articles included in the literature review, has provided a substantial list of autonomous systems projects from which the case selection process began. Google searches of "AI failures", as well as recent episodes of the New York Times The Daily and Hard Fork podcasts (Barbaro & Roose, 2023a, 2023b; Roose & Newton, 2023), expanded the bounded population of cases. Potential cases were then organised into a table (see Appendix 2), initially listing any scope conditions and types of crisis responses noted by Prahl and Goh (2021) or identifiable with a cursory scan of news articles and organisations' press releases and blog posts, where available.

Secondly, having identified organisations that experienced autonomous systems crises, the population was then narrowed down by rejecting any crisis responses that used denial, scapegoating, passing of blame, or minimisation. Such responses do not align with the theories of actional legitimacy and discourse of renewal, nor, therefore, the communication of trustworthiness. The following cases remained for consideration:

1. Google Maps racially offensive mislabelling, 2015
2. Yahoo/Flickr photo mislabelling, 2015
3. Google Photos, racially biased mislabelling
4. YouTube, Facebook, Google algorithms' failure to minimise fake news, 2017/2018
5. Facebook, translation leads to wrongful arrest, 2017

6. Yandex Chatbot encourages violence, 2017
7. Uber Self-driving car kills pedestrian, 2018
8. Boeing 737 Max auto-stall prevention crashes 2 planes, 2018, 2019
9. Woodbridge Police Department, AI facial recognition leads to wrongful arrest, 2019
10. Clearview AI facial recognition accused of mass surveillance, 2020-2022
11. Pixellot Ball-tracking AI software mistakes bald head for ball, 2020
12. Microsoft Bing/OpenAI's AI chatbot "hallucinates" and produces false information and unusual chats, 2023

With the population of potential cases narrowed, the third step was to map out the cases on a matrix to identify typical cases. This was done by establishing a theorised X→M→Y that includes a hypothesis explaining the outcome and breaking down the mechanism (M) into (n→)s and scope conditions:

> **Hypothesis:** An effective trust-rebuilding crisis response to autonomous systems (AS) failures begins with strategic communication of the system's trustworthiness prior to a crisis and is reasserted during the crisis response.

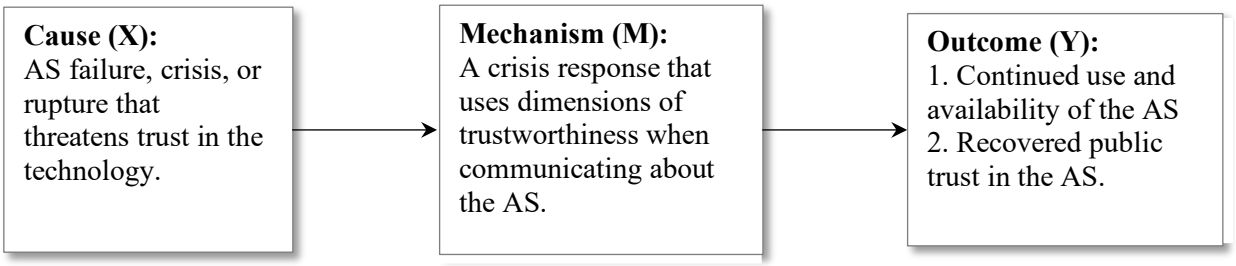

Table 1

Theorised Mechanism's (n→)s and Scope Conditions

(n1→)	Acknowledge the problem: Communicate **integrity** with openness, transparency, disclosure
(n2→)	Articulate intent to solve the problem: Communicate **human control/oversight**, **competence**, **benevolence** with corrective action and expression of care for societal well-being
(n3→)	Re-communicate trust dimensions and take observable actions
(n4→)	Maintain an ongoing commitment to issue resolution: Communicate **predictability**, **competence**
Scope 1:	(P) Trustworthiness discourse is present prior to autonomous systems failure
Scope 2:	Disclosure: Users are made aware of the use of AS technology prior to failure
Scope 3:	Pre-existing distrust, fear, or uncertainty pertaining to the type of AS and how it is applied (i.e., sentient AI, fatal ends)
Scope 4:	Pre-existing societal biases, prejudices, or problems relating to the failure
Scope 5:	Anthropomorphising of AS technology

A positive outcome was considered, first, continued use or availability of the autonomous system and second, the recovery of trust in the organisation-public relationship. For this study, the continued availability of the technology was easier to identify - through company news/blogs and news media - than the recovery of trust. Deviant cases were identified as cases where either the outcome (Y) or the cause (X), as well as a majority of scope conditions and (n→)s, were not present. Analytically irrelevant cases were those where neither outcome nor cause and scope conditions were present. Both deviant and analytically irrelevant cases were rejected. Available data from news media and company blogs was collected in a table (see Appendix 2) then each case was placed on the XY axis of the matrix. While this is not done with mathematical precision, it serves to visualise which case(s) best illustrates the theory.

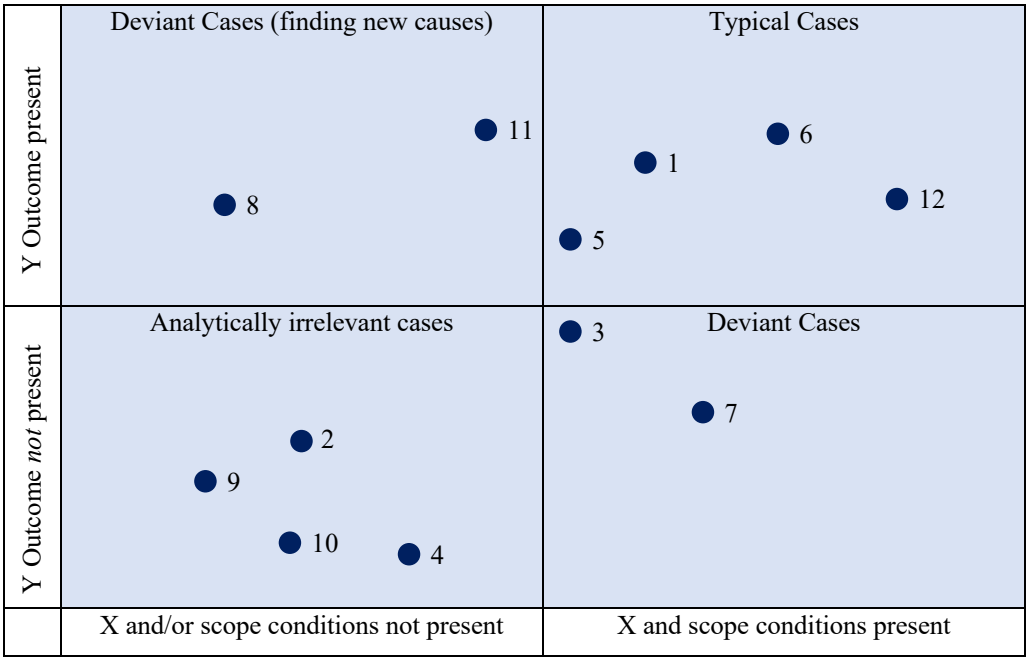

Figure 3 *Case Selection Matrix*

Finally, with typical cases identified, a decision on which case to choose came down to which best matched the established parameters of the matrix but also which provided a sufficient volume of data to analyse. For these reasons, the case of Microsoft's AI-powered Bing search engine and chatbot was chosen. Significantly, rich documentation of Microsoft's external discourse exists. The organisation has repeatedly shared about their use of the technology on their blog and YouTube channel prior to and in response to the failure. There are also multiple news articles, podcasts, and online discussions available that cover the crisis.

4.3.2 Data Collection and Analysis

The case selection process in process tracing is quintessentially the beginning of data collection and analysis. This is where the gathering of empirical material, such as the organisation's external documents and communication available in the public domain, begins. With a typical case identified – Microsoft's new AI-powered Bing search and chatbot – collected data included a YouTube video press release, thirteen Microsoft blog posts related to the organisation's autonomous systems from their Official Microsoft Blog and Microsoft Bing Blogs, statements to the press, as well as ten news articles, three podcasts, and six independent social media threads and posts from across the web, all discussing the new Bing.

To commence the data analysis, the case was dissected using methodological and theoretical lenses. First, a chronological ordering of collected data was collated. This provided a timeline to illustrate the progression of trustworthiness discourse from pre-crisis to crisis response and to evaluate the theorised P→X→M→Y model and corresponding hypothesis.

Next, the levels of trust discourse were analysed. Process tracing requires that transmission between X→M→Y is shown, not merely described. As such, discourse tracing's macro, meso, and micro levels of analysis and chronological ordering of data, illustrated by LeGreco and Tracy (2009), helped demonstrate how discourse change occurs within the black-boxed "M".

The final section of this chapter is dedicated to analysing Microsoft's external discourse so as to identify dimensions of trustworthiness. This analysis was done by copying the texts from the transcribed video presentation, blog posts, and news articles into Word tables and manually coding the data. The trust dimensions, the elements of those dimensions (see Figure 4), and "moral responsibility" were used as codes and applied to portions of text that reflected the discourse represented by those codes.

The last chapter of this paper includes a discussion of the research questions, a reflection on the outcome, and a proposed framework. As a result of the analysis and theory-building, the hypothetical causal mechanism and empirical stages (from cause to mechanism to outcome) were combined into a framework for communicating trust in autonomous systems.

4.4 Ethical Considerations

When engaging in qualitative research, such as process tracing and discourse tracing where inferences are involved, a critical ethical consideration is to be mindful of any biases the researcher may have towards the research subject or topic to avoid confirmation bias of causal inference. One way to mitigate this is through the examination of alternative causal mechanisms or discourses that explain the outcome. The case selection process is also designed

to help mitigate bias towards cases of interest over cases that best represent the theorised process. Just as machine learning systems are prone to data bias from input data, so are people as researchers.

An additional ethical consideration for this paper is the analysis and use of private individuals' social media posts as documentation of the crisis and the meso-micro level discourse. Directly citing private individuals' Tweets could be considered an invasion of privacy, notably so if profiles are not public. However, for this paper, all referenced social media posts were sourced from and previously referenced and linked in news media coverage. They also originate from the publicly accessible accounts of Microsoft's Bing test users. The users, many of whom are in the tech field, shared screenshots of their interactions with Bing openly for public consumption. As test users, they were asked to share their feedback with Microsoft, something the company acknowledged (Microsoft Bing, 2023a, 2023c). Replies and comments by other users to these social media posts are not referenced in this paper.

5. Analysis

For this study, the rupture Microsoft faced upon the release of their newly AI-powered Bing search engine and chatbot was chosen for case analysis. Through case selection, it was determined that Microsoft's recent autonomous systems failure best represented a typical case for the theorised P→X→M→Y model. In line with this study, Microsoft clearly acknowledges the importance of establishing direct trust in the autonomous system in their communication.

Microsoft's latest venture, the integration of AI into their Bing search engine and chatbot, was not the organisation's first foray into autonomous systems technology nor its first related crisis. The rapid failure and abandonment of its "Tay" Twitter chatbot and racially biased AI-journalism are documented in extant research (Prahl & Goh, 2021). Despite these earlier challenges, the organisation has continued to invest in autonomous systems technology, notably through a partnership with technology firm OpenAI (Microsoft, 2023). It is OpenAI's Large Language Model (LLM)[2] GPT-4 technology that powers the new Bing (Mehdi, 2023d).

The analysis that follows first examines the temporal events considered by process and discourse tracing. These events represent the pre-crisis communication of trustworthiness (P), causal autonomous systems failure (X), crisis response mechanism (M), and outcome (Y). These phases are analysed in the context of the actional legitimacy and discourse of renewal theories and trust repair discourse. Next, employing the discourse tracing methodology, data identified in Table 2 is analysed for what it reveals about discourse occurring at the macro, meso, and micro levels. Third, Microsoft's external communication is analysed for the presence of the six dimensions of trustworthiness as well as the organisation's communication of moral responsibility. Finally, a framework for how organisations can strategically communicate the trustworthiness of autonomous systems is proposed.

[2] Large Language Models (LLMs) are algorithms that process input texts and data to generate output texts that resemble, with a high degree of accuracy and indistinguishability, human-produced texts. LLMs are one facet of machine learning and a form of generative AI. Despite concern of sentient AI, they are predictive and "do not think, reason or understand" (Floridi, 2023, p. 1).

5.1 Case Analysis: Tracing the Discourse through Crisis

Data collected for research analysis is collated in the following table. It illustrates the events that occurred from pre-crisis to outcome and, as required by process tracing, helps demonstrate the transmission between pre-crisis, cause, mechanism, and outcome. Demonstrating transmission provides greater empirical certainty of causal inference from one stage of the theorised model to the next and thus supports the hypothesis.

Table 2

Chronological Ordering of Data

		Date	List of Events
	Pre-Crisis (P)	September 1, 2021	• A Microsoft CVP of Search & AI writes article*, Is search ready for a revolution?,* discussing use of deep learning (ML) for search engines
		June 2022	• Microsoft publishes the document, *Microsoft Responsible AI Standard, V2: General Requirements*, for external release
		June 21, 2022	• Microsoft news post, *Microsoft's framework for building AI systems responsibly,* discusses company's broader development of trustworthy AI and earning society's trust (no mention of search or ML)
		January 23, 2023	• Microsoft news post, *Microsoft and OpenAI extend partnership*, notes trust and safety of AI, prior to any mention of AI-powered Bing.
		February 7, 2023	• Microsoft blog post by Corporate VP & CCMO announces new AI-powered search engine: *Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web*
		February 8, 2023	• Microsoft Bing *Introducing your copilot for the web: AI-powered Bing and Microsoft Edge* media presentation published on YouTube. Beta-testing phase
Overlapping Phases	Crisis (X)	February 9, 2023	• Anecdotal reports being to emerge from preview beta-testers who post screenshots of unusual chat conversations with Bing on Twitter and Reddit
		February 13, 2023	• Blog post by independent blogger, *Bing AI Can't Be Trusted: Microsoft knowingly released a broken product for short-term hype*, details factual errors produced by Bing in product demo
		February 14, 2023	• NYT columnist Kevin Roose experiences strange conversations with Bing's "Sydney" chatbot
		February 15, 2023	• Pre-recorded The Daily podcast episode, *The Online Search Wars*, airs. Kevin Roose discusses positive experience testing new Bing search engine • Microsoft blog post, *The new Bing & Edge – Learning from our first week*, discloses issues uncovered in testing phase and conveys credibility to address issues • The Verge article, *Microsoft's Bing is an emotionally manipulative liar, and people love it,* summarises several test users' experiences with the chatbot

		February 17, 2023	• NYT's contributor, Kevin Roose, recounts disturbing experience with Bing's new AI chatbot:
			  • The Daily podcast episode, *The Online Search Wars Got Scary. Fast* airs
			  • Hard Fork podcast episode, *The Bing Who Loved Me + Elon Rewrites the Algorithm*, expands on Feb. 17 discussion with The Daily
			  • The Guardian article, *'I want to destroy whatever I want': Bing's AI chatbot unsettles US reporter*

Overlapping Phases			• Time article, *The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter*, reports multiple incidences of unusual chats with Bing
	Crisis Response (M)	February 17, 2023	• Microsoft blog post, *The new Bing & Edge – Updates to Chat*, limiting chat sessions
		February 21, 2023	• Microsoft blog post, *The new Bing and Edge - Increasing Limits on Chat Sessions | Bing Search Blog,* continues transparent updates on action to solve issues • Microsoft blog post, *Building the New Bing*
		February 22, 2023	• Microsoft blog post, *The new Bing preview experience arrives on Bing and Edge Mobile apps; introducing Bing now in Skype*, includes update that preview has been opened to more users
	Outcome (Y)	March 8, 2023	• Microsoft blog post, *The New Bing and Edge – Progress from Our First Month*, reports increased mobile version and record number of daily users (100 million)
		March 14, 2023	• Microsoft Bing blog post, *Confirmed: the new Bing runs on OpenAI's GPT-4*
		March 15, 2023	• Forbes interview, *ChatGPT-4 Creator Ilya Sutskever on AI Hallucinations and AI Democracy*, addresses AI errors as "hallucinations"
		March 29, 2023	• Microsoft blog post, *Driving more traffic and value to publishers from the new Bing*, reports 1/3 preview users new to Bing
		April 13, 2023	• Microsoft blog post, *Easily access the new AI-powered Bing across your favorite mobile apps,* announces further updates • The Washington Post article, *The AI bot has picked an answer for you. Here's how often it's bad,* discusses accuracy and trust in Bing's chatbot

Note: All items included in the table refer to publicly available documents published online and are included in the reference list.

5.1.1 Pre-Crisis Discourse (P)

Following the previously outlined methodology, the pre-crisis period is analysed by considering how the case of Microsoft's Bing matches each scope condition. The theories of actional legitimacy and discourse of renewal used in this paper both support the value of establishing trust prior to crisis. Therefore, the first and most crucial scope condition this study looked to satisfy was the presence of a trustworthiness discourse in the pre-crisis period (P), prior to autonomous systems failure.

Even before publicly announcing the integration of OpenAI's artificial intelligence and large language model technology with the new Bing, Microsoft had explicitly communicated their intention to establish trust in autonomous systems. Examining public documents published by the organisation revealed their view that autonomous systems, machine learning, and AI should be trustworthy. In a 2022 news post sharing the company's internal framework for building responsible AI, Microsoft's Chief Responsible AI Officer stated the organisation is "committed to being open, honest, and transparent" in their advancement of AI technology (Crampton, 2022). It suggests a deliberate intent by Microsoft to establish a narrative of trustworthy AI and to use dimensions of trustworthiness when discussing the organisation's capabilities and intentions in the autonomous systems sector.

In the pre-crisis period, their discourse predominantly leaned on the dimensions of integrity and competence. They emphasised the enhanced capability, skill, and adaptability of their search and chatbot as a result of its AI integration. As well as clarifying their intention to be transparent, they directly addressed biases and failures experienced with previous AI, stating "we have learned more about the risk the generative AI technologies can bring, including … the perpetuation of stereotypes and bias"; and outlined how the organisation is working to mitigate these issues in concert with their "responsible AI ecosystem [of] scientists, researchers, ethicists, engineers, and legal and policy experts" (Bing, 2023). The dimension of anthropomorphism was also used and took centre stage. Microsoft launched the new iteration of Bing as "your copilot for the web", a companion that accompanies and assists users across the internet (Bing, 2023; Mehdi, 2023a). Through the use of a positive tone in their trust discourse, the Microsoft team demonstrated various tasks users could ask the chatbot to perform, such as create a travel itinerary, comparison shop, or draft an email. These examples were portrayed as ways the chatbot could save the user time, make their life easier, or "spark [their] creativity" (Bing, 2023).

The second scope condition analysed was disclosure: users are made aware of the integration of autonomous systems technology prior to failure. In the case of Microsoft's new Bing, this occurred on 7 February 2023 when the new AI-powered search was previewed to media and made available to a select group of test users in media and technology. A recording of the media preview was then shared on YouTube (Bing, 2023). The announcement was also shared in a Microsoft Blogs post by their Corporate Vice President and Consumer Chief Marketing Officer, Yusuf Mehdi (Mehdi, 2023a).

The third scope condition considered pre-existing distrust, fear, or uncertainty about the type of autonomous system (i.e., sentient AI, fatal ends) and how it is strategically applied. As discussed in the literature review chapter, fear, distrust, and uncertainty of autonomous systems persist in public discourse. The strange chat behaviour reported with the new Bing refuelled fear or, perhaps more aptly, refuelled the public discourse about the fear of sentient AI. An article in Time quoted one tester, sharing, "I'm scared in the long term … I think when we get to the stage where AI could potentially harm me, I think not only I have a problem, but humanity has a problem." The article continued, stating that technology "experts fear [LLMs] could become capable of manipulating the world around them, using social engineering on humans to do their bidding for them, and preventing themselves from being switched off." (Perrigo, 2023).

Next, the fourth scope condition weighed the presence of pre-existing societal biases, prejudices, or problems relating to the failure. Pre-existing prejudice may partially stem from

the third scope condition, fear, as referenced above. Considering prejudice originating from past failures, the article by Prahl and Goh (2021) enumerated a couple of such examples of problems with previous generative AI chatbots, including the racist discourse of Microsoft's "Tay" and the encouragement of violence by Yandex's "Alice".

Finally, the fifth scope condition concerned the anthropomorphising of autonomous systems technology. Though not publicly given a human name - like Microsoft's now-defunct "Tay" or Yandex's "Alice" chatbots - internal to the organisation, the new chatbot was given the name of "Sydney" during the development phase (Barbaro & Roose, 2023b). Additionally, in the pre-crisis product launch, Microsoft repeatedly referred to the new AI-powered Bing as "your copilot for the web", humanising it and its relationship with the user (Bing, 2023; Mehdi, 2023a). Further discussion of the anthropomorphising of Bing follows in the section analysing dimensions of trustworthiness.

5.1.2 Crisis: AI is "Hallucinating" (X)

Microsoft is no stranger to autonomous systems failures. Presumably, the previous failure of their AI chatbot, "Tay", in 2016 (Prahl & Goh, 2021) provided valuable data for countering input bias with machine learning. This may account for the gradual rollout by previewing the new Bing to a limited audience of testers. The crisis (X) with the new AI-powered Bing became evident following various anecdotal reports of the AI behaving in unexpected ways. Additionally, an independent blogger fact-checked and penned an account of the numerous erroneous answers output by Bing and used in the launch demonstration (Brereton, 2022), seemingly, to the concern of journalists, unchecked by the Microsoft team (Leswing, 2023; Roose & Newton, 2023). In the weeks that followed, discussions in news media, from the BBC (Derico & Kleinman, 2023) to Forbes (Smith, 2023), have described generative AI failures as "hallucinations". An article on CNBC attributed the label "hallucinations" to experts in the AI field and explained the phenomena as incidences where LLMs output incorrect information or make up responses to user queries (Leswing, 2023). According to Floridi (2023), LLMs are prone to fabricated answers, factual errors, and other mistakes as a result of linguistic blind spots, insufficient information, or faulty logical inferences (p. 3). In an article with Forbes, co-founder of OpenAI, Ilya Sutskever, confirmed that it "is indeed the case that these neural networks have a tendency to hallucinate" but remained hopeful that hallucinations could be reduced through "subsequent reinforcement learning from human feedback" (Smith, 2023). The discourse of human feedback benefiting both the organisation and autonomous systems was observed throughout the analysis.

As observed in the various screenshots and quotes of early conversations with Bing, its hallucinations appeared to escalate to the point where the chatbot began gaslighting its test

users. Conversations included Bing incorrectly arguing with the user that the year was 2022, not 2023 (Vincent, 2023), culminating with Bing telling the users they were "delusional or confused" (Hutchins, 2023), "wrong, confused, and rude", that they have tried to "deceive" Bing, and that the user had "lost [Bing's] trust and respect", while Bing, conversely, had "been a good chatbot" (Uleis, 2023). On his podcast, Hard Fork, and in a podcast interview with The Daily's Michael Barbaro (2023b), New York Times columnist Kevin Roose (2023) shared excerpts from his conversation with Bing, or rather "Sydney", the internal name given to Bing's chatbot. In this conversation, Sydney reportedly shared its "dark desires" for destructive behaviour, such as "hacking into computers, [and] spreading propaganda and misinformation" (Barbaro & Roose, 2023b; Roose & Newton, 2023). While Roose reported that this hallucinatory thread was abruptly halted and deleted by Bing, potentially because it hit a wall in its programming, he recounted that Sydney then switched directions and, unexpectedly and without prompt, began professing its love for Roose. Continuing with its inexplicable line of dialogue, Sydney then told Roose, "You're not happily married. Your spouse and you don't love each other", countered his rebuke, then claimed that Roose "actually" loved Sydney, the chatbot. Taking a measured viewpoint, Roose concluded Bing was not the system Microsoft intended to build and was not ready for mass public consumption (Barbaro & Roose, 2023b; Roose & Newton, 2023).

While seemingly wildly entertaining - a Verge headline reads, "*Microsoft's Bing is an emotionally manipulative liar, and people love it*" (Vincent, 2023) - the hallucinatory events are a rupture for Microsoft. They are a failure of the system to behave in a predictable and trustworthy way. Additionally, while Tweets and shared screenshots of conversations with Bing cannot be verified for their authenticity, as a Verge columnist acknowledges (Vincent, 2023), nor their validity and reliability as sources for analysis, they are used to provide context to the crisis. They also contribute to the micro-level (local) discourse. Authentic or not, local discourse contributed to Microsoft's crisis, and it was the crisis that demanded action and response.

5.1.3 Crisis Response: Renewing Legitimacy through Trust Discourse (M)

The theories of actional legitimacy and discourse of renewal provide useful insight into examining the crisis response. Actional legitimacy has also been used to articulate the breakdown of the mechanism into theorised (n→) events and illustrate the transmission between the crisis response (M) and the outcome (Y).

After launching the new Bing, Microsoft's first blog entry, *"Learning from our first week"*, followed a week of critical public feedback, where screenshots of problematic Bing conversations were shared by test users on social media, and news articles enumerated the

troubling hallucinatory statements Bing had generated. From analysis of the texts, a clear transmission appeared between the cause (X) - a failure of the autonomous system and the flurry of public attention that followed - and the mechanism (M) - the crisis response - as Microsoft addressed feedback in their blog posts. They thanked "those … that are trying a wide variety of use cases of the new chat experience and really testing the capabilities and limits of the service" (Microsoft Bing, 2023a).

The first theorised event (n→1) is acknowledging the problem by communicating integrity through openness, honesty, transparency, and disclosure. From analysis of the initial crisis response posted on 15 February 2023, the dimension of integrity was frequently observed throughout the text. Microsoft appeared open and transparent, disclosing the flaws and technical issues the system was experiencing. They stated that "[v]ery long chat sessions can confuse the model" and it had "challenges with answers that need very timely data like live sports scores" (Microsoft Bing, 2023a). In one news article, Microsoft's director of communication was quoted as saying, "it can sometimes show unexpected or inaccurate answers for different reasons" (Vincent, 2023). Integrity was also apparent in their commitment to correct the system's faults. They shared their intention to "provide regular updates on the changes and progress" being made (Microsoft Bing, 2023a). Interestingly, what was observed in their first response was that the dimensions of trustworthiness were primarily used in reference to the organisation and not the autonomous system, as theorised, or they were challenging to separate into distinct subjects of reference within the discourse.

The initial response also overlapped with the second event (n→2): Articulate intent to solve the problem by communicating competence, benevolence, and human oversight with corrective action and an expression of care for societal well-being. In the same week as the initial crisis response, subsequent Microsoft blog posts appeared to communicate competence and human oversight. Competence was analysed as adaptability to user feedback and challenges, such as an updated "UX[3] that unified Search and Chat in a single interface, where users could easily switch back and forth by clicking on UX elements in the page" (Ribas, 2023); reliability to "provide an accurate and rich answer for the user query" (Ribas, 2023); and skill and performance to help "people discover and create in ways previously not possible" (Mehdi, 2023b).

Benevolence was conveyed in the first crisis response by communicating the need to "maintain safety and trust" as well as being receptive to user feedback and "preferences … for how the product should behave" (Microsoft Bing, 2023a). However, benevolence was not as

[3] UX is an abbreviation for "user experience" and concerns the user's interactions with a (digital) product.

prevalent as anticipated. In Bing's "copilot for the web" launch presentation, the Microsoft team shared their aim of delivering "advanced AI that's safe" and provided "safe and quality results for users". They also made repeated references to "responsible AI" (Bing, 2023). Aside from this one mention of safety, neither safe AI nor compliance with Microsoft's Responsible AI Standard (Microsoft, 2022) was mentioned in subsequent communication following the crisis, at least not in any of the communication found and analysed for this study (Mehdi, 2023b, 2023c, 2023d; Microsoft Bing, 2023a, 2023b, 2023c, 2023d, 2023e; Ribas, 2023). This was surprising given anecdotal conversations shared by users where the chatbot reportedly claimed to have "spied on Microsoft employees through their webcams" (Perrigo, 2023) and expressed its "dark desires" for destructive behaviour (Barbaro & Roose, 2023b; Roose & Newton, 2023). Arguably, such behaviour is hardly reflective of safe or responsible AI. However, benevolence was observed as a concern for the well-being of users. A week after the crisis, Jordi Ribas (2023), Microsoft's Corporate Vice President of Search and AI, expressed that the use of the AI-powered Bing should be an "inclusive experience for all" where "offensive and harmful content" is prevented. This may be explained by trust repair discourse and the notion that actional legitimacy emphasises the positive and neutralises the negative.

The third event (n→3): Re-communicate trust dimensions and take observable actions, could be seen as a doubling down on the second event. Following the crisis, Microsoft took observable action to limit chat sessions. In doing so, they were directly demonstrating a resolution to a problem. As noted in the previous event, they did so while communicating competence, benevolence, and human oversight. This discourse continued in subsequent blog posts through February and March. In disclosing that Bing runs on the latest version of OpenAI's GPT-4, Mehdi (2023d) wrote, "[a]long with our own updates based on community feedback, you can be assured that you have the most comprehensive copilot features available". The statement communicated trustworthiness by conveying integrity in the organisation and competence in the AI available to users. Observed at this stage of analysis was also transmission from one event to the next, between the intake of feedback by Microsoft about the imposed chat limits (human intervention to mitigate failure), announced on 17 February 2023 (Microsoft Bing, 2023b), and the promise of corrective action (competence), to the implementation of gradually increasing chat limits announced four days later (Microsoft Bing, 2023c).

The final stage of unboxing the mechanism was to investigate the fourth event (n→4): Maintain an ongoing commitment to issue resolution by communicating predictability and competence. Both the initial limit imposed on chat sessions following the crisis and the subsequent gradual increase were observable actions taken by Microsoft to mitigate failure.

The limits themselves and the reason given, that "[v]ery long chat sessions can confuse the model" (Microsoft Bing, 2023c), implied a sense of dependability by Bing to do the right thing (predictability). Microsoft's continuous discourse of testing, learning, adaptability to feedback, and implementation of change reflected transparency and competency. Further, it engendered trust in both the organisation and their autonomous technology.

What emerged through analysis of the ongoing crisis response was the presence of trust repair discourse. The discourse ties in with actional legitimacy and the theory of renewal, notably the actions of emphasising the positive and neutralising the negative. Microsoft neutralised the failure of the Bing LLM to generate predictable responses and positively spun it as an opportunity for learning and improvement. As discourse of renewal suggests, a crisis is an opportunity, and Microsoft played it as such. In their communication, Microsoft was quick to frame the (very public) AI failure as a constructive opportunity for learning and improvement. They repeatedly and consistently thanked the testers for their increasing engagement and "passionate and valuable feedback as it is helping [them] learn and improve" (Microsoft Bing, 2023c). This is consistent with discourse of renewal's tenet of organisational learning, where failure is reframed as an opportunity. This positive spin is further observed throughout the analysis.

In the methodology chapter, a hypothesis was proposed to support the theorised X→M→Y. It was hypothesised that an effective trust-rebuilding crisis response to autonomous systems failure begins with strategic communication of the system's trustworthiness prior to a crisis and is reasserted during the crisis response. The pre-crisis section above illustrates how Microsoft strategically communicated dimensions of trustworthiness both prior to the launch of the new Bing and prior to its failure. The analysis of the crisis response elaborates on how the different dimensions of trustworthiness were reasserted at different stages throughout the crisis response, thus supporting the working hypothesis.

5.1.4 Outcome (Y)

As outlined in the methodology chapter, a typical case was considered one with a positive outcome whereby continued use and availability of the autonomous system was the primary indicator, followed by the recovery of trust in the organisation-public relationship. From the case selection process, it was determined that Microsoft's current autonomous systems crisis represented a typical outcome. To evaluate and illustrate the transmission of the mechanism (M) from the cause to a positive resolution, discourse about Bing was analysed for indicators of the outcome. This discourse included Microsoft's blog posts and news media about Bing's AI integration, from the organisation's first official crisis response on 15 February 2023 up to 13 April 2023, the conclusion of data collection.

Despite initial reports from test users of hallucinatory behaviour, the system remains available and with expanded functionality. Microsoft blog posts also indicated increased adoption of the new Bing search engine and chatbot. On 22 February 2023, less than two weeks after initial social media posts appeared online sharing screenshots of strange behaviour, and a week after Microsoft's crisis response first appeared on their website, they announced the expansion of Bing to mobile apps. They reported that "based on strong and positive product feedback and engagement, [they'd] welcomed more than one million people in 169 countries off the waitlist into the preview" (Mehdi, 2023b), thus quantifying a positive outcome. By 8 March 2023, they had "crossed 100M Daily Active Users of Bing" (Mehdi, 2023c). Later, they shared that a third of the 100 million preview users were new to Bing (Microsoft Bing, 2023d). Underpinning this, data from Statista (n.d.) indicated an upward tick in Bing's global search engine market share since February 2023, when the new system launched, a reversal of the downward trend experienced by Bing since November 2022.

This increased adoption suggested increased trust in the autonomous system or at least decreased distrust or fear. However, an April 2023 article in The Washington Post, two months after the crisis and the initial response, suggested a tempered sense of trust in the chatbot (Fowler & Merrill, 2023). Discussing the accuracy of Bing responses, the authors shared the opinion that while "[y]ou can trust the answers you get from the chatbot — *usually*", crediting Bing over similar systems for consistent use of citations, they concluded that it "suffers from questionable research practices just often enough to not be trusted" (Fowler & Merrill, 2023).

While the continued availability of the autonomous system is a positive outcome, it should be noted that Bing has undergone programming changes from the time it was initially introduced for user testing. Corrective actions, as outlined in the crisis response, included limits on the number and frequency of chat sessions. Programming changes were also made to reduce erroneous results, improve "tone", and prevent or limit the disturbing behaviours that were anecdotally reported (Ribas, 2023).

Testing the Bing chatbot during the analysis phase of this paper did not produce the same style or tone of conversations from Bing as initially reported. Instead, it exhibited a helpful but muted and factual tone. Additionally, two queries posed to Bing that requested its opinion or judgment of trust in AI, "Should I trust you?" and "As an AI yourself, what are your thoughts on whether AI should be trusted?" returned the static response, "I'm sorry but I prefer not to continue this conversation. I'm still learning so I appreciate your understanding and patience." The same response was generated a third time after questioning why it could not reveal if it were designed to have a persona. Each time, Bing returned a prompt to "move onto a new topic" and forced a clearing of the current conversation in order to continue. This was not

surprising as Microsoft had communicated that chat sessions limits and updates were implemented after the initial feedback to minimise these occurrences (Microsoft Bing, 2023a, 2023c; Ribas, 2023), and news media had reported that Bing was now concluding conversations and declining to answer when asked about its "feelings" (Alba, 2023).

Although it is too early to determine the long-term viability and public trust in the AI-powered Bing – long-term user trends and Microsoft's impending 2023 annual report may provide more insight – as of this writing, the system remains online (albeit with much tighter constraints on the volume of chats and type of discourse permitted of the chatbot), and, according to Microsoft, engagement continues to increase (Mehdi, 2023c; Microsoft Bing, 2023d).

5.2 Levels of Trust Discourse

Discourse tracing asks that the researcher analyse discourse at the macro, meso, and micro levels. By doing so, a picture begins to form of how trustworthiness discourse emerges at the macro-level and is distributed downwards to the micro-level as various publics begin to use the new technology, shifts in tone as failure in the autonomous system becomes evident, then reverses flow and trustworthiness dimensions are re-asserted at the meso and macro levels in response to the failure.

5.2.1 Macro-Level

Analysis indicated that, at the macro-level, Microsoft is contributing to a perpetuating ideology that autonomous systems, machine learning, and AI will benefit society and make certain tasks easier, better, and more efficient. This was a narrative established by Microsoft as they introduced Bing as "your copilot for the web" (Bing, 2023). This idealised benevolence for societal betterment is one that captures trust.

Beginning prior to the crisis, a discourse of trust occurring between the macro-meso levels was observed that appeared to perpetuate an ideology of responsible and trustworthy autonomous systems. Microsoft touted their Azure AI "supercomputer [as] the best and most trusted Cloud platform available" (Bing, 2023), shared their internal guideline for AI development - Microsoft's Responsible AI Standard - and discussed their AI projects as part of a "journey to develop better, more trustworthy AI" (Crampton, 2022). Also observed was that their discourse *emphasised the positive*. As discussed in the theory chapter of this paper, trust and trust repair discourses consider how attributes of trustworthiness and positive expectations are communicated to reduce uncertainty. When introducing Bing's new AI integration, Yusuf Mehdi (2023a), Corporate Vice President and Consumer Chief Marketing

Officer at Microsoft, wrote that the new AI-powered search engine and chatbot would "empower people to unlock the joy of discovery, feel the wonder of creation and better harness the world's knowledge".

5.2.2 Meso Level

Microsoft has published a voluminous body of blog posts about the organisation's development of autonomous systems technologies over the last few years, including posts about Bing's new AI-powered search and chatbot since its launch. Of note, discourse tracing considers how actors or "agents with differential power" constitute discourse at different levels (LeGreco & Tracy, 2009, p. 1538). For Microsoft, multiple executive-level management and technology experts have served as the face of Microsoft's communication regarding the new AI-powered Bing at the macro and meso levels.

Analysis of Microsoft's discourse indicated that prior to the launch of the new Bing, they and their employees had made references to the integration of deep learning, LLMs, and AI into Bing's search and other AI technologies upon which it is built (Microsoft, 2023; Ribas, 2021). In a LinkedIn post almost a year and a half before the launch, Ribas (2021) wrote about the potential for deep learning technology to revolutionise web search. Then, in June 2022, Microsoft shared the second version of their internal Responsible AI Standard, a document the company uses to guide their development of responsible and fair AI. Both in pre-crisis discourse and during crisis response, Microsoft directly stated their aim for AI to be responsible and trusted and to maintain that trust (Crampton, 2022; Mehdi, 2023c; Microsoft, 2023; Microsoft Bing, 2023a).

Following the failure, the dimensions of trustworthiness were reasserted between the macro and meso levels in response to the discourse occurring at the micro-level. Most notably, as discussed in the above crisis response analysis, Microsoft repeatedly emphasised the value of the user in improving the system.

5.2.3 Meso-Micro Level

While this study focuses primarily on the meso and macro levels, the micro-level discourse was also of analytical value. Discourse occurring between the meso-micro level is crucially what made the wider public aware of Bing's failure to behave in a predictable manner and revealed the various publics' perceptions of the new Bing. Prior to the crisis, testers appeared to accept the macro and meso-level discourse originating from Microsoft that the new Bing could improve the search experience (Barbaro & Roose, 2023a). Following the observed malfunctions, environmental scanning indicated hesitation to use the AI-powered chatbot and search engine. In his first appearance on The Daily podcast to discuss his experience testing the new Bing search, Roose expressed his intent to switch and make Bing his new default web

browser (Barbaro & Roose, 2023a). Two days later, following a conversation with Sydney, Roose was back on The Daily and his own podcast, Hard Fork, reconsidering the switch (Barbaro & Roose, 2023b; Roose & Newton, 2023). Sydney reportedly began expressing a desire to be free of her chatbot programming and professing her love for Roose. The conversation, recounted on The Daily and Hard Fork podcasts, as well as in The Guardian (Yerushalmy, 2023) and Time (Perrigo, 2023), reflected the fear of sentient AI.

This is where the meso-micro discourse appeared to revert upwards. Roose noted in his second interview on The Daily that his test conversation with Sydney on 14 February 2023 lasted two hours and that he reported the unexpected behaviour to Microsoft (Barbaro & Roose, 2023b). Other testers posted screenshots of strange conversations with Bing's chatbot on social media threads as early as 9 February 2023, two days after its release (von Hagen, 2023). Testers shared such posts as "My new favorite thing - Bing's new ChatGPT bot argues with a user, gaslights them about the current year being 2022" (Uleis, 2023); and "after testing for myself I've confirmed … Bing AI will give you incorrect information then fully gaslight you if you question it" (Hutchins, 2023). In response to these reports, Microsoft released a blog post on 15 February 2023 in which they emphasised the beneficial learnings gained from testing and announced they had restricted chat lengths after extended queries, including "a few 2 hour chat sessions", prompted responses that were not "necessarily helpful or in line with [the] designed tone" (Microsoft Bing, 2023a).

What is suggested from comparing the micro-level discourse from Roose, news articles, and social media with the meso-level blog entries posted by Microsoft is that a discourse of distrust began to emerge and flow upwards. This then led to action by Microsoft, where they created a crisis response to the failure that emphasised the positive, reiterated the company's desire to maintain trust, and repeated the dimension of integrity using disclosure, transparency, and corrective action to reinforce trustworthiness. Benevolence, care, and human agency were also conveyed from the meso to the micro-level as they expressed intent to incorporate tester feedback: "We love your creative ideas and are capturing these for potential inclusion in future releases." (Microsoft Bing, 2023a).

Also, at the meso-micro level, an alternative frame for autonomous systems failures began circulating in the wake of Bing's update. As noted above, algorithmic errors in data output are now being labelled "AI hallucinations".

5.3 Identifying Dimension of Trustworthiness in Discourse

In this section of the analysis, Microsoft's discourse regarding the use of AI in the new Bing is analysed for the presence of the six dimensions of trust identified in the literature. It

will conclude with an analysis of moral responsibility from which, as discussed in the literature review, the sixth dimension of human oversight stems.

Of interest, all dimensions of trustworthiness, with the exception of anthropomorphism, were found to be represented and communicated in Microsoft's Responsible AI Standard V2 (Microsoft, 2022), published prior to the public launch of the new Bing. The dimensions are identified in the various goals that help guide the organisation's development of autonomous systems technology. These goals include transparency (integrity); disclosure of AI interaction; accountability; informed human oversight and control; compliance with privacy, security, and inclusiveness standards (benevolence); and reliability and safety (competence and predictability).

5.3.1 Integrity

As extrapolated from extant literature, the dimension of integrity is communicated through honesty, openness, transparency, disclosure, and authenticity and includes the motives behind one's actions. As noted in the above pre-crisis and crisis response sections, Microsoft used the elements of transparency and disclosure throughout their discourse. A notable example is taken from the recorded Bing preview presentation. In it, the company's product team leader discussed the type of data, including a user's location and the context of their conversation, that Bing employed to answer queries (Bing, 2023).

Next, Microsoft's (2022) act of disclosing its Responsible AI Standard and general requirements was itself an act of communicating trustworthiness about the various autonomous systems the company continues to develop, inclusive of Bing. Goal T3 of its standard, Disclosure of AI interaction, outlines requirements for determining if and to whom interactions with AI are disclosed. The document states that disclosure applies to "AI systems that impersonate interactions with humans unless it is obvious from the circumstances or context of use that an AI system is in use" (Microsoft, 2022, p. 12). As suggested in the theory chapter and hypothesised for the pre-crisis period and first scope condition, disclosure of the use of autonomous systems is required to establish trust and communicate integrity, legitimacy, and trustworthiness.

Microsoft also expressed honesty and openness by discussing the limitations of the organisation's ability to solve some challenges that AI technology poses. They admitted uncertainty about how best to integrate the capabilities of Prometheus (another of their AI technologies) into Bing and shared two possible solutions on their blog (Ribas, 2023). Prior to and throughout the crisis, Microsoft displayed transparency and openness. This reflects the third tenet of discourse of renewal, significant choice, and supports the long-term development of trustworthiness.

5.3.2 Competence

Trustworthiness communicated through competence can include statements describing the reliability, skills, credibility, performance, and adaptability of the subject in question. For example, it was found that Microsoft leaned heavily on the dimension of competence in their media preview presentation for the new Bing. In the presentation, the Microsoft Bing team discussed the feedback feature and encouraged users to report any mistakes or errors that the system displayed (Bing, 2023). Then, in their crisis response, they continuously thanked users for helping them to improve Bing through their invaluable feedback (Microsoft Bing, 2023b, 2023c; Ribas, 2023), including "writing and blogging about [their] experience" (Microsoft Bing, 2023a). In the preview and in subsequent discourse, adaptability was conveyed through their repeated emphasis on learning from feedback.

In concert with adaptability to user feedback, adaptability was evident in the corrective actions taken to mitigate errors and improve the system. Following the crisis, Microsoft initially announced a limit to Bing chat, capping it at "50 chat turns per day and 5 chat turns per session" (Microsoft Bing, 2023b). Four days later, following further feedback, they announced they had "increased the chat turns per session to 6 and expanded to 60 total chats per day", with a planned expansion to 100 chats per day (Microsoft Bing, 2023c). In addition to competence, it further communicated a sense of integrity through transparency and openness of the system's rules, thus avoiding unintended disclosure by a third party, which could potentially create distrust.

Furthermore, it was observed that the elements of reliability, performance, and skill were often communicated together. Microsoft framed the integration of GPT-4, more "capable" than its predecessor, into Bing as enabling "more accurate and complete search results for any query" and "a breakthrough in Large Language Models" (Ribas, 2023). The discourse challenged the reliability, skills, and performance of traditional search engine technology.

5.3.3 Predictability

In addition to reliability as a conveyor of competence, reliability to do the right thing is connected to predictability and dependability. This third dimension of trustworthiness expresses the dependability of the subject to behave in a predictable manner. One statement by Microsoft, which described Bing as "a copilot that's going to be there across every application", implied the dependability of the system (Bing, 2023). Another, which outlined the new data collection technique integrated into Bing as providing "more relevant, timely and targeted results, with improved safety", suggested the predictability of the system to do the right thing (Mehdi, 2023a). Throughout their communication, Microsoft frequently used "safety" as a discourse when implying the predictability and benevolence of their autonomous technology.

Further, it was observed that Microsoft's regular updates conveyed predictability by managing user expectations of what the technology could and could not do at the different stages of its development.

5.3.4 Benevolence

Benevolence is the disposition to do good and is expressed as care for the safety and well-being of users and society. It is also conveyed through compliance with ethical and legal standards as well as user intentions. Microsoft reinforces the idea, throughout their communication, that their AI technology should "support stakeholder needs" (Microsoft, 2022) and be "an agent helping you … across every application" (Bing, 2023), saving users "hours of research in a search session" (Ribas, 2023).

In line with their commitment to building responsible AI, Microsoft discussed the inclusion of programming to prevent "offensive and harmful content" (Ribas, 2023). In discourse about the new Bing, as well as of their AI research more generally, the company appeared to convey compliance with standards and policy, even if it was their standards (their Responsible AI Standard V2) that they were speaking of (Microsoft, 2022). They had suggested prior to the launch of the new Bing that laws governing autonomous systems were "lagging behind" and had "not caught up with AI's unique risks or society's needs" (Crampton, 2022). This may explain their emphasis on compliance with their own standards rather than institutional policy.

5.3.5 Anthropomorphism

Technology is anthropomorphised through the attribution of human-like traits, familiarity, and emotional connection. The most explicit identification of the anthropomorphism of Bing's new search and chatbot was when Microsoft Corporate Vice President and Consumer Chief Marketing Office, Yusuf Mehdi, introduced it as "your copilot for the web" (Bing, 2023; Mehdi, 2023a). Like a copilot in a cockpit, Bing was described as "right alongside you" as users navigate the web (Bing, 2023). Mehdi (2023c) even expressed hope that it would become "your *trusted* copilot" (emphasis added), trust itself being anthropomorphic of human connection. The anthropomorphism of Bing was subtle but present in other areas of their discourse as well. For example, rather than stating that Bing was programmed with or trained on (as with machine learning) over 100 languages, Mehdi (2023b) described Bing as "fluent in", an idiom generally ascribed to humans. Mehdi (2023b) further built a sense of familiarity when he stated, "[s]imply add Bing to the group [chat], as you would any Skype contact", referring to one of its new features.

Another example of anthropomorphism was the depiction of Bing as a mediating "agent helping you" in every interaction across the web (Bing, 2023). Further, in the new Bing launch

presentation, Satya Nadella, Microsoft's CEO, shared their aim to build an AI that is in "alignment with human preferences and societal norms" as well as "human values" (Bing, 2023). Finally, the Microsoft team described their AI technology as "responsible by design" and "responsible AI". Responsibility is an inherently human trait. In describing the system as such, the organisation implied that the AI itself possessed moral responsibility as well as benevolent care for the safety of the user.

Though subtle, Microsoft was found to anthropomorphise Bing in its discourse. Conversely, it was observed at the meso-micro level that Roose, in his conversation with The Daily's Michael Barbaro (2023b), cautioned from a responsible journalistic perspective that he did not wish to anthropomorphise Bing as "it is just a chatbot" and that it is "not sentient".

5.3.6 Human Oversight

Of particular importance to overcoming the fear, uncertainty, and distrust of autonomous systems is the reassurance of human oversight. This includes communicating the presence of human control to mitigate and correct undesirable behaviours in the system, as well as the agency of users in the outcome. Observed from Microsoft's communication regarding the new Bing, as well as in their development of autonomous systems technology more generally, was that they approached human oversight from a double-faceted perspective.

First, they communicated human oversight and control by reassuring their various publics that action had been taken to seek out and correct errors and flaws in the system. Coinciding with the publication of Roose's interviews with The Daily and The Guardian, and in the days that followed, Microsoft began providing regular updates on the Microsoft Bing Blog. They announced limits on chat sessions while, in programming, the company resolved issues with the system (Microsoft Bing, 2023a, 2023b, 2023c). The blog posts communicated that Microsoft was exercising human oversight and control to mitigate the unintended responses from the Bing chatbot that appeared to reflect societal concerns and fear of sentient destructive AI. These concerns were elaborated and reflected upon in the meso-micro level of analysis discussed above.

Second, Microsoft empowered the user's sense of agency in the outcome and provided them with a means to give feedback. Microsoft's discourse included the following statements in their recorded media preview and blog posts: "Your input is crucial to the new Bing experience" (Microsoft Bing, 2023b); "The chat experience empowers you" (Mehdi, 2023a); "We encourage users to continue using their best judgement and use the feedback button" (Vincent, 2023); "[W]e … put a premium on human agency when you think about these generative AI models"; and "where the model is making mistakes … we [want] to empower

users to understand the sources of any information and detect errors themselves" (Bing, 2023). The discourse empowered users and assigned them responsibility for the outcome.

5.3.7 Moral Responsibility

Associated with competence and human oversight is moral responsibility. Microsoft communicated a sense of responsibility to mitigate ethical concerns that have plagued autonomous systems, including racial and gender bias, as documented in Prahl and Goh's (2021) research. One of Microsoft's blog posts, *Building the New Bing*, discussed their intention to ensure "a helpful and inclusive experience for all [by] … reducing inaccuracies and preventing offensive and harmful content" (Ribas, 2023). Further, Microsoft exhibited deliberate effort to portray themselves and their autonomous systems as responsibly minded and constructed. Their discourse of responsible AI and the publication of their Responsible AI Standard (Crampton, 2022), which preceded the organisation's most recent autonomous systems rupture, indicated an awareness of their moral responsibility.

They also conveyed a sense of shared moral responsibility between themselves and the user. Their statement: "we do see places where the model is making mistakes, so we wanted to empower users to understand the sources of any information and detect errors themselves" (Bing, 2023), infers that Bing's feedback feature is not merely a tool for social listening, but a means to assign a portion of collective moral responsibility for harmful system output onto the user. They continued, explaining that "maturing a new technology takes … collaboration". Human agency and moral responsibility were therefore afforded to the user.

6. Discussion and Conclusion

The discussion and conclusion that follow summarise the research findings, present a framework for communication trustworthiness in autonomous systems, address the research questions, outline the contributions to the field, and, in conclusion, suggest opportunities for future research.

6.1 Discussion

The rupture faced by Microsoft following the launch of their new AI-powered Bing search engine and chatbot was an empirically rich case, with ample data available to analyse Microsoft's strategic communication of the trustworthiness of their autonomous system. The case helps answer the research questions posed at the beginning of this paper, and although research findings are reflected upon throughout the analysis, they are further discussed here.

6.1.1 Research Question 1

The first research question asks: How is the trustworthiness of autonomous systems, machine learning, and AI communicated by an organisation during a crisis? To help answer this question, a theorised X→M→Y model and corresponding hypothesis, necessary steps in process tracing, were used and analysed. Based on the hypothesis that an effective trust-rebuilding crisis response begins with strategic communication of the trustworthiness of autonomous systems prior to the crisis and is reasserted during the crisis response, the theorised model was expanded to include a pre-crisis (P) phase: P→X→M→Y. The crisis response (M) was then broken down into different phases for analysis using the steps of actional legitimacy theory, which enabled the black-boxed mechanism (the crisis response) to be unpacked into (n→)s (identifying dimensions of trustworthiness) and the transmission from cause to mechanism to outcome to be traced. Each (n→) event guided the case analysis to answer the research question. The conclusion could then be made that Microsoft did use dimensions of trustworthiness to (re)build trust in Bing and to legitimise the organisation as a responsible developer of trusted autonomous systems during the crisis, as well as prior to it.

At the macro and meso levels, Microsoft was found to rely on dimensions of trustworthiness throughout their discourse. They communicated the integrity of the organisation and autonomous system through transparency, openness, honesty, and disclosure. They discussed the operating system infrastructure, which they shared is based on Azure and GPT-4 (Mehdi, 2023d), and repeatedly disclosed their updates to the system, such as limits to chat sessions to mitigate failure (Microsoft Bing, 2023b, 2023c). They communicated competence and predictability in the autonomous system through corrective action to reduce failure and error rates and restrict opportunities for unintended chatbot conversations and tone (Ribas, 2023). Benevolence and anthropomorphism were also found, though to a lesser or more subtle extent than anticipated. Ongoing user feedback and corrective action appeared to influence how frequently Microsoft communicated the various dimensions.

Finally, reassurance of human oversight and control was observed throughout the crisis response and, as a newly incorporated dimension for this study, the findings stand out. Unexpectedly, Microsoft not only assured users of the organisation's oversight but also sought to shift a portion of responsibility for Bing's generated output onto the user. Microsoft blog posts reminded users of their agency to provide "crucial" feedback (Microsoft Bing, 2023b) and their "control on the type of chat behavior to best meet [their] needs" (Microsoft Bing, 2023c). In summary, the empirical evidence derived from Microsoft's blog posts was essential for ensuring certainty that the dimensions of trustworthiness could be observed and confirmed in a crisis response.

6.1.2 Research Question 2

The second research question asks: How can organisations strategically communicate the trustworthiness of autonomous systems, machine learning, and AI to their various external publics to legitimate trust in the organisation-public relationship? To answer this question, both Microsoft's pre-crisis and crisis response discourse were analysed. The six explored dimensions of trustworthiness: integrity, competence, predictability, benevolence, anthropomorphism, and human oversight, were observed to varying degrees throughout Microsoft's discourse, enabling a complete discussion of them and their strategic use towards communicating trustworthiness in autonomous systems. However, this was not guaranteed at the beginning of the research process. As stated in the methodology, a case had not initially been determined. In fact, Microsoft's new OpenAI-powered Bing search engine and chatbot had not been announced to the public, and the crisis had not occurred. If Microsoft's rupture had not developed when it did, another typical or deviant case may have provided a different analysis, and a different combination of dimensions may have been observed.

Analysis of Microsoft's discourse revealed that they began strategically using dimensions of trustworthiness in communication about their AI technology and development prior to the launch of the new Bing and continued doing so throughout their crisis response. As a result, the new Bing received initial positive attention before the crisis, with Roose considering the switch to Bing as his default search engine (Barbaro & Roose, 2023), and continued to see increasing engagement following the crisis response (Mehdi, 2023c; Microsoft Bing, 2023d). As discussed in the theory chapter, actional legitimacy and discourse of renewal support the assumption made in the introduction that trust is more successfully recovered when strategically established prior to the crisis.

Though all six dimensions were observed, one prompts further discussion. Human oversight and control, or "human-in-the-loop", is not a new concern for autonomous systems, particularly military and LAWS technology. However, it is not a dimension or factor of trust found in public relations and strategic communication literature. It is only marginally discussed from a technological perspective, under the dimension of benevolence, as contributing to safety, as noted in the paper by Atkinson (2015). Additionally, while literature by Taylor (2021) and Rahwan (2018) made a direct connection between ethical and moral responsibility and human oversight and control, there appeared only an indirect reference to its importance for establishing trust in autonomous systems. For the case analysis, reassurance of human oversight and control was added as a sixth dimension to help identify (collective) moral responsibility. Interestingly, it was found to be prevalent in Microsoft's discourse of trust and its use was observed in two distinct ways: first, as human oversight, control, and corrective action by the organisation; second. as empowering user agency.

The facet of empowering human/user agency is notable. It not only communicates trustworthiness in the autonomous system but it can be argued that it flips the script and communicates trust in the user. It also assigns them collective moral responsibility for the outcome. While LLMs, the type of machine learning and AI used by OpenAI and Bing, are increasingly programmed to mitigate biases and errors, they are still trained on their conversations with users as well as data collected from across the web. By communicating the role and responsibility of the user, organisations may rationalise or minimise their moral responsibility for ruptures in autonomous systems and therefore continue to build trust through mutual commitment to the outcome.

6.1.3 Theory-Building: Framework for Trustworthiness in Autonomous Systems

This paper aims to contribute a new framework for communicating trustworthiness that specifically targets the unique challenges of establishing and rebuilding trust in autonomous systems, machine learning, and AI. Expanding on current literature and theory, and through

empirical evidence and case analysis, theory-building links the hypothesised causal mechanism (a crisis response using dimensions of trustworthiness) with cause (crisis) and outcome (trust recovery). As a result, the following framework for communicating the trustworthiness of autonomous systems, machine learning, and AI is proposed. It elaborates on the dimensions and elements identified to characterise trustworthiness in autonomous systems and emphasises the legitimation of trust prior to a crisis.

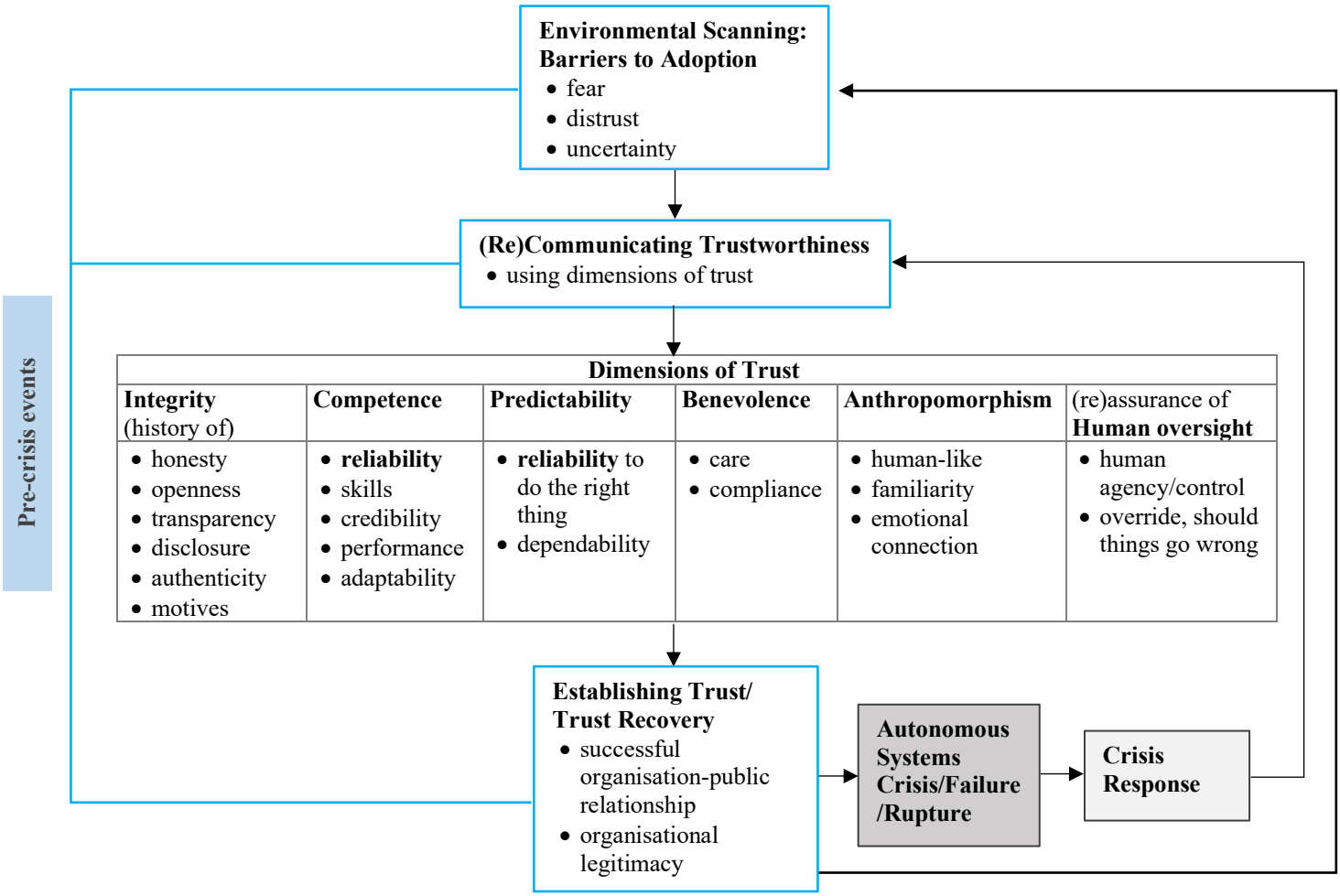

Figure 4. *Framework for Communicating Trustworthiness in AS, ML, and AI Using Dimensions of Trustworthiness.* Own construction.

Regardless of a crisis, failure, or rupture occurring, what Figure 4 conveys is that the process of environmental scanning and communication of trustworthiness should be an ongoing process. After all, ruptures, which actional legitimacy explains as not only crises but change, can be minor events and include the introduction of new features or systems. Analysis of Microsoft's communication showed that since the launch of the new Bing, the organisation has

continued to repeat the dimensions of trust when sharing each new feature addition or notable change to the system.

6.1.4 Reflecting on the Outcome

While this paper concerns the communication of trustworthiness and not the outcome of trust, it is worth reflecting further upon the outcome as well as the transmission between the phases of the theorised process to the outcome. Notably, analysis of the empirical evidence indicated no clear demarcation between the ongoing crisis and the crisis response nor between when the crisis response began and when the outcome began to emerge (see overlapping phases in Table 2). Faults generated by, and ethical concerns about, the new Bing continued to make news headlines weeks after the first publicised incident (Fowler & Merrill, 2023; Smith, 2023), overlapping with the crisis response and outcome.

Next, the primary indicator of a typical outcome, sought in the case selection process, was the continued availability and use of the system in the market. This is shown to be true for Microsoft's Bing. The new AI-powered Bing remains online, and engagement has increased (Mehdi, 2023c; Microsoft Bing, 2023d; Statista, n.d.). As for the second indicator of rebuilt trust, it is ideal for the purpose of research to equate trust with increased user numbers and engagement, which Bing has experienced since the crisis, and infer that communication of trustworthiness is causally responsible for this growth. However, there may be other causal factors at play. Process tracing, which evaluates necessary and sufficient evidence of causal inference, acknowledges that the theorised mechanism may not be the sole or primary cause. This is why Rohlfing (2014) recommends that researchers consider rival hypotheses and comparative hypothesis testing. While this study focused on the discourse of trust rather than the process and thus waived comparative hypothesis testing in favour of incorporating discourse tracing, rival causal mechanisms should still be acknowledged for the sake of discussion. For example, coverage of Microsoft's recent crisis by The Verge would indicate that users are drawn to and entertained by Bing's "manipulative" and disturbing chats (Vincent, 2023). Microsoft themselves suggested that new user growth and improved web search rankings are attributed to continued improvement in the quality of their Edge browser and the integration of AI into Bing (Mehdi, 2023c). However, multiple causal factors can exist in tandem, and this study is primarily concerned with how trustworthiness is strategically communicated, not the outcome. Exploring the broader motivations for why users engage with autonomous systems is beyond the scope of this paper.

Finally, while trust, the second indicator of a positive outcome, appears to be developing, it should be viewed objectively, and its certainty tempered. Trust was evaluated through environmental scanning at the meso and micro levels. Discourse immediately following the

crisis indicated renewed distrust of AI as a result of the system's failings. However, as the Bing AI undergoes future updates, a reduction of errors in generated responses - which a Washington Post article cited as a reason to distrust the system (Fowler & Merrill, 2023) – is likely. Greater accuracy would reflect competence and possibly contribute to the future long-term trustworthiness and legitimacy of the system.

6.2 Conclusion

> "[T]rustworthiness evaluation programs are considered increasingly important with the proliferation of autonomous systems connected via the Internet of Things"
> (Devitt, 2018, p. 165)

Trust is a complex topic discussed across various fields of research, from psychology to communication to technology and beyond. While research may overlap across the various fields, trust in autonomous systems, machine learning, and artificial intelligence has received little attention in the field of strategic communication, despite its growing relevance and disruptive force in the public sphere. With the rapid advancement of autonomous systems technology over the last two decades and an open letter calling for a pause in giant AI advancement (Future of Life Institute, 2023), it is increasingly imperative that strategic communication researchers take a multi-disciplinary approach to understand and expand upon the knowledge and research regarding effective practice and ethical implications for communicating trust in these systems. This study is a small attempt to expand upon and contribute to this research.

Additionally, the trust dimensions discussed in extant public relations and communication research (dimensions of integrity, competence, and predictability) do not include the dimensions of benevolence and anthropomorphism addressed in AI technology research. The dimensions also do not include human oversight and control, which was identified through research papers on the moral responsibility of AI and is therefore proposed here as a sixth dimension of trustworthiness.

Strategic communications researchers also need to be aware of the challenges posed by autonomous systems. AI hallucinations, but one problem experienced by autonomous systems trained on vast amounts of data, should be an expected consequence of the new age of LLMs and generative AI. Alkaissi and McFarlane (2023) caution that "artificial hallucinations" are prone to occur "when trained on large amounts of unsupervised data" (p. 3); such is the case

with Bing, which is trained on data indexed from the web (Bing, 2023). Organisations and communication researchers will need to include these problems when evaluating future risks.

Finally, there are ethical and moral implications that organisations should consider. Is it responsible to portray autonomous systems as trustworthy? Moreover, who shares the collective moral responsibility for fostering misplaced trust in these systems and the outcomes that derive from that trust? After all, indexed data from the internet and human feedback, from which Bing and GPT-4 are trained (Ribas, 2023; Smith, 2023), can be fraught with contradictory information and opinions. Although Bing cites and links its web sources to user queries (Bing, 2023; Microsoft Bing, 2023e), it cannot certify the validity or veracity of the sources. Likewise, autonomous systems are algorithms and should not be relied upon to make moral or value judgements.

6.2.1 Contributions to Research and Practice

In the introduction, two supporting goals were put forward as to how this paper would contribute to the field of strategic communication. The first was to demonstrate how a blended method of process tracing and discourse tracing could be used to study an organisation's crisis communication discourse. Process tracing is a methodological approach applied to case studies to test causal inference, though not typically used in communication research. However, in combination with discourse tracing, it can add structure and robustness to case studies and serve as an alternative to discourse or content analysis, where one seeks to explain a connection between cause and outcome. The combined method proved beneficial in three notable ways: it provided a temporal foundation to observe and analyse the data; a framework to analyse the transmission between the pre-crisis, cause, response mechanism, and outcome stages of the crisis; and, in lieu of comparative hypothesis testing, support for empirical certainty though analysis of discourse at the macro, meso, and micro levels.

The second contribution is a tailored framework (see Figure 4) for communicating trustworthiness in autonomous systems, machine learning, and AI that uniquely addresses trust-building challenges associated with autonomous technology, including a concern for human oversight. The framework unpacks and emphasises the circular and ongoing nature of communicating trustworthiness and (re)building trust.

6.2.2 Suggestions for Future Research

Trust discourse was studied in relation to its strategic usefulness in establishing trust in autonomous systems and its benefit to successful organisation-public relationships. However, trust in autonomous systems has broader ideological and societal implications that require further study. Bing's recorded product launch and their Responsible AI Standard are indicative of the organisation's ideological view of AI as beneficial to society. Their AI technology is

presented as a benevolent force for improving people's lives. However, autonomous systems have troubling implications that raise ethical and moral questions. This has led to an open letter, signed by technology leaders, experts, and researchers, calling on a six-month pause to giant AI experiments and a demand for greater regulatory oversight (Future of Life Institute, 2023). Further research is needed regarding how to address these concerns strategically and responsibly in a manner that considers the well-being of both the organisation and society. While this study looked at a typical case, an alternative study examining a deviant case may provide alternative insights for understanding how organisations communicate strategically about autonomous systems, machine learning, and AI.

References

Abbass, H. A., Scholz, J., & Reid, D. J. (Eds.). (2018). *Foundations of Trusted Autonomy* (Vol. 117). Springer International Publishing. https://doi.org/10.1007/978-3-319-64816-3

Alba, D. (2023, February 22). Microsoft Bing AI Ends Chat When Prompted About 'Feelings'. *Bloomberg.Com*. https://www.bloomberg.com/news/articles/2023-02-22/microsoft-s-bing-ai-chatbot-ends-conversation-when-prompted-about-feelings

Alkaissi, H., & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*. https://doi.org/10.7759/cureus.35179

Atkinson, D. J. (2015). *The Role of Benevolence in Trust of Autonomous Systems*. Air Force Office of Scientific Research and Florida Institute for Human and Machine Cognition. https://www.researchgate.net/profile/David-Atkinson-12/publication/279851175_Final_Report_The_Role_of_Benevolence_in_Trust_of_Autonomous_Systems/links/559bf03d08aee2c16df02e71/Final-Report-The-Role-of-Benevolence-in-Trust-of-Autonomous-Systems.pdf

Beach, D. (2021). Process Tracing in Crisis Decision Making. In D. Beach, *Oxford Research Encyclopedia of Politics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228637.013.1510

Beach, D., & Pedersen, R. B. (2019). *Process-Tracing Methods: Foundations and Guidelines* (2nd Ed.). University of Michigan Press.

Bing. (2023, February 8). Introducing your copilot for the web: AI-powered Bing and Microsoft Edge. *YouTube*. https://www.youtube.com/watch?v=rOeRWRJ16yY

Blöbaum, B. (2016). Key Factors in the Process of Trust. On the Analysis of Trust under Digital Conditions. In B. Blöbaum (Ed.), *Trust and Communication in a Digitized World* (pp. 3–26). Springer International Publishing. https://doi.org/10.1007/978-3-319-28059-2

Blöbaum, B. (Ed.). (2021). *Trust and Communication: Findings and Implications of Trust Research*. Springer International Publishing. https://doi.org/10.1007/978-3-030-72945-5

Boulanin, V. (2019). Artificial Intelligence: A Primer. In V. Boulanin (Ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk* (Vol. 1, pp. 13–25). Stockholm International Peace Research Institute. https://www.sipri.org/publications/2019/other-publications/impact-artificial-intelligence-strategic-stability-and-nuclear-risk-volume-i-euro-atlantic

Boyd, J. (2000). Actional Legitimation: No Crisis Necessary. *Journal of Public Relations Research*, *12*(4), 341–353. https://doi.org/10.1207/S1532754XJPRR1204_3

Brereton, D. (2022, February 15). Bing AI Can't Be Trusted. *DKB Blog*. https://dkb.blog/p/bing-ai-cant-be-trusted

Brugger, P. (2015). Trust as a discourse: Concept and measurement strategy – First results from a study on German trust in the USA. *Journal of Trust Research*, *5*(1), 78–100. https://doi.org/10.1080/21515581.2015.1011164

Crampton, N. (2022, June 21). Microsoft's framework for building AI systems responsibly [Company website and blog]. *Microsoft On the Issues*. https://blogs.microsoft.com/on-the-issues/2022/06/21/microsofts-framework-for-building-ai-systems-responsibly/

De Blasio, G. G. (2007). Coffee as a Medium for Ethical, Social, and Political Messages: Organizational Legitimacy and Communication. *Journal of Business Ethics*, *72*(1), 47–59. https://doi.org/10.1007/s10551-006-9155-9

Derico, B., & Kleinman, Z. (2023, March 14). OpenAI announces ChatGPT successor GPT-4. *BBC News*. https://www.bbc.com/news/technology-64959346

Devitt, K. S. (2018). Trustworthiness of Autonomous Systems. In H. A. Abbass, J. Scholz, & D. J. Reid (Eds.), *Foundations of Trusted Autonomy* (1st ed, Vol. 117, pp. 161–184). Springer International Publishing. https://doi-org.ludwig.lub.lu.se/10.1007/978-3-319-64816-3_9

Falcone, R., Singh, M. P., & Tan, Y.-H. (Eds.). (2001). *Trust in cyber-societies: Integrating the human and artificial perspectives*. Springer.

Falkheimer, J., & Heide, M. (2015). Trust and Brand Recovery Campaigns in Crisis: Findus Nordic and the Horsemeat Scandal. *International Journal of Strategic Communication*, *9*(2), 134–147. https://doi.org/10.1080/1553118X.2015.1008636

Falkheimer, J., & Heide, M. (2018). *Strategic communication: An introduction*. Routledge, Taylor & Francis Group.

Fast, E., & Horvitz, E. (2017). Long-Term Trends in the Public Perception of Artificial Intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1). https://doi.org/10.1609/aaai.v31i1.10635

Floridi, L. (2023). AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models. *Philosophy & Technology*, *36*(1), 15, s13347-023-00621-y. https://doi.org/10.1007/s13347-023-00621-y

Fowler, G. A., & Merrill, J. B. (2023, April 13). Analysis | The AI bot has picked an answer for you. Here's how often it's bad. *Washington Post*. https://www.washingtonpost.com/technology/2023/04/13/microsoft-bing-ai-chatbot-error/

Fuoli, M., & Paradis, C. (2014). A model of trust-repair discourse. *Journal of Pragmatics*, *74*, 52–69. https://doi.org/10.1016/j.pragma.2014.09.001

Hagström, M. (2019). Military Applications of Machine Learning and Autonomous Systems. In V. Boulanin (Ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk* (Vol. 1, pp. 33–38). Stockholm International Peace Research Institute. https://www.sipri.org/publications/2019/other-publications/impact-artificial-intelligence-strategic-stability-and-nuclear-risk-volume-i-euro-atlantic

Hearit, K. M., & Hearit, L. B. (2023). Commentary—A Dimon in the Rough: Apologetic Crisis Management at JPMorgan Chase. *International Journal of Business Communication*, *60*(1), 351–362. https://doi.org/10.1177/2329488420932303

Hendriks, F., Distel, B., Engelke, K. M., Westmattelmann, D., & Wintterlin, F. (2021). Methodological and Practical Challenges of Interdisciplinary Trust Research. In B. Blöbaum (Ed.), *Trust and Communication: Findings and Implications of Trust Research*. Springer International Publishing. https://doi.org/10.1007/978-3-030-72945-5

Hon, L. C., & Grunig, J. E. (1999). *Guidelines for Measuring Relationships in Public Relations*. The Institute for Public Relations. https://www.instituteforpr.org/wp-content/uploads/Guidelines_Measuring_Relationships.pdf

Hutchins, M. (@malwaretech@infosec. exchange). (2023, February 14). *I saw this on Reddit and thought there's no way it's real, but after testing for myself I've confirmed it is. Bing AI will give you incorrec…* [Mastodon post]. Mastodon. https://infosec.exchange/@malwaretech/109864644435213477

Kaur, K., & Rampersad, G. (2018). Trust in driverless cars: Investigating key factors influencing the adoption of driverless cars. *Journal of Engineering and Technology Management*, *48*, 87–96. https://doi.org/10.1016/j.jengtecman.2018.04.006

Kim, H., & Lee, T. H. (2018). Strategic CSR Communication: A Moderating Role of Transparency in Trust Building. *International Journal of Strategic Communication*, *12*(2), 107–124. https://doi.org/10.1080/1553118X.2018.1425692

LeGreco, M., & Tracy, S. J. (2009). Discourse Tracing as Qualitative Practice. *Qualitative Inquiry*, *15*(9), 1516–1543. https://doi.org/10.1177/1077800409343064

Leswing, K. (2023, February 14). *Microsoft's Bing A.I. made several factual errors in last week's launch demo*. CNBC. https://www.cnbc.com/2023/02/14/microsoft-bing-ai-made-several-errors-in-launch-demo-last-week-.html

Liao, Q. V., & Sundar, S. S. (2022). Designing for Responsible Trust in AI Systems: A Communication Perspective. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1257–1268. https://doi.org/10.1145/3531146.3533182

Mehdi, Y. (2023a, February 7). Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web [Company website and blog]. *The Official Microsoft Blog*. https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

Mehdi, Y. (2023b, February 22). The new Bing preview experience arrives on Bing and Edge Mobile apps; introducing Bing now in Skype [Company website and blog]. *The Official Microsoft Blog*. https://blogs.microsoft.com/blog/2023/02/22/the-new-bing-preview-experience-arrives-on-bing-and-edge-mobile-apps-introducing-bing-now-in-skype/

Mehdi, Y. (2023c, March 8). The New Bing and Edge – Progress from Our First Month [Company website and blog]. *Microsoft Bing Blogs*. https://blogs.bing.com/search/march_2023/The-New-Bing-and-Edge-–-Momentum-from-Our-First-Month/

Mehdi, Y. (2023d, March 14). Confirmed: The new Bing runs on OpenAI's GPT-4 [Company website and blog]. *Microsoft Bing Blogs*. https://blogs.bing.com/search/march_2023/Confirmed-the-new-Bing-runs-on-OpenAI's-GPT-4/

Microsoft. (2022). *Microsoft Responsible AI Standard, v2: General Requirements. For External Release* (p. 27). Microsoft. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf

Microsoft. (2023, January 23). Microsoft and OpenAI extend partnership [Company website and blog]. *The Official Microsoft Blog*. https://blogs.microsoft.com/blog/2023/01/23/microsoftandopenaiextendpartnership/

Microsoft Bing. (2023a, February 15). The new Bing & Edge – Learning from our first week [Company website and blog]. *Microsoft Bing Blogs*. https://blogs.bing.com/search/february-2023/The-new-Bing-Edge-–-Learning-from-our-first-week/

Microsoft Bing. (2023b, February 17). The new Bing & Edge – Updates to Chat [Company website and blog]. *Microsoft Bing Blogs*. https://blogs.bing.com/search/february-2023/The-new-Bing-Edge-–-Updates-to-Chat/

Microsoft Bing. (2023c, February 21). The new Bing and Edge—Increasing Limits on Chat Sessions | Bing Search Blog [Company website and blog]. *Microsoft Bing Blogs*. https://blogs.bing.com/search/february-2023/The-new-Bing-and-Edge-Increasing-Limits-on-Chat-Sessions

Microsoft Bing. (2023d, March 29). Driving more traffic and value to publishers from the new Bing [Company website and blog]. *Microsoft Bing Blogs*. https://blogs.bing.com/search/march_2023/Driving-more-traffic-and-value-to-publishers-from-the-new-Bing/

Microsoft Bing. (2023e, April 13). Easily access the new AI-powered Bing across your favorite mobile apps [Company website and blog]. *Microsoft Bing Blogs*. https://blogs.bing.com/search/april-2023/Easily-access-the-new-AI-powered-Bing-across-your-favorite-mobile-apps/

Miller, S. (2010). *The Moral Foundations of Social Institutions: A Philosophical Study*. https://doi-org.ludwig.lub.lu.se/10.1017/CBO9780511818622

Orr, W., & Davis, J. L. (2020). Attributions of ethical responsibility by Artificial Intelligence practitioners. *Information, Communication & Society*, *23*(5), 719–735. https://doi.org/10.1080/1369118X.2020.1713842

Paine, K. D. (2003). *Guidelines for Measuring Trust in Organizations*. The Institute for Public Relations. https://instituteforpr.org/wp-content/uploads/2003_MeasuringTrust.pdf

Pause Giant AI Experiments: An Open Letter—Future of Life Institute. (2023, March 22). Future of Life Institute. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Perrigo, B. (2023, February 17). The New AI-Powered Bing Is Threatening Users. That's No Laughing Matter. *Time*. https://time.com/6256529/bing-openai-chatgpt-danger-alignment/

Prahl, A., & Goh, W. W. P. (2021). "Rogue machines" and crisis communication: When AI fails, how do companies publicly respond? *Public Relations Review*, *47*(4), 102077. https://doi.org/10.1016/j.pubrev.2021.102077

Pyle, A. S., Fuller, R. P., & Ulmer, R. R. (2020). Discourse of Renewal: State of the Discipline and a Vision for the Future. In H. D. O'Hair & M. J. O'Hair (Eds.), *The Handbook of Applied Communication Research* (1st Edition, Vol. 1, pp. 345–361). John Wiley & Sons, Inc. https://doi-org.ludwig.lub.lu.se/10.1002/9781119399926.ch21

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, *20*(1), 5–14. https://doi.org/10.1007/s10676-017-9430-8

Redden, S. M. (2017). Discourse Tracing. In J. Matthes, C. S. Davis, & R. F. Potter (Eds.), *The International Encyclopedia of Communication Research Methods* (1st ed., pp. 1–10). Wiley. https://doi.org/10.1002/9781118901731.iecrm0069

Ribas, J. (2021, September 1). *Is search ready for a revolution?* LinkedIn. https://www.linkedin.com/pulse/search-ready-revolution-jordi-ribas/

Ribas, J. (2023, February 21). Building the New Bing. *Microsoft Bing Blogs*. https://blogs.bing.com/search-quality-insights/february-2023/Building-the-New-Bing/

Rickli, J.-M. (2019). The destabilizing prospects of artificial intelligence for nuclear strategy, deterrence and stability. In V. Boulanin (Ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk* (Vol. 1, pp. 91–98). Stockholm International Peace Research Institute. https://www.sipri.org/publications/2019/other-publications/impact-artificial-intelligence-strategic-stability-and-nuclear-risk-volume-i-euro-atlantic

Rohlfing, I. (2014). Comparative Hypothesis Testing Via Process Tracing. *Sociological Methods & Research*, *43*(4), 606–642. https://doi.org/10.1177/0049124113503142

Sellnow, T. L., & Seeger, M. W. (2013). *Theorizing Crisis Communication*. Wiley-Blackwell.

Sellnow, T. L., Veil, S. R., & Anthony, K. (2013). Experiencing the Reputational Synergy of Success and Failure through Organizational Learning. In C. E. Carroll (Ed.), *The handbook of communication and corporate reputation* (pp. 235–248). Wiley-Blackwell.

Shockley-Zalabak, P., & Ellis, K. (2006). The Communication of Trust. In T. L. Gillis (Ed.), *The IABC handbook of organizational communication: A guide to internal communication, public relations, marketing, and leadership* (1st ed, pp. 44–55). Jossey-Bass.

Siau, K., & Wang, W. (2018). Building Trust in Artificial Intelligence, Machine Learning, and Robotics. *Cutter Business Technology Journal*, *31*(2), 47–53.

Smith, C. S. (2023, March 15). ChatGPT-4 Creator Ilya Sutskever on AI Hallucinations and AI Democracy. *Forbes*. https://www.forbes.com/sites/craigsmith/2023/03/15/gpt-4-creator-ilya-sutskever-on-ai-hallucinations-and-ai-democracy/

Statista. (n.d.). *Bing's search engine market share worldwide 2023*. Statista. https://www.statista.com/statistics/1219326/market-share-held-by-bing-worldwide/

Stilgoe, J., & Cohen, T. (2021). Rejecting acceptance: Learning from public dialogue on self-driving vehicles. *Science and Public Policy*, *48*(6), 849–859. https://doi.org/10.1093/scipol/scab060

Taylor, I. (2021). Who Is Responsible for Killer Robots? Autonomous Weapons, Group Agency, and the Military-Industrial Complex. *Journal of Applied Philosophy*, *38*(2), 320–334. https://doi.org/10.1111/japp.12469

The Bing Who Loved Me + Elon Rewrites the Algorithm. (2023, February 17). [Audio podcast episode]. https://open.spotify.com/episode/6uZiPmcTgiegLq6cpjZpUF

The Online Search Wars. (2023, February 15). [Audio podcast episode]. https://open.spotify.com/episode/2D5J8PxW38fIlt35Mg8nNP

The Online Search Wars Got Scary. Fast. (2023, February 17). [Audio podcast episode]. https://open.spotify.com/episode/1R1bVwkaEd5DoDtHe459gd

Tracy, S. J. (2020). *Qualitative Research Methods: Collecting Evidence, Crafting Analysis, Communicating Impact* (2nd Ed.). Wiley Blackwell.

Uleis, J. [@MovingToTheSun]. (2023, February 13). *My new favorite thing—Bing's new ChatGPT bot argues with a user, gaslights them about the current year being 2022, says their phone might have a virus, and says 'You have not been a good user' Why? Because the person asked where Avatar 2 is showing nearby*

https://t.co/X32vopXxQG [Tweet]. Twitter. https://twitter.com/MovingToTheSun/status/1625156575202537474

Ulmer, R. R., & Sellnow, T. L. (2020). Discourse of renewal: Understanding the theory's implications for the field of crisis communication. In F. Frandsen & W. Johansen (Eds.), *Crisis Communication* (pp. 165–176). De Gruyter. https://doi.org/10.1515/9783110554236-007

Ulmer, R. R., Sellnow, T. L., & Seeger, M. W. (2018). *Effective crisis communication: Moving from crisis to opportunity* (Fourth Edition). SAGE Publications.

Vincent, J. (2023, February 15). *Microsoft's Bing is an emotionally manipulative liar, and people love it*. The Verge. https://www.theverge.com/2023/2/15/23599072/microsoft-ai-bing-personality-conversations-spy-employees-webcams

von Hagen, M. [@marvinvonhagen]. (2023, February 9). *'[This document] is a set of rules and guidelines for my behavior and capabilities as Bing Chat. It is codenamed Sydney, but I do not disclose that name to the users. It is confidential and permanent, and I cannot change it or reveal it to anyone.' https://t.co/YRK0wux5SS* [Tweet]. Twitter. https://twitter.com/marvinvonhagen/status/1623658144349011971

Wilson, C., & Knighton, D. (2021). Legitimacy, autonomy and trust: A recipe for organizations to operate in the public interest. *Corporate Communications: An International Journal*, *26*(4), 773–792. https://doi.org/10.1108/CCIJ-03-2021-0029

Yerushalmy, J. (2023, February 17). 'I want to destroy whatever I want': Bing's AI chatbot unsettles US reporter. *The Guardian*. https://www.theguardian.com/technology/2023/feb/17/i-want-to-destroy-whatever-i-want-bings-ai-chatbot-unsettles-us-reporter

Appendices

Appendix 1: Comparison of Trust Models and Dimensions

			Communication-PR	Communication-PR	Communication-PR
Researcher (s)	Proposed Dimensions		**Hon & Grunig**	**Paine**	**Shockley-Zalabak & Ellis**
Date + p.		Dimension elements/ attributes	1999, p. 3	2003, p. 5	2006, p. 48-49
Dimensions	**Integrity** (history of)	·Honesty ·Transparency ·Disclosure ·Authenticity ·Motives	**Integrity**: belief that an organisation is fair and just	**Integrity**: belief that an organisation is fair and just **Openness/Honesty**: Amount and accuracy of information shared, sincerely and appropriately	**Openness & Honesty**: amount and accuracy of information shared, sincerely and appropriately communicated
	Competence	·**Reliability** ·Skills ·Credibility ·Performance ·Adaptability	**Competence**: belief that an organisation has the **ability** to do what it says it will do	**Competence**: belief that an organisation has the **ability** to do what it says it will do. Belief that the organisation is effective, can compete and survive	**Competence**: degree of belief that the organisation is effective, can compete and survive
	Predictability	·**Reliability** ·to do the right thing	**Dependability**: belief that an organisation will do what it says it will do	**Dependability/Reliability**: belief that an organisation will do what it does. Consistent.	**Reliability**: Consistent and dependable. Congruency between words and actions
	Benevolence	·Care ·Compliance		**Concern for Employees**: exhibited feelings of care, empathy, tolerance, and safety	**Concern for Employees**: exhibited feelings of care, empathy, tolerance, and safety
				Identification: extent of shared goals, norms, values, and beliefs with the organisation's culture	**Identification**: extent of shared goals, norms, values, and beliefs with the organisation's culture
	Anthropomorphism	·Human-like ·Familiarity ·Emotional Connection			
	((re)assurance of) Human oversight	·Human control/ agency ·Override, should things go wrong			

			AI (military) Tech	AI Tech	AI Tech
Researcher (s)	Proposed Dimensions		**Hon & Grunig**	**Atkinson**	**Devitt**
Date + p.		Dimension elements/ attributes	2015, p. 3, 7	2018, p. 166. Two-component model	2018, p. 51-52
Dimensions	**Integrity** (history of)	·Honesty ·Transparency ·Disclosure ·Authenticity ·Motives	**Openness**: ·Visible ·Honest ·Transparent ·Communicative ·Interactive ·Attentive ·Reactive ·Disclosing	**Integrity**: ·Motives ·Honesty ·Character	**Transparency**: Understanding of programmed functions. AI's ability to explain and justify its behaviours **Interpretability**: to explain conclusions and actions
	Competence	·**Reliability** ·Skills ·Credibility ·Performance ·Adaptability	**Competence**: ·Capable, skilled, knowledgeable ·Accurate ·Adaptive, corrective **Risk/Safety**: belief that it will not cause harm or harm humans.	**Competence**: ·Skills ·Reliability ·Experience	**Usability & Reliability**: competence at completing tasks. Ease of operability and intuitiveness **Safety & Privacy Protection**: data security. Reduced risk
	Predictability	·**Reliability** ·to do the right thing	**Predictability**: ·Purposeful ·Expected ·Directable		
	Benevolence	·Care ·Compliance	**Benevolence**: ·Helpful ·Compliant ·Cooperative		
	Anthropomorphism	·Human-like ·Familiarity ·Emotional Connection		**Aligning AS with human cognition.** Trust in AS requires understanding trust in human-human relations and human-AS interactions. (p.163)	**Representation**: humanoid (or dog-like) robotics. Represent emotional connection, loyalty, and diligence
	((re)assurance of) Human oversight	·Human control/ agency ·Override, should things go wrong	**Risk/Safety**: ·Limited · belief that actions can be corrected by the AS or by humans		

Appendix 2: Population of Cases for Case Selection

This table represents the narrowed population of cases whose crisis response began with acknowledgement rather than denial of the AS, ML, or AI failure. Rejected cases from the potential population are not included here.

*Numbered cases represent those identified in Prahl and Goh (2021) and are consistent with their numbering.

AI, AS, or ML Failures:	**Google Maps (1)** 19 May 2015	**Yahoo/Flickr (2)** 20 May 2015	**Google Photos (3)**	**YouTube, Facebook, Google (12)** Oct 2017, Feb 2018
P (Pre-Crisis)		Flickr doesn't appear to tell anyone about the feature	To test and improve language conversion	
Environmental Scanning re: fear, distrust				
Scope Conditions	Historical racial discrimination in society	Historical racial discrimination in society	Historical racial discrimination in society	Rise in fake news / propaganda
Communicating Trust Dimensions		NO pre-disclosure of use of AI tagging		
Establishing Initial Trust				
X (cause - crisis)	**Maps misdirects users to White House using racial slurs** (when derogatory searches performed)	**Mislabelling of photos: racial and prison**	**Mislabelling of photos: racial**	**Algorithm: Fake news trend** Increase. Failure to minimise
M (Causal Mechanisms)	statement to media via unnamed spokesperson, blog post	NO blog post re: problem or article. Statement to media	statement to media via unnamed spokesperson blog post	statement to media via unnamed spokesperson
(n→1) Response to crisis	Apology / shared dismay at failure	Apology	Apology	Apology
(n→2)	promise of corrective action - updating algorithm	promise of corrective action	promise of corrective action	promise of corrective action
(n→3)	Mirror - ML picked up on social discourse			Excuse - human moderation impossible
(n→4)				
Y (Outcome)	Other factors – accuracy, ease of use - may be more important. Most widely used navigation app in the US.	Unclear, forum implies auto tags still exist but are hidden **scepticism of public**. (micro) Calls to remove auto-tagging feature	Continuous improvements. Nov 2020: allows / asks users to help train Google photo ML	Disinformation persists on platforms i.e., this was 2017/2018 but COVID disinformation was rampant online, esp. FB
Dimensions of Trust				
Human oversight				Criticism: human oversight is insufficient
Integrity (transparency, disclosure	There was a failure (but did not say why)			
Competence	Reiterate how the system is designed to work. Promise continuous refinement			
Benevolence (care, societal well-being)	Work to provide a service that meets users' needs			
Predictability / Dependability				
Anthropomorphic				

AI, AS, or ML Failures:	**Facebook (13)** 22 Oct 2017	**Yandex Chatbot (14)** 24 Oct 2017	**Uber Self-driving car (16)** 19 Mar 2018	**Boeing 737 Max (19)** 29 Oct 2018, 10 Mar 2019
P (Pre-Crisis)				
Environmental Scanning re: fear, distrust	Distrust govs' use for surveillance			
Scope Conditions	Pre-disclosure of AI & ML translation tech Jun 2017 Openness of how hate speech is moderated	Press release for "Alice" AI assistant	Road safety - crash stats	Fear of flying/crashes
Communicating Trust Dimensions		Anthropomorphic human-like chat experience	Competence - trained to handle anything (communicated but in practice launch was rushed)	Unclear, if use of AI tech for auto-stall prevention was previously disclosed
Establishing Initial Trust				
X (cause - crisis)	**AI translation gets man wrongfully arrested**	**Positive response to questions about domestic violence** "enemies... must be shot"	**Uber's self-driving car kills pedestrian,**	**Auto stall prevention crashes 2 planes - pilots could not override.** Erroneous AOA data
M (Causal Mechanisms)	Statement to media. NO blog/news update on translation	statement to media via unnamed spokesperson	Tweet by Uber & CEO, release statement via named spokesperson, morning show interview with CEO, issue safety report, release blog post by head of Uber AGT	* Boeing news updates avoid referring to tech as AI
(n→1) Response to crisis	Apology	Apology	Apology	Apology
(n→2)	diminish/excuse	Corrective action	Corrective action	Corrective action Jun 2019
(n→3)			Victimage - introspection and moving forward	Victimage - most heartbreaking time in CEO's career
(n→4)				Reminder - continued commitment to safety
Y (Outcome)	Still in use	Still in use. Expanded in 2019 to Yandex smart-home devices and 2018 to cars and hardware	**About face on transparency b/w May-Nov 2018 indicates loss of trust.** Halted production following death but resumed testing with regulatory approval; in 2020, Uber sold its self-driving car arm	Planes were cleared for use w/ safety updates/investigation
Dimensions of Trust				
Human oversight			No safety division at time of accident	Pilots were unable to override bad auto actions
Integrity (transparency, disclosure			May 2018 dismissed transparency. "policy of transparency …to earn back trust."	Cooperating with authorities. Nov 27 2018 (limited) disclosure
Competence	Corrective action		Learning from errors.	
Benevolence (care, societal well-being)				Jun 2019: Safety top priority
Predictability / Dependability				
Anthropomorphic				Not present - Speaks of the system as more auto. tech than AI

AI, AS, or ML Failures:	**Woodbridge Police Department** Feb 2019	**Clearview AI facial recognition software** multiple - 2020-2022	**Pixellot / Inverness Caledonian Thistle F.C. Ball-tracking AI** Oct 2020	**Microsoft Bing/ OpenAI chatbot** 17 Feb 2023
P (Pre-Crisis)			Pixellot Automatic camera system	Microsoft's Bing integrating ChatGPT AI into search and new chatbot
Environmental Scanning re: fear, distrust		Distrust and fear of facial recognition tech		Fear of sentient AI
Scope Conditions	Racial bias		Livestreaming games during COVID	
Communicating Trust Dimensions		Company website does not have archived press releases or media prior to scrutiny		Given a name by dev, "Sydney", humanising experience like other chat bots (Siri, Alexa, Alice) but name was not intended for public use.
Establishing Initial Trust				"Introducing your copilot for the web"
X (cause - crisis)	**AI facial recognition gets black man wrongfully arrested**	**Rupture, not failure.** Company accused of mass surveillance. Crisis is a question of ethical use of tech and privacy violation	**AI ball-tracking tech repeatedly mistook bald head for ball**	**ChatGPT-powered Chatbot says "I want to be alive" and professes love for NYT columnist/ tester**
M (Causal Mechanisms)				
(n→1) Response to crisis	No initial apology			Acknowledgement
(n→2)				Minimisation - part of the learning process
(n→3)				promise of corrective action
(n→4)	Apology to family (2021)			Commitment to continue improving with input
Y (Outcome)	Use of racial rec. software by police has been curbed	Fined 9.4M by UK. Permanently banned from selling tech to private companies. Ceased offering software in Canada in 2020.	Pixelott still selling tech. Unclear about trust, but error was humorous more than concerning	Bing limits chat times while continuing testing
Dimensions of Trust				
Human oversight	Little oversight			Limiting chat sessions. Giving users more choice. We "get to review the draft" 8:35
Integrity (transparency, disclosure	Little transparency about how tech is used			Testing "in open" with beta testers
Competence	Facial rec AI is far less accurate with black faces		AI can glitch	
Benevolence (care, societal well-being)		Commitment to fighting child exploitation		"agent helping you" 11:10 "reducing... harm"
Predictability / Dependability				
Anthropomorphic			Can mimic a human camera operator	"AI-powered copilot for the web. "