Cost Benefit Analysis of Online Harms

KAROLINA MARKEVICIUTE

Cost Benefit Analysis of Online Harms

Karolina Markeviciute

Durham University Business School

Durham University

A thesis submitted for the degree of

Doctor of Philosophy

2025

Abstract

The rapid evolution of digital technologies and algorithmic content curation has transformed the online information ecosystem, enabling unprecedented connectivity and knowledge exchange while intensifying harms such as privacy erosion, algorithmic targeting, and the exploitation of user vulnerabilities. These developments have deepened challenges surrounding information integrity, trust, and individual agency, particularly as online platforms facilitate the circulation of both misinformation (inadvertent inaccuracies) and disinformation (proactive deception). The latter presents acute risks in politically volatile and strategically contested contexts where informational asymmetry may be deliberately weaponised.

This thesis constructs a behavioural game-theoretic model to examine online information consumption under uncertainty and the cognitive burden of costly verification. Drawing on decision theory, behavioural economics, and contract theory, the model elucidates how users allocate cognitive effort to assess content in the presence of unreliable sources and malicious strategic intent designed to influence belief formation. The user decision process is formalised in a dynamic optimisation problem, where verification effort is governed by the trade-off

between the perceived value of accurate information and the escalating cognitive and emotional toll of engagement. While the framework provides an analytical distinction between misinformation and disinformation, its empirical validation is focused on the latter due to its acute strategic significance. Accordingly, the model is calibrated using qualitative data from semi-structured interviews with Ukrainian individuals navigating the volatile information space during the ongoing 2022 Russian invasion of Ukraine, where disinformation is widespread, substantiating its capacity to capture nuanced behavioural adaptation in high-stakes settings characterised by uncertainty and strategic manipulation.

The analytical findings reveal that users adopt diverse strategies in response to rising verification costs and informational risk, including reliance on heuristics, selective scrutiny, and withdrawal from engagement. These behavioural patterns reflect forms of constrained optimisation rather than irrationality, driven by cognitive and emotional limitations under conditions of sustained stress. Crucially, the analysis indicates that increased exposure to disinformation does not consistently prompt greater verification effort; in certain circumstances, disengagement becomes an adaptive means of preserving cognitive equilibrium and stability. These findings complicate prevailing assumptions about information resilience, highlighting that user responses may reflect survival-oriented behaviour rather than active empowerment.

By integrating empirical insight with formal modelling, the study advances understanding of informational vulnerability and user decision-making in strategically manipulated online environments. The findings emphasise the need for interventions that alleviate cognitive overload and support strategic verification behaviours. Such measures are essential for strengthening collective resilience and safeguarding the integrity of information ecosystems in the presence of coordinated digital manipulation and epistemically corrosive content.

Declaration

I, Karolina Markeviciute, hereby declare that this is entirely my own work unless referenced to the contrary in the text. No part of this thesis has previously been submitted else where for any other degree or qualification in this or any other university.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Julian Williams, for his unwavering guidance and patience throughout this project. His insightful feedback and steadfast support were invaluable, particularly in navigating some of the more complex challenges. I am also profoundly grateful to my family, whose continuous encouragement and trust served as a constant source of strength.

Contents

List of Figures

Nomenclature

Symbol	Description
$\boldsymbol{s}$	A possible state of the world
$\boldsymbol{S_s}$	Number of possible states of the world s
$\boldsymbol{\tilde{x}, \tilde{y}, \tilde{z}}$	A set of prospects/options
$\boldsymbol{x}$	Vector of wealth outcomes associated with a prospect $\tilde{x}$
$\boldsymbol{x_s}$	Wealth outcome associated with a prospect in state s
$\boldsymbol{\tilde{x}_i}$	ith available prospect/option
$\boldsymbol{x_i}$	Vector of wealth outcomes associated with a prospect $\tilde{x}_i$
$\boldsymbol{y}$	Vector of wealth outcomes associated with a prospect $\tilde{y}$
$\boldsymbol{z}$	Vector of wealth outcomes associated with a prospect $\tilde{z}$
$\boldsymbol{P}$	Vector of probabilities associated with outcomes x_s occurring in S states of the world
$\boldsymbol{p_s}$	Probability of sth state of the world occurring
$\boldsymbol{U(\cdot)}$	Utility function
$\boldsymbol{E(\cdot)}$	Expectation operator
$\boldsymbol{w}$	Vector of the level of wealth
$\boldsymbol{U^k(\cdot)}$	kth order derivative of the utility function $U(\cdot)$
$[\cdot]'$	First-order derivative
$[\cdot]''$	Second-order derivative
$[\cdot]'''$	Third-order derivative
$\boldsymbol{\pi}$	Risk premium associated with a prospect
$\boldsymbol{CE}$	Certainty equivalent associated with a prospect
$\boldsymbol{\sigma^2}$	Variance
$\boldsymbol{A(w)}$	Absolute risk aversion
$\boldsymbol{P(w)}$	Absolute prudence
$\boldsymbol{R(w)}$	Relative risk aversion
$\boldsymbol{\hat{\pi}}$	Relative risk premium associated with a prospect
$\boldsymbol{c}$	Wealth scalar in the set of quadratic utility functions
$\boldsymbol{\gamma}$	The coefficient of the degree of relative risk aversion
$\boldsymbol{I}$	The interval of the distribution of outcomes associated with a prospect
$\boldsymbol{w_s}$	Possible outcome of final wealth in state s
$\boldsymbol{\tilde{w}}$	Vector of final wealth outcomes associated with a prospect

Symbol	Description
$\boldsymbol{F_i(w_s)}$	ith continuous cumulative distribution function
$\boldsymbol{F_i(w_s)}$	ith continuous cumulative distribution function
$\boldsymbol{S(w)}$	Single-crossing property of mean-preserving spread
$\boldsymbol{L}$	The amount of loss associated with a prospect
$\boldsymbol{g}$	The amount of invested in prevention
$\boldsymbol{p(g)}$	The probability of loss after investing in risk preventing or reducing measures
$\boldsymbol{C(g)}$	The cost function of prevention
$\boldsymbol{U(g)}$	The utility function with respect to amount invested in prevention
$\boldsymbol{U^{i0}}$	The utility function in the absence of information
$\boldsymbol{U^{i}}$	The utility function in the presence of information
$[\cdot]^{\boldsymbol{i0}}$	Absence of any information
$[\cdot]^{\boldsymbol{i}}$	Presence of any information
$\boldsymbol{p^{i0}}$	Subjective probability of success associated with an action in the absence of any information
$\boldsymbol{p^{g}}$	Posterior probability of success when a good signal is expected
$\boldsymbol{p^{b}}$	Posterior probability of success when a bad signal is expected
$\boldsymbol{q}$	The ex-ante expected probability of receiving a good signal about success
$\boldsymbol{p_c}$	Benchmark probability required to match the expected utility of the hedging option derived
$\boldsymbol{U^{g}}$	The utility function associated with posterior probability of success when a good signal is expected
$\boldsymbol{U^{b}}$	The utility function associated with posterior probability of success when a bad signal is expected
$\boldsymbol{K}$	Informational premium
$\boldsymbol{CE^{i}}$	Certainty equivalent associated with a prospect in the presence of information
$\boldsymbol{CE^{i0}}$	Certainty equivalent associate with a prospect in the absence of information
$\boldsymbol{\alpha}$	A decision variable
$[\cdot]_{\boldsymbol{t}}$	Stamp of time period
$\boldsymbol{T}$	The number of time periods t = 0,...,T
$\boldsymbol{V(\cdot)}$	Indirect utility function

Symbol	Description
$\boldsymbol{P^{i0}}$	Distribution P_{i0} of the states of nature in the absence of information
$\boldsymbol{m}$	An informational signal
$\boldsymbol{M}$	The number of possible informational signals m
$\boldsymbol{V^{i}}$	Indirect utility function under presence of information
$\boldsymbol{V^{i0}}$	Indirect utility function under absence of information
$\boldsymbol{P^{m}}$	Posterior probability distribution of states of the world under the reception of signal m
$\boldsymbol{p_s^m}$	Posterior probability of sth state of the world occurring under the reception of signal m
$\boldsymbol{B(\alpha)}$	The opportunity set of decision variables
$\boldsymbol{r}$	Discount rate

Chapter 1

Introduction

1.1 Background and Overview

As this is the first-time cost-benefit and behavioural and regulatory analysis are applied in the area of online harms, I will review the core results and provide some examples to demonstrate the application in this field.

Confining the resources, time, and context, not only each substantial but also day to day occasion often evolves into a very distinctive situation which establishes an environment and a set of circumstances within which decision makers are subjected to choose the best available option that allows generating the most desirable outcome improving allocative efficiency [Mishan and Quah, 2020]. A versatile method providing a means not only to design and implement the decision-making process but to also investigate its subsequent consequences is the Cost Benefit Analysis (CBA) which may be conducted at different times over the life cycle of a decision [Boardman et al., 2017] and from the standpoint of different actors, levels and types of impact.

As such, irrespective of being personal or public, the notions of CBA are extensive enough to be employed to investigate any matter. In all these endeavours, at the core of conducting CBA methodology is the enablement to quantify the value of the consequences of a decision in monetary terms that pursuing a specific course of action may generate. As much as to put in perspective, it is important to establish the key foundational principles of the Cost-Benefit evaluation as well as closely related concepts and frameworks so as to also simultaneously provide the basis to build a custom assessment tool warranted by the diverse nature and characteristics of various issues and questions, and as it particularly relates to online harms.

Essentially, the general idea of the Cost benefit analysis revolves around determining whether the benefits associated with implementing a course of action exceed the accompanying costs and may extend to comparing the benefits of the other best possible options which otherwise, in many instances, may impose costs on the choice of foregone alternative opportunities. To this end, the analysis necessitates to identify and estimate all the benefits and costs of the available options. Confounding this quest, however, the precise benefits and costs may not only themselves be very uncertain and prone to variation over time which may arise from a variety of different sources such as those of technical nature, but also variable in their significance contingent on individual perspectives.

The purpose of this chapter is to establish the theoretical foundations that support the modelling of online harms as decision problems under uncertainty. It introduces cost–benefit analysis (CBA) and decision theory as analytical tools for understanding how individuals evaluate trade-offs between effort, information accuracy, and potential exposure to harm in digital environments. The scope of

the analysis is limited to individual-level decision-making rather than collective or societal aggregation. The chapter also outlines how these theoretical concepts connect to the modelling framework developed later in the thesis, which focuses on the optimal user verification effort under uncertain information quality.

For clarity, the term information in this thesis refers to digital content, specifically text, images, or data, encountered or shared by users within online platforms, rather than metadata or system-level information unless explicitly stated. Online harms denote welfare-reducing consequences that arise from exposure to misleading, manipulative, or harmful content, including disinformation and privacy intrusions. Verification effort refers to the time, cognitive resources, and opportunity costs an individual expends to evaluate the credibility of online information.

1.1.1 Cost-Benefit Analysis in Guiding Online Behaviour

Within the decision-making in an online environment, users are handed an unlimited source of content and activities to choose from, however, often not only just in exchange for provision of personal data but the permission to access and even control it. As a result, complex trade-offs in the online decision making related to sharing and hiding as well as exploiting and protecting personal data emerge wherein both the users and data holders, respectively, become faced with the evaluation and balancing of the associated benefits and costs.

In this thesis, the cost–benefit framework is applied at the individual level, where the choice variable x represents the decisions users make such as the level of verification effort, privacy protection, or information disclosure, undertaken

in a given online interaction. Each action incurs a private cost $C(x)$, reflecting time, attention, and cognitive resources, and yields an expected benefit $V(x)$, such as improved accuracy or social engagement. The objective of the user is to select x that maximises expected utility $U(V(x)-C(x))$. This formulation allows the CBA framework to capture user-specific trade-offs that underpin the models developed in later chapters.

Empirical literature highlights some of the costs and benefits of private information protection and sharing. Protected data may carry benefits and costs that mirror or are dual to the costs and benefits associated with disclosed data for both data subjects and data holders, at both the individual and societal levels [Acquisti et al., 2016]. Disclosed personal information can result in economic benefits for both data holders in the form of savings, efficiency gains, increased revenues through consumer tracking and as personalisation, targeted offers and promotions for data subjects. At the same time, such disclosures can be costly, and the data holders may incur costs of data breach or misuse and investment in data encryption infrastructure whereas the consumers may experience tangible and intangible costs from identity theft, spam or discrimination, and stigma or psychological discomfort respectively [Feri et al., 2016; Stone and Stone, 1990]. Similarly, protected data may have both benefits and costs for both parties and such benefits and costs may often be dual; that is, the inverse of the benefits and costs highlighted above in choosing to disclose data.

To give a brief outline into the online actions and the subsequently expected outcomes, individuals daily choose to engage in transactions involving their personal data. By querying on a search engine, the user is implicitly selling information about their interests in exchange for relevant results. By using an

online social network, in addition to their interests, users are implicitly selling information about their activities, emotions, work history, location, demographics and networks of friends and acquaintances in exchange for a new method of interacting with them. However, besides the willing online participation in social media which itself facilitates a culture of personal data disclosure, online users are often substantially unaware of the extent of behavioural targeting for which their personal data is collected and identified online via the constantly evolving technologies, such as cookies, web bugs and others allowing advertisers, website operators, social networks or search engines to track user online behaviour [McDonald and Cranor, 2010]. Regardless that such intensification of data collection may subject the user to significant enhanced benefits including reduced search costs or personalised content matching, in addition to potential increases in the costs of experiencing identity theft or discrimination, the surreptitious nature of the data collection and its applications add to the concerns of privacy which users may also suspect on the grounds of being unauthorised [Acquisti et al., 2016]. On the other hand, although research indicates that users are concerned about their privacy, including but not limited to the ambiguous dissemination of data and its use by third parties [Smith et al., 2011], they appear to have a proclivity for risky and privacy compromising behaviour online [Acquisti, 2004; Barnes, 2006; Barth and De Jong, 2017] in literature referred to as privacy paradox.

Effectively, while users tend to maintain a theoretical interest and positive attitude to the protection of privacy, they rarely translate into actual protective and preventive behaviour measures [Joinson et al., 2010; Pötzsch, 2008]. For instance, most users are informed about privacy risks online, however, they still appear to share private data in exchange for retail value and personalised ser-

vices [Acquisti and Grossklags, 2005; Sundar et al., 2013]. Although users seem to deploy privacy settings on social media to restrict and conceal circulation of their data, they seem to have no concern for the data collection in the background [Young and Quan-Haase, 2013]. A strand of research suggests that rational processes may well account for the observed paradoxical online behaviour by way of conscious-analytic profit-loss calculations in which users weigh the perceived costs and benefits associated with privacy disclosure as a means to reach a decision on a specific online action.

In particular, and against the backdrop of users navigating and deploying the ever advancing and growing digital environment driven by the consequentially increasing amount of collectable, storable, analysable and redeployable individual information, much of the research on online privacy relies on the so called privacy calculus model [Culnan and Armstrong, 1999; Dinev and Hart, 2006], which closely to the conventional CBA, regards privacy related decision making as guided by rationality where individuals weigh the anticipated costs of disclosing personal data against the potential benefits [Gerber et al., 2018]. However, no formal modelling within the literature of online behaviour and online harms has been proposed to account for the most up-to date evidence on the cross section of risks faced by online users under a variety of context dependent circumstances as the networked technology is becoming increasingly pervasive.

Building on the cost–benefit perspective, the following section introduces decision theory as the formal mechanism for analysing these trade-offs under uncertainty. It provides the mathematical foundation for how individuals evaluate risky prospects, thereby operationalising the intuitive logic of cost–benefit reasoning within a rigorous framework of expected utility.

1.2 Decision Theory

In an endeavour to fill this gap, a careful foundational account of human behaviour in the decision making is required. Although much of psychological, economics and finance research suggests that most individuals typically exhibit aversion toward uncertainty and risk [Schneider and Lopes, 1986], beside the nonconformist minority, the choice over the same actions among different individuals is expected to differ on the basis of different tastes, attitudes and preferences as well as the specificity of situation and context [Eckel and Grossman, 2008].

Illuminating the decision making process, a contribution by Bernoulli [1954] posits that mathematical expectations as those based on prices and wealth are an inadequate measure of value as they ignore the particular circumstances and standpoint of the appraising individual and render the results of valuations equal and uniform to every person. To put the postulations made by Bernoulli into context of benefits and costs from their original price and wealth perspective, it is the expected and subjective satisfaction extracted from a monetary outcome and not the monetary outcome itself that determines the ranking of available actions.

Bernoulli [1954], and later expanded by Von Neumann and Morgenstern [2007] encapsulates that the degree of satisfaction derived from a course of action varies with individual subjective preferences adding to the complexity of the decision-making process as they recalibrate the objective mathematical expected value of the consequent benefits and costs into a subjective change in welfare specific to the factors and influencing personal attitudes and perspective, and effectively, imply a non-linear relationship between the identified monetary outcome and the

satisfaction attained from it.

To characterise this relationship in technical terms, for every level of monetary outcome x_i associated with a specific course of action, a Utility Function $U(\cdot)$, which is a mathematical embodiment of personal tastes and preferences,- is used to quantify the level of satisfaction, obtaining utility $U(x_i)$ from the x_i monetary outcome. Since benefits and costs are often uncertain, the utilities of each possible monetary outcome x_i from an action are then weighed according to the probability p_s of that particular outcome x occurring in state s of S states of the world and summed in order to arrive at the expectation of $U(x)$. Essentially, the straightforward intuition behind this relationship is the maximisation of expected utility, however, proving it was a complicated matter.

In online environment, the value of maintaining some and certain personal data private and the value of disclosing it almost always are entirely dependent on the context as well as contingent on intrinsically uncertain combinations of states of the world [Acquisti et al., 2013]. Furthermore, privacy preferences and attitudes are subjective and idiosyncratic and the information which may be construed as sensitive and hence its value is likely to vary across individuals [Tufekci, 2008]. Regardless of the subjective valuations and preferences, however, personal data also bear substantial economic value differentiable on different dimensions such as circumstances, context, time or granularity, which companies have a vested interest in extracting [Chellappa and Sin, 2005; Norberg et al., 2007]

To provide some perspective into the essence and fit of Expected Utility (EU) methodology for CBA, although having formally been coined within the confines of ethical philosophy as a constituent of utilitarianism providing a gauge in the intrinsic pursuit of pleasure maximisation, the conceptual conjectures of utility

were already cropping up in the study of economics and finance. Inducing some etymological confusion, the term was borrowed and acquired an alternative technical meaning in the fields of economics and finance as either determining or representing tastes and preferences of different individuals. Within this thesis, Expected Utility provides the formal structure for the cost–benefit reasoning introduced earlier by representing online user decisions as choices that maximise expected utility over uncertain benefits and costs.

Upon this adaptation departing from philosophical contentions, utility was repurposed as tool of universal application in economic and financial analyses [Schoemaker, 1982], including but not limited to the decision theory wherein utility is used to model the subjective value of a choice based on the premise of rationality according to which individuals are capable of ranking their selections in a consistent order of their preferences to obtain optimal levels of benefits, thus facilitating the quantification of trade-offs in risk and reward for diverse decision-making contexts.

More specifically, developed over the course of the twentieth century to operationalise mathematical analysis of economic and financial problems, [Hicks and Allen, 1934; Pareto, 1972] the axiomatic utility theory was laid out formally proving that preferences satisfying a set of technical conditions may be represented by a utility function, which by illustration of two options ascribes a higher utility to the one of which generally more is preferred in the classical sense of rational utility maximisation in the absence of uncertainty. The principles governing such rational preferences over prospects were established in the two following axioms:

Axiom 1. *Completeness.*

For any $\tilde{x}$, $\tilde{y}$*: either* $\tilde{x} \preceq \tilde{y}$ *or* $\tilde{y} \preceq \tilde{x}$.

Axiom 2. *Transitivity.*

For any $\tilde{x}$, $\tilde{y}$, $\tilde{z}$: *if* $\tilde{x}$, $\preceq \tilde{y}$ *and* $\tilde{y} \preceq \tilde{z}$ *then* $\tilde{x} \preceq \tilde{z}$.

To account for the uncertainty of prospects, the approach extended to the expected utility theory (EUT), which, following the early conceptualisations by Bernoulli, was formalised by Von Neumann and Morgenstern [2007] imposing an enhanced set of axioms proving the existence of a utility function that may be constructed to depict preferences characterising rational expected utility maximising behaviour under risky circumstances. In this endeavour, complementing the axioms of Transitivity and Completeness, Von Neumann and Morgenstern [2007] introduced further two conditions on the rational preferences over prospects:

Axiom 3. *Continuity.*

Let $\tilde{x} \preceq \tilde{y} \preceq \tilde{z}$. *Then there is a* $p \in [0,1]$ *such that:* $\{p\tilde{x}, (1-p)\tilde{z}\} \sim \tilde{y}$

Axiom 4. *Independence.*

Let $\tilde{x} \preceq \tilde{y}$. *Then for any* $\tilde{z}$, *and any* $p \in [0,1]$: $\{p\tilde{x}, (1-p)\tilde{z}\} \preceq \{p\tilde{y}, (1-p)\tilde{z}\}$

In effect, the latter axioms underlie the departure from the original theory in that the utility function takes the expected utility form where the utility attached to an uncertain prospect is the probabilistic expectation of the utilities assigned to the possible outcomes of that prospect, however, in the same vein, delineating preferences and thus selecting the expected utility maximising prospects [Broome, 1991]. Despite this formal structure, in online environments, these axioms are only partially satisfied, as users face ambiguous and fast-changing information, limited attention, and context-dependent preferences that undermine the stability assumed by the classical framework.

An essential distinction in the analysis of rational preferences over prospects is the presence of two main types of utility functions referred to as the ordinal utility function and the more information-rich interval-valued cardinal utility function. The conceptions of cardinal utility trace to the marginal revolution of 1870, however, the notion of a cardinal utility function as unique only up to a positive transformation, via ongoing heated deliberations [Lange, 1934a,b; Moscati, 2013; Samuelson, 1937], was essentially stabilised and propelled into use in finance and economics by the Expected Utility Theory [Fishburn, 1970; Von Neumann and Morgenstern, 2007] on which it was derived.

In the interim, however, criticisms against the cardinal measurability and quantifiability of utility as untenable were advanced [Robbins, 1932], escalating into the conception of utility ranking preferences in an ordinal scale [Hicks and Allen, 1934; Pareto, 1972]. As opposed to cardinal utility which, in addition to rank ordering of preferences, communicates the relative magnitudes of the intervals - desirability distances- between the options in accordance with the scale of utility embodying those preferences and, thereby, the related strength of preferences, the only information conveyed by the ordinal utility is the ranking of the order of preferences from least to most preferable. That said, however, the utilities of options, either cardinal or ordinal, may only be established relative to the utilities of other options, and, by the same token, both utility functions are interpersonally incommensurable with regard to levels and units of utility. Thereby, neither allows for meaningful interpersonal comparisons [Elster and Roemer, 1993]. On the one hand, in as much as the application of either cardinal or ordinal utility may be deemed as nearly immaterial under certainty, mapping preferences into real numbers at their interval scale is essential in the expected

utility framework not only to allow for the operation of mathematical expectation but also capture the underlying attitude toward risk. On the other, being cardinal from the measurement perspective, in terms of preferences, EUT is ordinal as it is invariant to any increasing transformation and thus, only furnishes ordinal rankings of prospects [Schoemaker, 1982].

While a set of axioms are imposed on preferences to ensure that individuals behave as if they were trying to maximise the expected utility, the expected utility function is not of a unique form, and may assume different shapes as to capture heterogeneity in preferences not only across individuals, influenced by factors such as age, gender, income, education, [Dohmen et al., 2011], but also within the same individual under varying circumstances and across different decision and choice domains [Schoemaker, 1990; Weber et al., 2002].

For instance, interpersonal differences in risk preferences are embodied in the reluctance and readiness to partake in high-risk sports such as skydiving, which may also extend an example of domain-specific risk attitudes where the discrepancy occurs when an individual eager to partake in skydiving is hesitant about making risky financial decisions such as those related to insurance or investment. Another domain-specific but also highly characteristic example is purported by MacCrimmon and Wehrung [1990] who found managers to have different risk attitudes toward decisions involving personal as opposed to business money, or when evaluating financial versus recreational risks.

In this respect, of note is the salient feature of the expected utility theory capacitating insight into the implications of ubiquitous heterogeneity in risk preferences for decision-making across a broad range of concerns of different natures and thereby providing a systematic approach to investigate decisions under uncer-

tainty. More specifically, granted that utility functions preserve the descending order of preferences over uncertain options, the expected utility theory allows capturing a range of heterogenous attitudes toward risk which are generally categorised into risk averse, risk seeking and risk neural.

1.2.1 The Expected Utility Framework

To set the expected utility analysis into motion, if an individual who is assumed to live for a single period has a utility function $U(\cdot)$ and initial level of wealth w chooses to engage in an action $\tilde{x}$, which, although possibly bearing psychological outcomes, in continuance with the economic viewpoint of CBA, results in costs or benefits only construed as monetary outcomes $x = (x_1, ...x_s, ..., x_S)$ occurring with probability $p = (p_1, \ldots, p_s, ..., p_S)$. Living only for one period, the individual is assumed to immediately deploy all final wealth derived from initial wealth w and the outcome of $\tilde{x}$ Following the process of rational decision making, rather than simply taking the expectation of the action $\mathbb{E}(x)$, which obtains utility $U[\mathbb{E}(x)]$, the individual evaluates the expected utility $\mathbb{E}[U(x)]$.

For example, if a user chooses a verification effort level x that incurs cost $C(x)$ and faces a probability ρ that a piece of content is inaccurate, expected utility may be written as

$$EU(x) = (1 - \rho)\, U(V_{True} - C(x)) + \rho\, U(V_{False} - C(x)),$$

illustrating how the cost–benefit trade-off in online decisions is analysed within the expected-utility framework.

Assume a simplified setting in which a user decides whether to expend verifica-

tion effort x before sharing online content. The benefit from accurate information is $V_{True} = 10$, the loss from false information is $V_{False} = 0$, and verification incurs cost $C(x) = 2x$. The prior probability that the content is inaccurate is $\rho = 0.2$, so the probability that it is accurate is $1 - \rho = 0.8$.

Without verification ($x = 0$), expected utility is

$$EU(0) = 0.8\,U(10) + 0.2\,U(0).$$

If the user verifies ($x = 2$), new evidence revises the belief about inaccuracy to $\hat{\rho} = 0.1$, meaning the probability of accuracy is now 0.9, representing a Bayesian-type update based on improved information quality. Expected utility becomes

$$EU(2) = 0.9\,U(10 - 2) + 0.1\,U(0 - 2) = 0.9\,U(8) + 0.1\,U(-2).$$

The increase $\Delta EU = EU(2) - EU(0)$ reflects the value of information gained from verification. The larger this expected-utility improvement relative to cost, the greater the motivation to acquire information, particularly for risk-averse users who anticipate substantial losses from misinformation. This simplified reasoning anticipates the belief-updating and effort-cost mechanisms formalised later in the thesis, particularly in the general model of information (Section 1.3.3).

When risk averse preferences are observed, individuals exhibit tendency toward risk avoidance and preference for outcomes with lower uncertainty over those with higher uncertainty regardless whether the latter obtain expected outcomes of greater monetary value. To that effect, a risk averse individual always prefers a certain outcome. In the expected utility framework, at any level of wealth, such

individuals dislike every prospect providing an average outcome of zero: $\forall w$, $\forall x$ with $\mathbb{E}(x) = 0$, $\mathbb{E}[U(w + x)] \leq U[\mathbb{E}(w)]$, defining risk preferences of individuals who favour a certain over an uncertain prospect with equal or greater expected value [Rosen et al., 2003; Tversky and Fox, 1995]. For any risk averse expected utility maximiser, this equivalently translates into

$$\mathbb{E}[U(w + x)] \leq U[\mathbb{E}(w + x)]$$

Which satisfies Jensen's inequality condition by which the curvature of utility functions displaying risk averse attitudes is concave for all outcomes x and levels of wealth w. Accordingly, the expected utility framework is able to embed risk averse preferences into the shape of a utility function which, as well as being concave, in line with the basic property of rational behaviour to prefer more over less in the context of monetary decision making to elicit higher utility levels is increasing in wealth outcomes x. In mathematical terms, the increasing slope and the concavity of risk averse utility functions implies positive first- and negative second-order derivatives, respectively, $U'(x) > 0$ and $U''(x) < 0$. In simple and intuitive terms, risk averse utility is always increasing with wealth but at a decreasing rate because each additional unit of wealth yields an ever-smaller increase in subjective utility. This feature particular to risk averse preferences denotes diminishing marginal utility.

The decision making under the expected utility framework by a risk averse individual facing an uncertain prospect is investigated graphically in figure 1.1. The utility appears to be overestimated when the expected value is used for a risk averse individual. The mathematical expectation of the outcome as such

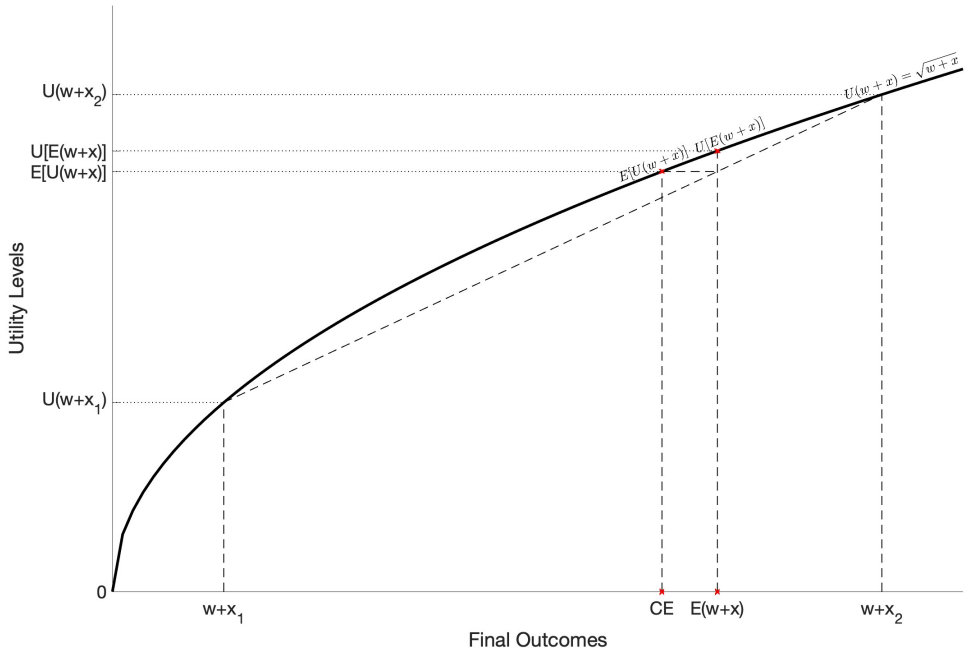

Figure 1.1: Expected Utility Analysis of Risk-Averse Individual

ignores that a person may prefer a lower risk. By weighing each possible outcome by associated probability, the expected utility takes into account the risk averse preferences by downward adjusting the value of utility.

While experimental and survey research appears to generally find the majority of individuals to display risk aversion with only few disposed to risk seeking across a variety of financial and non-financial contexts [Dohmen et al., 2011; Holt and Laury, 2002], understanding risk driven choices is warranted to establish preventive or mediating contingencies as the magnitude and significance of the repercussions associated with risky decisions may be colossal. Besides individuals still being regularly witnessed to partake in risk-taking behaviour, deviations from preferences typical to any specific individual may also occur, as exemplified by the patterns underlying the 2009 financial crisis, spreading across large portions

of the population and generating decision making bearing significant economy wide welfare implications. Therefore, modelling risk seeking behaviour may prove useful in predicting and regulating market behaviour.

To the extent that risk seeking behaviour seems to exhibit tendencies opposite to those linked to risk averse preferences, its mathematical characterisation in the Expected Utility Theory likewise runs in reverse direction:

$$\mathbb{E}[U(w+x)] \geq U[\mathbb{E}(w+x)]$$

In which case, by Jensen's inequality, the utility function is convex, although maintaining a positive slope associated with the rationale that the increasing wealth always provides a higher subjective value, and the following $U'(x) > 0$ and $U''(x) > 0$ characteristics of its first and second derivatives as represented in figure 1.2.

Having discussed the curvatures of convex and concave shapes, this also calls for a review of utility functions of a simpler linear form such as $U = a + b(w + x)$. Essentially, besides risk attitudes leading to either avoidance or pursuit of risky actions, there is another group of individuals who are indifferent in the choice to or not to take an action characterised by uncertain outcomes.

This type of behaviour ties into the linear representation of utility functions U which encapsulate risk neutrality of an individual whose preferences for uncertain prospects in the expected utility theory are ranked in the order of their expected outcome as the value of the expected utility coincides with the utility of the

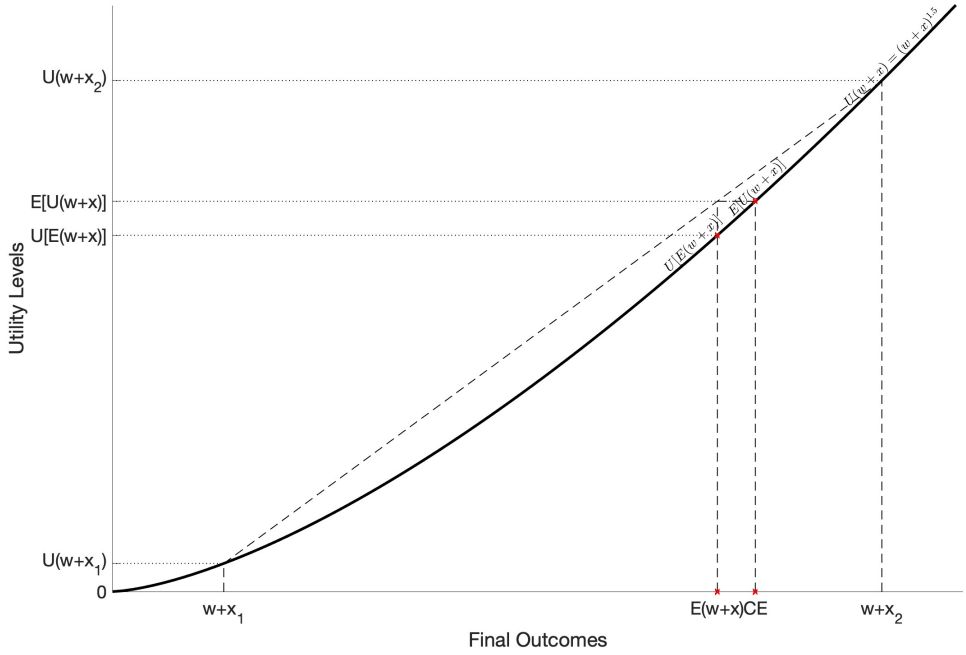

Figure 1.2: Expected Utility Analysis of Risk-Seeking Individual

expected outcome:

$$\mathbb{E}[u(w+x)] = \mathbb{E}[a + b(w+x)] = a + b(w + \mathbb{E}(x)) = U(w + \mathbb{E}(x))$$

As presented in figure 1.3, the shape of the utility representing a risk neutral individual is linear and only the level of wealth matters.

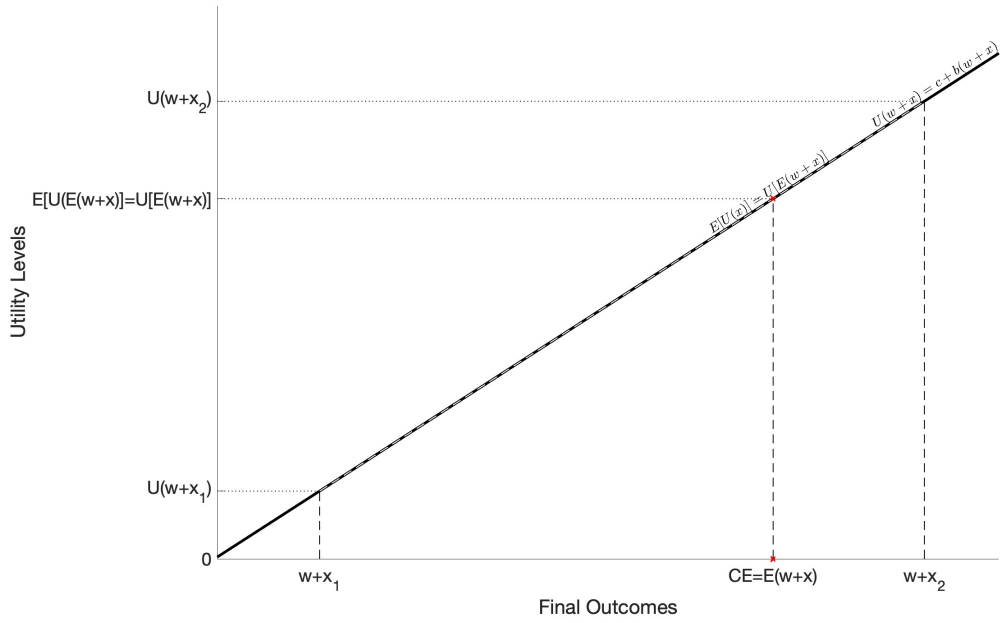

Figure 1.3: Expected Utility Analysis of Risk-Neutral Individual

1.2.2 Measuring Aversion to Risk: Risk Premium, Certainty Equivalent, Absolute Risk Aversion, Relative Risk Aversion

Whilst disliking options bearing zero-mean risks, risk averse individuals may still be involved in actions yielding sufficiently satisfactory and positive outcomes, implying a trade-off between the expected benefits and the degree of risk [Eeckhoudt et al., 2011]. Essentially, exposure to risk is inevitable and in addition to its far- and wide-reaching impact on welfare through self-directed or subrogated decision making, such as that concerning portfolio management or public policies, also imbue day-to-day tasks and choices. Despite the prevalence of risk averse preferences, the scope and the form of countermeasures taken to get rid of different

risks varies across the risk averse population. To this end, quantifying the degree of risk aversion may inform and lead to enhanced decision making and hence welfare. In the context of expected utility theory, the degree of risk aversion may be determined in relation to the monetary sacrifice an individual is willing to incur to eliminate a zero-mean risk, that is, a fair bet wherein the expected gain or loss is neutral. That is, for function U, $\forall w$, $\forall z$ and $\mathbb{E}(x) = 0$,

$$\mathbb{E}[U(w + x)] = U(w - \pi) \tag{1.1}$$

where π is the monetary sacrifice associated with the risk being eliminated, also referred to as risk premium and may be thought of as the minimum compensation required to induce readiness to bear the risk associated with an action or, conversely, the maximum payment to avoid that risk altogether. The risk premium has a convenient property of being measured in the same units as the uncertain prospect and unlike the utility itself, it is invariant to the scaling of the utility function, thereby providing a means to compare preferences and quantify the degree of risk aversion for the same risk among individuals.

Under many risky and uncertain circumstances, however, decision making is subjected to expected outcomes which differ from zero. Against this backdrop, the expected utility theory provides an insightful concept of certainty equivalent CE which measures the monetary amount that provides the same level of utility as the expected utility of an uncertain prospect:

$$\mathbb{E}[U(w + x)] = U(w + CE) \tag{1.2}$$

To put in perspective, figure 1.4 depicts CE under the setting of risk averse

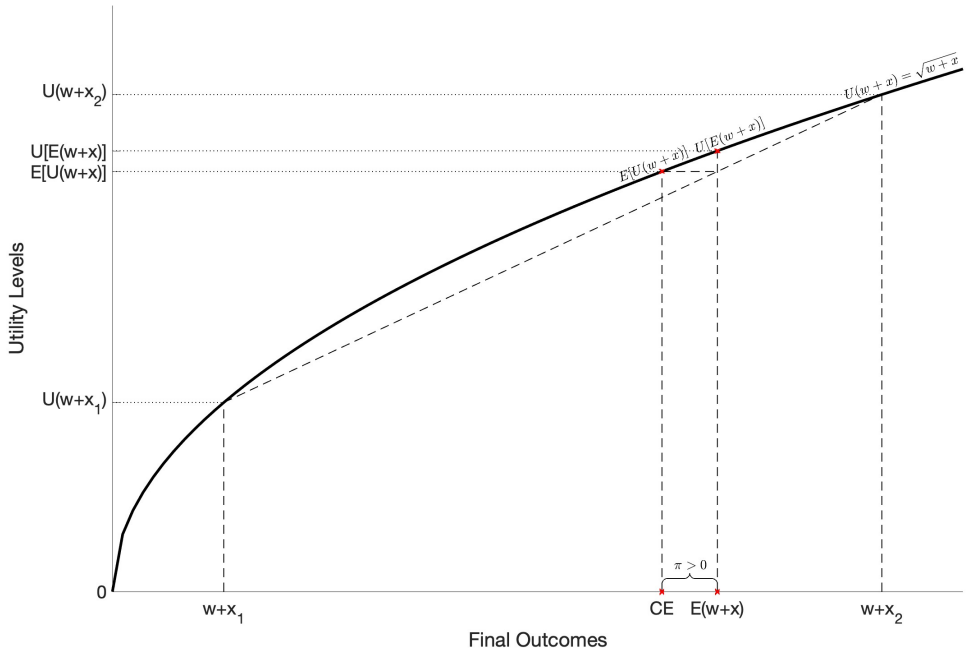

Figure 1.4: Risk Premium $\pi > 0$ of a Risk-Averse Individual

preferences. Since risk-averse individuals are prepared to accept a lower but guaranteed to a higher but uncertain outcome, CE is visibly below the expected outcome. When $\mathbb{E}(x) = 0$, comparing 1.1 and 1.2 leads to $CE = \mathbb{E}(x) - \pi$ which can be rearranged into $\pi = \mathbb{E}(x) - CE$ indicating that the distance between the expected outcome and the certainty equivalent may also serve as alternative means to measure the risk premium. Further complementing the analysis, it can be subsequently deduced that $\pi > 0$, $\pi = 0$ and $\pi < 0$ identify with concave, linear and convex utility functions representing risk-averse, risk-neutral and risk-loving preferences respectively.

As another bonus to the approach, a simple and convenient expression for the risk premium can be derived for a local analysis of small risks. This requires delving back into the equality $\mathbb{E}[u(w + x)] = U(w - \pi)$ to solve for the risk

premium π that instigates indifference toward the risk aversion driven dislike for zero mean risk $\mathbb{E}(x) = 0$, however, to which end, complicated by appearing inside the function, π has to be isolated on the left-hand side. To action this, using a second order and a first order Taylor approximation for the respective left and right-hand sides of equality, the following obtains:

$$\mathbb{E}[U(w+x)] \simeq \mathbb{E}[U(w)] + (xU'(w) + 0.5[x^2U''(x)] =$$
$$= U(w) + U'(w)\mathbb{E}(x) + 0.5U''(w)\mathbb{E}(x^2) = U(w) + 0.5\sigma^2U''(w)$$

Where $\sigma^2 = \mathbb{E}(x^2)$ is the variance of the outcome associated with the prospect x and $U(w-\pi) \simeq U(w) - \pi U'(w)$

By substituting the two into the equality, the following closed form expression for the risk premium π referred to as Arrow Prat approximation can be attained:

$$\pi \simeq \frac{1}{2}\sigma^2\frac{-U''(w)}{U'(w)} = \pi \simeq \frac{1}{2}\sigma^2 A(w) \tag{1.3}$$

It can be first observed that the risk premium is approximately proportional to the variance of the associated prospect. While variance may pose as a good measure to determine the riskiness of a prospect as seemingly borne out by its frequent use to model behaviour under risk wherein individual risk attitudes are assumed to only depend upon the mean and the variance of the underlying risks. However, the appropriateness of estimating risk by variance is contingent upon the normal distribution of outcomes or the assumption of quadratic risk preferences. Having that in mind, the validity of the approximation in 1.3 to

measure the risk premium depends on its accuracy which is generally only true when the risk is small or in very special cases [Eeckhoudt et al., 2011]. As an elaboration on its fallacy, the risk associated with a prospect may be non-normally distributed and hence, the risk premium may also depend on the other moments of the outcome distribution such as skewness and kurtosis affecting the desirability of a risk. Effectively, despite having the same mean and variance, the risks characterised by different levels and directions of skewness, and, in the same line of argument, different concentration levels within the tails of the associated outcome distributions may not necessarily yield the same risk premium.

Secondly, $\frac{-U''(w)}{U'(w)} = A(w)$ is the coefficient of the absolute risk aversion developed independently by Arrow [1963] and Pratt [1964] which measures sensitivity to risk in monetary terms. Essentially, $\frac{-U''(w)}{U'(w)}$ accounts for the curvature of the utility function which captures the rate at which marginal utility changes when wealth is increased by one monetary unit. In its own right, the first derivative of the absolute risk aversion measure $A'(w)$ specifies the change in risk preferences as a response to the change in income by the following $A'(w) < 0$, $A'(w) = 0$ and $A'(w) > 0$ results coinciding with the decreasing, constant and increasing absolute risk aversion abbreviated as DARA, CARA, IARA, respectively, and which, put in simple terms, correspond with an individual becoming more, no more or less, or less willing to accept a particular risk when their wealth increases.

Although several independent contributions were made to the decision theory in defining the notions of an increase in risk aversion and of decreasing absolute risk aversion, the most advanced by far are owed to Pratt [1964]. In later contributions, Segal and Spivak [1990] introduced the orders of risk aversion. The formulation of the degree of risk aversion, however, was not left unchallenged.

Under uncertain initial wealth, Kihlstrom et al. [1981] and Nachman [1982] disproved that the individual who is more risk averse in the sense of Arrow-Pratt will be willing to pay a higher risk premium to rid of another risk. Furthermore, Ross [1981] characterised a set of more strict conditions on the utility functions as an enhanced alternative accounting for the uncertainty of initial wealth which may also be correlated with the risk under investigation and allowing for the comparison of different degrees of risk aversion between individuals.

In light of wealth considerations, it is also intuitive that the same fixed risk tends to be perceived as more trivial by wealthier people as they generally appear to be less inclined to pay for its elimination [Pratt, 1964] The according changes in the risk premium implied by the changes in initial wealth may be captured by a set of utility functions bearing the property of prudence. In particular, the risk premium may be shown to fall as wealth increases if the condition $P(w) > A(w)$, where $P(w)$ defined as $\frac{-U'''(w)}{U''(w)}$ is the degree of absolute prudence of an agent with utility U, holds uniformly for all w, or equivalently, if $A'(w) < 0$ given that $A''(w) > 0$, or else $P(w) < A(w)$, obtains. That is, the risk premium of any risk will be decreasing in wealth if and only if either the utility function is characterised by an increasing DARA or a uniformly larger degree of absolute prudence than absolute risk aversion.

As a unit free analogy to Absolute Risk Aversion that expands the potential associated with gauging sensitivity to risk for better understanding of decision making under risk, the Relative Risk Aversion $R(w)$ (RRA) denotes the rate at which marginal utility changes when wealth is increased by one percent in the following

$$R(w) = \frac{-U''(w)}{U'(w)}w = wA(w)$$

and its first derivative $R'(w)$ similarly specifying the change in risk preferences in terms of relative risk aversion of decreasing $R'(w) < 0$, constant $R'(w) = 0$ or increasing $R'(w) > 0$ nature as a response to a percentage rather than unit-based change in wealth.

Complementarily, a unit-free relative risk premium $\hat{\pi}$ for proportional risk can also be derived as

$$\hat{\pi} \simeq \frac{\pi(wx)}{w} \simeq \frac{\frac{1}{2}w^2\sigma^2 A(w)}{w} = \frac{1}{2}\sigma^2 R(w)$$

which in addition to providing a means to analyse the portion of wealth individuals are prepared to spare to insure against any given proportional risk wherein higher risk-aversion results in a larger risk premium, may also be of value in establishing a range of acceptable degrees of risk aversion, or alternatively, a range of reasonably disposable share of wealth to eliminate zero mean risk.

Whilst absolute risk aversion is generally considered to be decreasing, there is, however, no consensus on how RRA changes with wealth due to a set of two contradictory effects. On the one hand, the intuition behind DARA dictates that, by becoming wealthier, one becomes less risk-averse, and thus risk premium becomes reduced. On the other, becoming wealthier also means exposure to a larger absolute risk x which simultaneously raises the risk premium. Therefore, there is no clear distinction which effect may dominate.

1.2.3 Most Common Utility Functions

In finance and economics literature, researchers often subject the Expected Utility analysis to a subset of utility functions to obtain tractable solutions. However, the application of any specific utility function may have significant implications for the analysis due to the assumptions imposed on risk preferences. Contingent on the choice of the precise functional form of a utility function, some results may be robust to extend to all risk-averse preferences, whilst others may only apply to a narrow class of preferences. In the face of these implications, however, several types of utility functions, have dominated the economics and finance research. For instance, a set of utility functions used throughout the momentous advances in economic and finance theory has been of the following quadratic form:

$$U(w) = cW - \frac{1}{2}w^2, \text{for } w \leqslant c$$

This set of utility functions may be operationally convenient as the expected utility generated by any distribution of the outcomes is a function of the associated mean and variance only. As EUT simplifies to a mean-variance approach to decision-making under uncertainty, however, it is highly unlikely to be appropriate to capture the risk preferences among different prospects. Moreover, the Quadratic Utility Functions are flawed in their requirement of w to be smaller than some scalar of wealth level c for U to be non-decreasing, and even more problematic is their increasing absolute risk aversion.

Just as prevalent is another classical group of the Constant Absolute Risk Aversion (CARA) Utility functions defined by their exponential functional form:

$$U(w) = \frac{exp(-cw)}{c}$$

Where c is some positive scalar. By definition, these functions exhibit $A(w) = c$ for all w and can be shown to obtain the exact Arrow-Pratt approximation when the outcomes are normally distributed. Although its invariance to changes in income may prove useful in the analysis of several alternative choices, the restrictive nature imposed on preferences by its functional format has been one of its main criticisms since DARA is generally assumed to be the plausible property exhibited by risk preferences [Hamal and Anderson, 1982; Sandmo, 1971].

Another and by far most predominantly used set of utility functions is in the form of Power Functions which by restricting risk preferences to exhibit DARA and CRRA properties have been well-received and in literature commonly referred to as CRRA class of preferences. The entire set of all CRRA Power utility functions can be defined by:

$$U(x) = \frac{w^{1-\gamma}}{1-\gamma} for \quad \gamma \geqslant 0, \gamma \neq 1, \tag{1.4}$$

$$ln(w) for \quad \gamma = 1. \tag{1.5}$$

The scalar γ is such that $\gamma > 0$, $\gamma \neq 1$ in 1.4 and can be demonstrated to be equal to the degree of RRA as $A(w) = \frac{\gamma}{w}$ and $R(w) = \gamma$ for all w. While 1.4 prohibits $\gamma = 1$, it can be easily shown that 1.5 satisfies the condition of $R(w) = 1$ for all w.

As an advantage, CRRA class of utility functions offsets any income effects in the decisions under risks, thereby considerably simplifying the analysis. On another note, empirical evidence on the shape of relative risk aversion as a function of wealth is mixed [Cohn et al., 1975; Friend and Blume, 1975; Guiso and Paiella, 2008]. Despite the widespread assumption of CRRA in estimating RRA, fewer studies directly investigate whether RRA increases or decreases with wealth [Brunnermeier and Nagel, 2008; Paravisini et al., 2017]

1.2.4 Summarising the Expected Utility Framework: Problems, Paradoxes, Alternatives

Providing reformatory foundation to the analysis of decision making under risk, the expected Utility Theory may be construed as a major paradigm shift in the theory of rational choice. However, since its conception, its axiomatisation of preferences has been the focus of much theoretical and empirical contention. The adequacy of axioms has been questioned by a series of a priori arguments or on the grounds of experimental and empirical violations, or some combination of the two, along with doubts cast by the architects of the framework [Von Neumann and Morgenstern, 2007, pp. 630] themselves. In pursuit to account for the various problematic and overly restrictive rationality constraints as well as for the empirically and experimentally observed patterns of preferences, a considerable number of generalisations of the EUT have been devised in which the standard EUT axioms have been weakened, replaced, or otherwise modified [Bolker, 1966, 1967; Jeffrey, 1965; Savage, 1954]. Given that EUT continues to retain its status as the key building block of a vast range of economic and finance theory

and analysis and as a long standing benchmark in the developments of decision-making theories, it is essential to understand the motivation behind the enormous amount of effort invested into developing alternatives to EUT. In this manner, the most representative theory of choice may be devised for application in an online environment.

Manifesting a form of impracticality for decision-making, the standard Expected Utility criterion is limited to evaluating options which come with a probability distribution over outcomes. This evokes an important distinction drawn by the economists between choice under *risk* and choice under *uncertainty*. Effectively, the former refers to situations in which knowledge or firm beliefs about objective probabilities are held, whereas the latter embody events whose objective numerical probabilities are unspecified, although, in which case, individuals may assume subjective probabilities decided on their own probabilistic beliefs [De Finetti, 1972; Finetti, 1992; Knight, 1921; LeRoy and Singell Jr, 1987; Ramsey, 1931].

This distinction between risk and uncertainty has significant implications for the CBA given that many decisions tend to be made without definitive knowledge about the costs and benefits stemming from the majority of ordinary actions, and particularly when they relate to online contexts. As a result, the decision-making process may require making a probability judgement on the relative likelihood of possible outcomes. As a variant of EUT, Savage [1954] introduced Subjective Expected Utility (SEU) positing a homogenous account of both decisions under *risk* and *uncertainty* where the preferences over prospects are ranked by the sum of utilities weighted by subjective probabilities assigned to the occurrence of outcomes due to a performance of an action and contingent on the degree of confi-

dence, which may or may not be an assessment of their objective probabilities. In this manner, preferences are translated into a subjective probability-weighted sum of the utilities of the possible outcomes associated with a prospect. Savage [1954] derived a specific set of individual necessary constraints on preference orderings guaranteeing the existence of a pair of subjective probability and utility functions relative to which the respective beliefs and desires characterising preferences may be represented as maximising subjective EU in line with the principles of rational decision making. While amounting to a system of six axioms, only three of these are necessary for SEU to represent the preferences of an expected utility maximiser, namely, the conditions of Weak Order, Weak Comparative Probability and Sure-Thing principle, from which, in particular, the latter performs a pivotal role in the derivation of Independence required for the additive representation of preferences.

1.2.4.1 Allais Paradox

In one of the most enduring counterexamples to the EUT referred to as the Allais paradox and manifested via a choice problem, Allais [1953] provided experimental evidence whereby decisions of the participants violated the Independence axiom of EUT, crucial for the representation of preferences by a function linear in probabilities. The subsequent accumulation of further experimental evidence confirmed the violation to be highly robust to variations in experimental parameters [Carlin, 1992; Kahneman and Tversky, 1979a; Slovic and Tversky, 1974]. Essentially, the emergence of the so called Allais paradox is illustrated in table 1.1 by comparing the choices of participants in a pair of decision problems exemplified in Kahneman and Tversky [1979a] wherein a choice is made between gambles A and B,

however, by a more subtle variation of the original experiment of Allais [1953], considering moderate as opposed to extremely large gains:

Table 1.1: Allais Paradox from Kahneman and Tversky [1979a]

Problem 1	Choose between				
A :	2 500 with probability	.33	B:	2 400 with certainty	
	2 400 with probability	.66			
	0 with probability	.01			
N = 72	[18]			[82]*	
Problem 2					
C:	2 500 with probability	.33	D:	2 400 with probability	.34
	0 with probability	.67		0 with probability	.66
N = 72	[83]*			[17]	

The investigation revealed that 82 per cent of the participants chose B in problem 1, but 83 per cent of the participants opted for C in problem 2, exhibiting a pattern of preferences violating the EUT in the same vein described in Allais. Based on EUT, the problems ought to yield two orders of preferences with one embodied in the following inequalities:

$$U(2,400) > 0.33U(2500) + 0.66U(2400) \text{ and } 0.34U(2400) > 0.33U(2500) \quad (1.6)$$

and the other in inequalities with reversed signs of 1.6.

Thereby demonstrating systematic inconsistencies with the observed choice, the Allais Paradox challenged the validity of the expected utility theory and of its counterpart Subjective Expected Utility as descriptive and predictive models on the basis of their restrictive Independence and Sure-Thing constraint on preferences, setting the stage for further debates, especially pertinent to the requirements of rationality itself, and thereby theoretical developments to account

for actual choice behaviour. In this thesis, such violations of the Independence and Sure-Thing principles are taken as evidence that users may distort probabilities or evaluate online outcomes in a reference-dependent manner, motivating the incorporation of behavioural elements into the later model of online harms.

1.2.5 Behaviour responses

Challenging the bedrock of most classical economic theories, in a line of arguments against the principles of perfect rationality, Simon [1978] posited the conception of bounded rationality whereby certain principles of cognition, such as perception and memory capacity, as well as constraints on resources, namely, time, money and information, delimit the ability of maximising decision-making. Motivated by bounded rationality considerations, a wide range of descriptive, normative and prescriptive accounts of decision-making departing from the assumptions of perfect rationality including some or all of the axiomatic properties of the EUT, expanding into experimental and behavioural economics. It is worth mentioning, however, that although an impressive array of models attempt to accommodate the behavioural departures in the observed choice patterns from the EUT, they tend to only give little to none consideration to the mental constraints discussed by Simon [1978] as well as being only able to account for some, but by far not all the regularities in the observed decisions. In this light, it is also notable that in consideration of decision-making online, research documents paradoxical conduct which manifests in a discrepancy between the expressed concern and the actual behaviour of users widely referred to as the phenomenon of Privacy Paradox. This discrepancy between intention and behaviour has engendered multiple theoretical

accounts of the observed patterns in decision-making online from the perspectives of perfect as well as bounded rationality, with both being guided by cost-benefit calculations [Barth and De Jong, 2017].

1.2.5.1 Prospect Theory

Often cited as a leading alternative model of choice under risk, the Prospect Theory was devised by Kahneman and Tversky [1979a] to account for the documented behavioural inclinations enacting systematic violations of EUT, such as that of Allais Paradox, evidenced by a series of experiments in laboratory settings. According to the findings of Kahneman and Tversky [1979a], individuals perceive outcomes as gains and losses rather than final levels of wealth with respect to a neutral reference point, whose position, however, may be influenced by both the framing of the offered prospect and expectations of the choice maker. As such, the manner in which a choice problem is structured, alongside the establishment of a reference point, becomes fundamental to understanding decision outcomes, owing to the distinct cognitive processes underpinning the evaluation of gains and losses [Levy, 1992]. In addition, individuals are argued to have a tendency to be risk-averse regarding gains and risk-seeking concerning losses, with the same amount of loss aggravating more than pleasing when the same amount is gained. To reflect these observed behavioural patterns, Kahneman and Tversky [1979a] propose a value function characterised by the following properties: (i) it is defined over deviations from a reference point rather than absolute wealth levels, such that any shift in the reference point results in a corresponding shift in the value function; (ii) it is generally concave for gains and convex for losses, reflecting risk aversion in the domain of gains and risk seeking in the domain of losses; and (iii)

it is steeper for losses than for gains, capturing the phenomenon of loss aversion, whereby the marginal utility of gains diminishes more rapidly than the marginal disutility of losses, thus producing an S-shaped value function.

These behavioural extensions to EUT have direct implications for analysing online behaviour. Users may over- or under-weight small probabilities when assessing the risk of encountering false or harmful content, or when evaluating the likelihood that verification effort will yield accurate information. Such distortions explain why users often neglect credible warnings or engage with dubious content despite awareness of risks. The modelling framework in later chapters incorporates this insight by allowing subjective probability distortions in users' belief-updating processes, often engendered by costly cognition, thereby embedding behavioural realism into the theoretical analysis of online harms.

In analogue fashion to EUT, the values of outcomes are then multiplied by corresponding decision weights, highlighting that Kahneman and Tversky [1979a] propose a weighting function which does not simply measure the perceived likelihood of events, but also the impact of these events on the desirability of prospects. From a technical point of view, decision weights assigned to an event may be influenced by factors other than probability such as ambiguity or uncertainty about the levels of risk or uncertainty [Ellsberg, 1961; Kahneman and Tversky, 1979a; Levy, 1992]. Kahneman and Tversky [1979a] outline several characteristics of the weighting function relating decision weights to stated probabilities. More specifically, the weighting function is not well-behaved near its endpoints as the variance in the region near 0 or 1 is large and not constant, reflecting unpredictable behaviour under extremely small or extremely large probabilities. As an aftermath feature, these regions witness a sharp indeterminant increase in

the weighting function owing to which changes in probabilities near 0 or 1 bear disproportionately large impact on the evaluation of prospects.

The next characteristic relates to the slope of the weighting function viewed as a gauge for sensitivity of decision weights and, by association, preferences, to changes in probability, as being, aside from the small region near the endpoints, less than 1 across its entire range. As a result, in the regions bar the endpoints, preferences are generally implied to be less sensitive to variations in probability than dictated by the expectation principle, resulting in the reflection of the certainty effect as the sum of the decisions weights linked to complementary events is generally less than the weight allocated to the certain event. This leads to the final features of the weighting function that large probabilities are underweighted and smaller probabilities are overweighted as borne out by experimental evidence, and thus, the property for all $0 < p < 1$, $w(p)+w(1-p) < 1$, which, referred to as subcertainty by [Kahneman and Tversky, 1979a, p.281] Kahneman and Tversky, simply put suggests the decision weights to not sum to 1 for decision between two options.

In an important departure from the standard theory of risky choice, Kahneman and Tversky [1979a] identified two phases in the decision-making process which begins with the editing phase involving a preliminary analysis of the choice problem wherein the available options, possible outcomes and their individual consequences, as well as the associated values and probabilities are identified, organised and reformulated. In turn, the edited prospects are passed onto the evaluative phase during which the prospect of highest value is selected. Notably, the former editing stage is an essential component for prospect theory to be able to rationalise violations of invariance, preference reversal, intransitivities, and other

axioms of preference [Abelson, 1985]. More precisely, it encompasses a number of mental accounting operations facilitating the decision problem referred to by Kahneman and Tversky [1979a] as coding, which identifies the reference point framing the subsequent gains and losses, simplification, which rounds off probabilities or outcomes as well as removes extreme outcomes altogether, detection of dominance, which pursues and eliminates dominated alternatives, combination, which combines probabilities of identical outcomes, segregation which separates the prospect into risk-free and risky components to evaluate the deviation between the two, and cancellation, which eliminates common components or irrelevant alternatives in prevention of preference reversals and violations of invariance.

Finally, in the evaluation phase of prospect theory, it is of note that the preferences for risk are determined jointly by the value and the weighting functions. For instance, the overweighting of probabilities is a necessary but insufficient condition to obtain risk-seeking and risk-aversion in the respective domains of gains and losses. In prospect theory, such reversal of risk attitudes may only be possible in the range of small probabilities. The specific range in which it will occur, however, depends on the relative shapes of the value and the weighting functions. It is of note, however, that Prospect Theory is not immune to critique. In particular, like many choice theories, the sum of the decision weights falling short of 1 engenders widespread violation of stochastic dominance, which stipulates that a shift of probability from less to more favourable outcomes ought to obtain an improved prospect [Fennema and Wakker, 1997; Quiggin, 2014]. On this note, such criticisms pertaining to both theoretical and empirical validities appear to be inherent in choice modelling research as borne out by the continual developments and revisions not only of EUT, but of its generalisations with

decisions theories currently numbering well into double digits [Starmer, 2000].

1.2.5.2 Other theories

To give an indication of the intensity of effort expended on constructing models of choice in closer conformity with the facts, the vastness of alternatives to EUT to account for its violations and observed regularities such as the Allais preferences goes well beyond the Prospect Theory, encompassing other widely debated and scrutinised models, namely, Rank Dependent Utility, Regret Theory as well as the Cumulative Prospect Theory. For instance, stimulated by the violations of dominance in New Theory of Cardinal Utility by Handa [1977], Quiggin [1981, 1982] furnished an alternative approach to probability weighting referred to as Rank-Dependent Utility (RDU) for decisions under risk with known probabilities, with a subsequent complementary extension by Schmeidler [1989] for decisions under uncertainty with unknown probabilities .

Limiting discussion to the mode of operation, the key idea of RDU revolves around subjecting probability weighting to depend on the rank order of the outcomes associated with a prospect wherein only unlikely extreme outcomes rather than all unlikely events are overweighted. To address this from a technical standpoint, RDU derives decision weights as a function of cumulative probability distribution [Quiggin, 1982, 2012], rather than individual probabilities, which were applied in the previous decision weight analyses such as that of Handa [1977]. Notably, to resolve the violation of the first order stochastic dominance, Tversky and Kahneman [1992] incorporated the rank-dependent weighting method for transforming probabilities into their original prospect theory to arrive at Cumulative prospect theory 10 years later recognised by the Nobel Memorial Prize in

Economic Science.

1.2.5.3 Ellsberg Paradox

To further emphasise the pervasiveness of systematically incongruous human behaviour within consequently ceaseless developments of choice models, the Ellsberg Paradox constitutes another classic contradiction not only between actual choices and the standard implications of EUT, but also its extensions and variants such as SEU despite their plausible intuition. In this endeavour, in the same manner as Allais [1953] displayed robust violations of EUT for choice under objective risk, Ellsberg [1961] demonstrated counter examples for choice under uncertainty by conducting a number of thought experiments, which, specifically, involved ball drawing gambles in either of the scenarios with a single or with two urns. In the first scenario, participants are presented with two urns, each containing 100 red and black balls, but one in an unknown proportion and the other in a 50:50 split between the red and black. Participants were asked to choose an urn and bet on the colour to be drawn based on which receiving a $100 payoff if the chosen colour is drawn, and $0 otherwise. Sparing the detail, the participants were observed to exhibit a pattern of choice indicating preference for prospects characterised by known (subjective) probabilities or quantifiable risk over those with unknown probabilities or incalculable risks. In the alternative scenario, participants were presented with an urn containing 30 red balls and a combination of 60 black and yellow balls in an unknown ratio. Similarly, individuals received a $100 payoff given the colour bet on was drawn or $0 otherwise, however, now betting in two sets of gambles outlined in the following:

Table 1.2: Ellsberg Paradox

Set 1	*Gamble i*	*Gamble ii*
	Red	Black
Set 2	*Gamble iii*	*Gamble iv*
	Either red or yellow	Either yellow or black

Where *gamble i* and *gamble ii* present choices between a red and a black ball, and *gamble iii* and *gamble iv* displaying choices to bet that either a red or yellow, or that either a black or yellow ball will be drawn, respectively. In the sense of the subjective and standard EUTs, individuals are assumed to make a probability judgment regarding the yellow and black balls, and subsequently, evaluate the expected utility of the two gambles. It then follows that in both instances an individual will prefer betting on the *gamble i* and the *gamble iii* if they overall believe that drawing a red rather than black ball is more likely. However, documenting a violation of subjective and standard EUTs, and, particularly, their respective Sure thing and Independence conditions, Ellsberg found that although individuals strictly preferred *gamble i* over *gamble ii*, they also strictly preferred *gamble iv* over *gamble iii*, displaying a pattern of preference for gambles with a known in lieu of an unknown number of balls. More generally, such relative preference for events with a known rather than unknown ambiguous probability embodies a phenomenon which has become known as ambiguity aversion. The robustness of choice patterns as demonstrated in the Ellsberg paradox and ambiguity aversion more generally were confirmed in a series of experiments [Becker and Brownson, 1964; MacCrimmon and Larsson, 1979; Slovic and Tversky, 1974].

In response to Ellsberg preferences, a substantial number of SEU generalisations were delevoped to accomodate ambiguity aversion. Limiting the discussion

to a brief introduction into this line of theories, one of the most prominent formulations is the Maxmin Expected Utility model introduced by the by Gilboa and Schmeidler [2004]. More specifically, Gilboa and Schmeidler [2004] propose a Maxmin-EU criterion which, instead of a single probability distribution over the states of the world, assumes that an individual believes in the existence of a set of possible probability distributions associated with a prospect, and as a form of extreme pessimism, chooses the worst possible probability distribution. Put in a more technical light, Maxmin Expected Utility theory characterises an individual by a utility function and a set of posterior probabilities over which the minimum expected utility is calculated and, in turn, the prospect maximising the minimum expected utility is selected.

Essentially, a vast array of incongruities has been unearthed throughout the continual development of the choice theory only to be challenged and further propelled by the next pattern of behaviour, which, however, can also frequently emerge and only be observed in specific circumstances and contexts. Whilst there may be no ideal decision theory to account for all the possible scenarios, being able to draw on a plethora of extant theoretical and empirical research on choice behaviour, an appropriate framework for the analysis of decisions online against the backdrop of potentially harmful consequences such as those related to social media arising in the form of hatred, disinformation, abuse, manipulation, harassment, incitement, theft inflicting subsequent physical, emotional and financial repercussions, may be developed.

1.3 Stochastic Approaches to Risk

Having focused on the tools to investigate the subjective value of prospects with uncertain outcomes which may consist of both or either costs or benefits, understanding different types of risks as well as the distribution of any specific risk itself is necessary to better gauge the properties of risk preferences. Particularly, while each individual may exhibit very different risk attitudes not only to different risks but also to any specific risk evaluated under different circumstances or different domains, the desirability for some risks may be shared by several individuals characterised by the same class of preferences . In this endeavour, the restrictions on preferences are weakened to capture the risk attitudes not limited to a single utility function but of a whole class such as that of all risk averse or only prudent individuals who identified with their particular group are analysed in unison to determine the constraints on changes in risk to which all individuals within their respective group respond in a similar fashion. This is not only salient in understanding individual behaviour, but also for decision making designed to benefit a group, such as a regulatory body making decisions on behalf of online users in relation to data protection and privacy legislation.

To this end, the theory of stochastic dominance considers certain statistical properties of distributions associated with the outcomes of available prospects allowing to infer whether there is a unanimous agreement on risk among certain classes of preferences. More specifically, stochastic dominance is a mathematical notion stemming from the theory of probabilities [Blackwell, 1953] which provides a means to compare distributions. It has been used to solve decision problems under uncertainty [Hanoch and Levy, 1975], to characterise portfolio

choices [Fishburn, 1977] as well as to compare income distributions [Atkinson et al., 1970].

1.3.1 Increases in Risk

To incite unanimity of choice across all risk averse individuals, the changes in risk may require to be mean-preserving and thereby maintaining the expected outcome. For emphasis, the focus is further cast on only those changes in risk which generate mean preserving increases in risk and aggravate all risk-averse individuals. There are at least three equivalent approaches to define such shifts in risk to make all risk-averse individuals worse off.

1.3.1.1 Zero Mean Noise and Mean Preserving Spread

As one method, uncertainty about the prospective final wealth in relation to the choice from available options $\tilde{x_i}$ may be increased by adding zero mean noises $\mathbb{E}(\tilde{\epsilon_s}) = 0$ to the different possible outcomes w_s associated with the final prospect of wealth $w + x_i = \tilde{w_i}$, each of which may obtain with probability p_s. By compounding $\tilde{w_1}$ with zero-mean noises $\tilde{\epsilon_s}$, an alternative wealth distribution $\tilde{w_2}$ may be obtained for the different wealth outcomes w_s of $\tilde{w_1}$ by replacing w_s of $\tilde{w_1}$ with $\tilde{\epsilon_s}$ where $\mathbb{E}(\tilde{\epsilon_s}) = 0$, leading to $\tilde{w_2} = w_s + \tilde{\epsilon_s}$. Therewith amplifying uncertainty, zero-mean noises always reduce the expected utility of all risk-averse individuals:

$$\mathbb{E}[U(\tilde{w_2})] = \sum_{s=1}^{n} p_s \mathbb{E}U(w_s + \tilde{\epsilon_s})] \leqslant \sum_{s=1}^{n} p_s U(w_s) = \mathbb{E}[U(\tilde{w_1})] \tag{1.7}$$

Another technique is the Mean Preserving Spread (MPS) which allows to construct a probability distribution preserving the mean but increasing the risk.

More specifically, it describes an operation which preserves the mean of either discrete or continuous distribution, but the probability mass $f_i(w)$ or probability density $f_i(\cdot)$ of $\tilde{w}_i$, respectively, is partially removed from some interval I to be transferred outside of it. Defined formally, $\tilde{w_2}$ is a MPS of $\tilde{w_1}$ if i. $\mathbb{E}(\tilde{w_2}) = \mathbb{E}(\tilde{w_1})$ and ii. there is an interval I such that $f_2(w) \leqslant f_1(w)$ for all wealth levels w in I and $f_2(w) \geqslant f_1(w)$ for all wealth levels w outside I.

Additionally, MPS can be translated into a condition on the cumulative distribution functions of the prospective final wealth $\tilde{w_1}$ and its more risky MPS $\tilde{w_2}$. Given the continuous cumulative distribution functions:

$$F_i(w) = \int^w f_i(w)dw \tag{1.8}$$

The property of mean-preserving spread where $F_2(w)$ is the MPS of $F_1(w)$, implies the following single-crossing condition

$$S(w) = \int^w [F_2(w) - F_1(w)]dw \geqslant 0 \tag{1.9}$$

for all w. To put it in less technical terms, the size of the area when $F_2(w)$ is above $F_1(w_s)$ is larger than or equal to the size of the area when $F_1(w)$ is above $F_2(w)$. In essence, this suggests that the probability density of $F_2(w)$ is more spread out than the probability density of $F_1(w)$, implying that $\tilde{w_2}$ is riskier and also has a higher variance than $\tilde{w_1}$. Effectively, $S(w) \geqslant 0$ and, equivalently, $\int^w F_2(w_s)dw \geqslant \int^w F_1(w)dw$ are both a necessary and sufficient integral condition for mean-preserving changes in risk to reduce the expected utility of all risk-averse individuals, and thereby, guarantee that every risk-averse individual unanimously dislikes it.

Accordingly, the concept of mean-preserving spreads also provides a stochastic ordering of prospects with equal means in terms of their respective degrees of risk as characterised by their probability distributions. However, an ordering as such is partial since any one of the two prospective outcomes with equal mean values may not necessarily be a MPS of the other, or have a higher variance than the other. On this note, ranking prospective outcomes by MPSs is a special case of ranking by second-order stochastic dominance, which in the instance of $\tilde{w}_2$ being a MPS of $\tilde{w}_1$, suggests that $\tilde{w}_1$ second-order stochastically dominates $\tilde{w}_2$ and all individuals represented by a set of risk-averse utility functions favour it.

1.3.1.2 Downside Risk

Another type of exacerbating changes in risk are the so called increases in downside risk which have the property of preserving both the mean and the variance associated with a prospect, by means of transferring a zero-mean risk from a richer to a poorer state of the world. In this light, however, it might not necessarily be unanimously disliked by risk averse individuals; some of the risk averse individuals may prefer it while others may dislike it. This being said, experiments show that most people generally dislike this type of shift in risk Eeckhoudt et al. [2011]; Mao [1970], implying risk aversion to downside risk. In the context of expected utility analysis, however, to ensure this result, the utility functions have to manifest the condition of prudence, indicating that in the EUT sense, all prudent individuals are averse to downside risk.

1.3.1.3 First-order Stochastic Dominance

Although extending the analysis of preferences beyond a single specific utility function, the changes in risk had to be constricted to be mean-preserving or increasing in downside risk for all risk averse or all prudent individuals, respectively, to dislike them, thereby exacting strong requirements on risk itself. By imposing unanimous preferences in an even broader group of individuals, ever stricter constraints on risk may be required. Additionally and more generally, in most decision-making situations under uncertainty a trade-off between risk and expected outcome arises and thus, prospects and actions involving higher risk are less likely to obtain outcomes with similar mean values. For instance, by agreeing to share more personal data online, higher benefits may be reaped, however, simultaneously increasing the exposure to cybersecurity risks. Against the backdrop of such shifts in risk, the theory of stochastic dominance may provide a practical means to establish more general tendencies and patterns of preferences within a population. More specifically, under a sole assumption of continuous and monotonically increasing utility functions, unanimous dislike for a prospect held by a considerably large group of individuals may be investigated using the First-Order Stochastic Dominance (FSD).

From a technical perspective, the cumulative distribution function $F_1(w)$ of prospective wealth $\tilde{w}_1$ is said to first-order stochastically dominate the cumulative distribution function $F_2(w)$ of prospective wealth $\tilde{w}_2$ if

$$F_2(w) \geqslant F_1(w) \tag{1.10}$$

suggesting that for every individual with an increasing utility function $U' > 0$,

$\tilde{w_1}$ dominates $\tilde{w_2}$, and $\mathbb{E}[U(\tilde{w_2})]$ is lower than $\mathbb{E}[U(\tilde{w_1})]$. This simply implies that acquiring a lower value of the final outcome $\tilde{w}$ is more likely under $F_2(w)$ as it is always above $F_1(w)$. If the two distributions cross, however, there is no FSD.

Moreover, viewed in the context of probability distributions, the mean value of $F_1(w)$ is higher than that of $F_2(w)$, with $F_1(w)$ being to the right of $F_2(w)$. Generally, when the first prospect FSD dominates the second one, the first will have a higher mean. However, if there is a third option which has a higher mean than the first, FSD cannot be infer solely based on the mean values of distributions.

1.3.2 Optimal Prevention

While uncertainty may resolve over time as well as insurance may be taken, it is also often possible to alter risk itself. Navigating online harms, individuals acquire antivirus software to reduce the risk of a virus or set up a two-factor authentication to prevent unwanted access and data breaches. Such risk-reducing actions are generally referred to as loss control. However, the precise manner in which distribution is modified by risk reduction mechanism might be rather complex as, for example, software itself may have risks of their own.

Moreover, self-protection and loss prevention is an endeavour itself which although may allow to decrease the probability of an adverse outcome, it may be costly to pursue. This generates a trade-off balancing which requires to strike the optimal EU maximising level of effort. However, frequently cost-benefit analysis of prevention is analysed under risk neutrality where only the expected value of loss is considered. This overlooks the desire to reduce variability of losses. There-

fore, risk aversion should be taken into consideration when analysing the costs and benefits of preventative actions. It is possible that in the extreme case of loss prevention, the risk might be avoided entirely. For instance, some individuals might decide not to sign up to a website to gain some free access if it requires to input card details on the website as either it might later on charge a fee or because of security concerns.

To gain perspective, a base case of risk neutrality is first considered where a risk-neutral individual is subject to the risk of losing an amount L with probability p. Assuming there is a preventative mechanism in which individual may invest a monetary amount g, the probability of damage L then becomes $p(g)$, whereas p is assumed to be differentiable twice, decreasing and convex function characterised by $p' < 0$ and $p'' \geqslant 0$. The objective of the choice problem is to select e to minimise the net expected cost of the risk accounting for the cost of prevention $C(g)$ which may be written as:

$$g^n \in arg \min_{g \geqslant 0} C(g) \equiv g + p(g)L \tag{1.11}$$

The solution of the optimal preventive investment g^n for the risk-neutral individual is defined by

$$-p'(g^n)L = 1$$

where the left-hand side of the equality is the marginal benefit of prevention, which implies the expected reduction of loss when one more monetary unit is invested in prevention as well as denotes the classical optimality condition of marginal cost being equal to the marginal benefit. Because full elimination of risk is usually very costly, the probability of damage typically remains positive,

$p(g^n) > 0$. It is worth noting, however, that risk neutrality may only serve as a good approximation when the risk is small or can be diversified away.

1.3.2.1 Risk Aversion and Optimal Prevention

Shifting away to a more general EU framework wherein a risk averse individual who is endowed with wealth w faces the risk of losing the amount L with probability $p(g)$, the decision problem under which preventative measure may be take can be expressed as:

$$g^* \in arg\min_{g \geqslant 0} U(g) = p(g)U(w-g-L) + (1-p(g))U(w-g) \tag{1.12}$$

Although it might seem intuitive that risk-aversion may induce more investment in risk prevention, and vice versa under risk-seeking preferences, it may not necessarily be the case. The examination of the result of risk-neutral g^n against that of risk-averse g^* can confirm it, particularly, when $U(g)$ is assumed to be concave, a higher g^* than g^n obtains if and only if $U'(g^n) > 0$. Sparing the detail of derivation explicitly laid out in Eeckhoudt et al. [2011], risk aversion may increase the optimal amount invested in prevention if and only if the probability of loss optimal for the risk-neutral individual p^n is below a critical threshold $\hat{p}$ defined to be equivalent to:

$$\hat{p} \equiv \left(\frac{1}{L}[U(w-g-U((w-g)-)]-U'(w-g)\right)[U'((w-g)-L)-U'(w-g)] \tag{1.13}$$

Put in simple terms, risk aversion may not necessarily increase the optimal investment in prevention because any form of risk aversion increases preventative measures only if more prevention obtains a second-order dominant shift in the final wealth distribution. This, however, never occurs as more prevention also leads to a reduction in wealth in the worst-case scenario when damage L is incurred. Thereby, prevention reduces wealth in both the good and bad states of the world, where prevention heightens the likelihood of the better of the two states. At a sufficient degree of risk-aversion, lowering wealth in the worst state may be construed as extremely painful in respect of utility loss. Namely, an infinitively risk-averse individual maximising the minimum final wealth withholds from investing in preventative measures altogether.

It is noted, however, that the critical threshold $\hat{p}$ depends on a specific utility function. When risk attitudes are defined by a quadratic utility function and the degree of prudence is zero ($U''' = 0$), the critical threshold $\hat{p}$ of 0.5 is measured at the maximum of risk variance. It then follows that when the risk neutral $p^n < \hat{p} = 0.5$, increasing loss prevention diminishes both p and σ^2, which is desirable under risk-averse quadratic preferences and thus reinforcing investment in preventative measures. On the other hand, when $p^n > \hat{p}$, while p falls, σ^2 rises and thus, lower spending on loss prevention is induced. In the limit, when the risk-neutral individual chooses $p^n = 0.5$, the impact on the variance is nil for small changes in risk prevention and all quadratic individuals select $p^* = 0.5$. Furthering this line of arguments, although it may too seem that prudent individuals should be making larger investments in risk prevention, conversely, prudence leads to increased marginal value of wealth and thus reduced willingness to expend wealth on prevention. That is, a prudent individual ($U''' > 0$) values

precautionary savings more than an imprudent person ($U''' < 0$), and hence, prefers saving more as a form of protection against loss at the expense of lower investment in preventive measures.

1.3.3 Information

Having inspected a spectrum of continuously growing academic literature on decision theory set to address the theoretical and empirical shortcomings of the standard EUT, most of the alternative accounts of choice appear to pivot on the pursuit of an appropriate formal conceptualisation of risk and uncertainty whereby to capture subjective beliefs about the probabilities of events. However, many real-life decision situations are dynamic and change over time in response to a variety of factors within but also outside the control of the decision-maker, who in most instances is concurrently subjected to inflows of new information, namely, knowledge, evidence, and experience on which to base or amend the decision-making process and thereby make better-informed choices allowing to increase expected utility. Therefore, there may be significant economic value embedded in information; however, by typically being contingent not only on its features and nature alone but also the specificity of the circumstances, context, and the parties in the known, information may have private, commercial as well as public implications [Acquisti et al., 2016].

Although economists have long recognised the importance of information as a valuable resource, it has only gained a prominent role in economic analysis in recent decades. More specifically, economic literature on information was predominantly motivated by the notions of Hayek [1945] wherein highly unlikely to

be fully acquirable for utilisation in central planning, knowledge and information are not only dispersed across separate individuals, but also ineffably local and unique to the circumstances of the fleeting moment, and hence more efficiently processable by a price system, coordinating the plans and actions of different individuals, and in turn, allowing for the most effective allocation of resources. Thereby inspired, a series of deliberations culminated in the seminal works such as those on the informative role of prices in market economies [Stigler, 1961]; the generation of knowledge and incentives to innovate [Arrow, 1962]; the ubiquity of asymmetric information and the consequent adverse selection [Akerlof, 1978]; the communication and acquisition of private information through the respective signalling [Spence, 1973] and screening [Rothschild and Stiglitz, 1978; Stiglitz, 1975] activities; and the arrangement of incentivising contracts to mitigate the moral hazard of asymmetric information [Holmström, 1979; Stiglitz, 1983] which amalgamated into a separate branch of study as it laid the foundations for what has become referred to as information economics [Acquisti et al., 2016].

Coinciding with the transition of modern economies toward extraordinary and ever-increasing capabilities of informational technologies, especially, the advent of the internet, enlarging the vastness of individual information available to be collected, stored, analysed, and repurposed for new uses, a rapid growth of the area of information economics followed, soon spanning the concepts of privacy sharing and privacy protection. Effectively, throughout the intense digitalisation of the economy, individuals are no longer simply the consumers of information, but also producers of frequently highly personal data, revealing the interests, actions and intentions at a breadth and detail which may bear substantial economic value deployable as a profitable business asset in targeting and else how influenc-

ing unsuspecting individuals by infiltrating their decision-making process in an inconspicuous but often interfering manner which, may produce a diverse set of effects of high, little, no or even negative relevance to the decision maker simultaneously exposed to the detriments associated with data security. As a result, complex and inherently intertemporal trade-offs between potentially dual and often ambiguous benefits and costs ingrained in privacy disclosure and protection emerge, which in addition to tangible elements, may also entail intangible aspects. Thereby, the multifaceted dimensions of vast inflows of digital information convolutes the extraction of value associated with information by complicating determining its relevance for management of risks in decision-making.

To put into perspective, risk has a nature of being sensitive to the arrival of new information which may affect the perception of riskiness associated with an act. Particularly, in conjunction with the Bayesian updating of risk, information is also useful as it allows for Bayesian updating of probability distributions, by which better decisions can be made than in the absence of information. For instance, in an online environment, before making a purchase, reviews about the seller may be consulted to avoid or minimise the potential disappointment of it being a fraud. Using online banking, the bank checks the account number against the name of the payee entered providing information whether they are sending money to the right person. Effectively, receiving an informative signal before the final decision is made may affect the response to the risks and, consequently, welfare. Therefore, information may have embedded value because it may enable better management of risk, which, more specifically, is analogous to introducing a mean-preserving spread in the probability of success. In the case of abundant digital collection of private data whose extent and usage are often uncertain, however, there may

be complex trade-offs curtailing the degree to which information may be utilised to optimally manage risk. To inspect the value of information in the context of economic analysis more generally and as it particularly relates the theory of choice, the framework of standard expected utility provides a starting building block whereon relevant modifications may be adapted to fit the context of online harms.

1.3.3.1 Value of Information

To gain preliminary insight, suppose a situation in which although subjective probability of success associated with an action in the absence of any information is believed to be p^{i0}, access to information may also be gained, which may not only allow to revise the course of action but also to investigate the necessity and eligibility of potential hedging strategies against a possible fallout. For instance, a government may intervene through policies that either reduce the probability of exposure to misleading material or provide an informational signal, such as a credible assessment of source reliability or an officially generated risk indicator, which functions as an additional piece of evidence that individuals incorporate into their posterior beliefs, thereby increasing the expected utility associated with the ensuing course of action.

Before the information is obtained, suppose there is an ex-ante expectation of receiving either a good signal or a bad signal about success characterised by the respective probabilities q and $1 - q$. By using Bayes's rule, the posterior probability of success may be computed as either $p^g > p^{i0}$ if the good signal is received, or $p^b < p^{i0}$ if the bad signal is received, whereas $p^{i0} = qp^g + (1 - q)p^b$ reflecting that the subjective probability of success remains the same before

the signal is observed. The Utility of undertaking an action in the absence of information may be calculated as $U^{i0} = p^{i0}U + (1 - p^{i0})U$. An individual may also have an option to evaluate pursuing a hedging strategy insuring against misfortunate outcome. In this endeavour, given the absence of information, the EU obtainable with the option to hedge the bad outcome of an action may be compared with the EU associated with this action without hedging in order to derive the benchmark probability of success p_c, below which, the EU is lower than that incorporating a hedging strategy and therefore, choosing to hedge is always optimal.

It thereby becomes of interest to determine whether expecting information makes the decision maker better off ex-ante, however, to solve which, applying backward induction is required. Effectively, the decision problem is first processed for each possible signal during which the expected utilities U^g and U^b for the corresponding p^g and p^b linked to the good and bad signals, respectively, are estimated. In this pursuit, the degree of confidence of success expressed in the form of probability transmitted by the expected news signals is also considered against the alternative courses of action, and the expected utility of the option yielding the optimum result is chosen. For instance, instead of taking insurance or using antivirus software or increasing privacy settings, if the level of confidence embedded in the expected information that a successful outcome will be acquired is high enough, an individual may maximise by allowing nature to run its course. However, given low confidence of success expected to be carried by a bad signal, the best alternative option may be chosen. Using these contingent values, before the informative signals of either success or misfortune, expected with probabilities q and $(1 - q)$, respectively, are observed, the unconditional expected utility U^i is

estimated by $U^i = qU^g(p^g) + (1-q)U^b(p^b)$. Finally, serving as a benchmark for decision making, the expected utility U^{i0} with p^{i0} in the absence of information is also assessed. If the result of $U^i > U^{i0}$ obtains, the expected utility may be seen to increase by changing the decision to that favoured by the information. Subsequently, the monetary value associated with information may also be derived as the difference between the decisions in the presence and absence of information.

$$U(CE^i) = U(CE^{i0} + K) \tag{1.14}$$

In equation 1.14, the left-hand side $U(CE^i)$ represents the utility associated with receiving a certain payoff that yields the same satisfaction as acting with additional information. The right-hand side $CE^{i0} + K$ expresses the utility of the certain payoff corresponding to acting without that information, adjusted by the monetary amount K, which compensates the user for the information gap. When both sides are equal, the user is indifferent between having and not having the new information, and K quantifies the monetary value of information.

Here, K is the monetary amount which an individual is willing to either pay to obtain information or, equivalently, be compensated with for the lack of it, and CE^i and CE^{i0} are the certainty equivalents associated with the expected utility in the presence and absence of information, respectively.

In the present thesis, CE^i is interpreted as the certainty equivalent of an individual's online decision after investing in verification or other protective actions, such as fact-checking, privacy adjustments, or security tools, whereas CE^{i0} denotes the certainty equivalent associated with acting on baseline, unverified information. Their difference therefore captures, in monetary terms, the value of

information generated by effort that reduces the probability or severity of online harms.

Notably, however, the information is valuable due to the ex-post decision being sensitive to the signal. In the case of insurance, threshold p_c may be established relative to the option to fully insure risk associated with undertaking an action, and against which the expected good and bad informational signals may be evaluated in the decision making. Indicatively, when either $p^{i0} < p_c$, or both $p^g < p_c$ and $p^b < p_c$, taking the insurance is preferred as information does not generate additional expected utility.

In the context of this thesis, the same logic applies when individuals decide whether to incur verification or acquisition costs before acting on online content. For example, paying for access to a trusted news source, installing security software, or devoting time to cross-checking the credibility of a post. The comparison between expected utility with and without such information acquisition provides the individual-level cost–benefit benchmark for assessing whether additional effort to reduce exposure to online harms is worthwhile.

1.3.3.2 General Model of Information

A general insight into the preliminary analysis is that the value of information is nonnegative and independent of the specific decision problem or the structure of information. To formalise the notion of information value, in any decision problem where the final utility $U(S, \alpha)$ is a function of a decision variable α and the state of the world s assuming S number of possible states of nature. In this context, α represents the information acquisition effort, that is, the extent of cognitive or temporal resources devoted to verifying or collecting additional information

before making a decision. CE^i and CE^{i0} denote the certainty equivalents of the user with and without the newly acquired information respectively. The difference between them measures the value of information, expressed as the expected increase in utility resulting from improved decision quality once the new information is incorporated.

The uncertainty then can be described by a vector of $P = (p_1, ..., p_s, ..., p_S)$, where $\sum_s p_s = 1$. Indirect utility function $V(P)$ can be defined as:

$$V(P) = \max_{\alpha} \sum_{s=1}^{S} p_s U(s, \alpha) \tag{1.15}$$

The decision problem without information, given the distribution P^{i0} of the states of nature, can be described as $V_{i0} \equiv V(P^{i0})$ determining the maximum EU. Assuming that decision maker is able to observe a signal before making decision on α, there may be M number of possible signals $m = 1, .., M$, where the probability of receiving signal m is denoted by q^m, with $\sum_m q^m = 1$. The posterior probability distribution of states of the world under the reception of signal m is denoted $P^m = (p_1^m, ...p_s^m, ...p_S^m)$. Since the signals are not yet observed, as before, the unconditional probability of state s is $\sum_m q^m p_s^m$ equals to that of P_s^{i0} in state s under no information, implying

$$P^{i0} = \sum_{m=1}^{M} q^m P^m, \tag{1.16}$$

suggesting that the underlying risk under the two circumstance is the same. It follows that the expected indirect utility regarding the choice of maximising action α before the informational signal is observed can be written as

$$V^i = \sum_{m=1}^{M} q^m \max_{\alpha} \sum_{s=1}^{S} p_s^m U(s,\alpha) = \sum_{m=1}^{M} q^m V(P^m), \tag{1.17}$$

whereas the value of information in the EU framework will be nonnegative when

$$\sum_{m=1}^{M} q^m V(P^m) \geqslant V \sum_{m=1}^{M} q^m (P^m), \tag{1.18}$$

although holding true if and only if the function 1.15 is convex in P.

The implications of this inequality are that an informed decision maker may always do at least as well as that who is uninformed by choosing to ignore the information. This also encapsulates the intuition behind the argument that incorporating information into the choice process may allow to adapt the decision to the given circumstances in a more efficient manner and thereby improve the management of risk.

In this respect, it might appear intuitive that the value of information for risk averse individual may be higher, however, it is generally not true. By comparing expected utility of the decision in the presence relative to that in the absence of information in terms of their respective certainty equivalents, the monetary value K, which may be construed as a compensating premium can be derived as:

$$V(CE^i) = V(CE^{i0} + K) \implies K = CE^i - CE^{i0} \tag{1.19}$$

Put differently, K induces indifference between being informed or acquiring the compensating premium. Given that the degree of risk aversion in EUT is determined by the concavity of utility function U, in anticipation that a more risk averse individual may require or, equally, be willing to pay a higher premium

to compensate or reduce risk, the impact on K relative to changes in concavity of U can be inspected more closely. In a concise analysis, Eeckhoudt and Godfroid [2000] provide a numerical example which dispels the intuition of a monotonic relationship between risk aversion and value of information which is shown to not necessarily obtain.

What is more, not only the degrees of risk aversion, but more generally, risk preferences appear to have no role in the inequality result of 1.18 since information always maintains a nonnegative value, whether an individual is risk-loving, risk-neutral or risk-averse [Eeckhoudt et al., 2011]. In so far as it concerns widening the spread between the posterior probabilities p^g and p^b away from the mean p^{i0}, and, overall, any mean-preserving spread within the range of posterior probabilities of the information structure, given the convexity of $V(P)$, the value of information and thus the expected indirect utility will increase. On the other hand, the inequality result is contingent on the linearity of EU in respect of probabilities which, however, constitutes the basis for the intense criticism and widely documented violations of EUT with subsequent emergence of accommodative non-linear EUT variants.

1.3.3.3 Comparative Statics

Having concentrated on welfare, the discussions is directed toward the implications of information for behaviour, particularly in light of optimal actions which are decided ex ante, before the informational signals are observed. For this, the analysis moves from a single period to a two-period decision model of the form:

$$\max_{\alpha_0} U(\alpha) \sum_{m=1}^{M} q^m \max_{\alpha \in B(\alpha_0)} \sum_{s=1}^{S} p_s^m U(s, \alpha, \alpha_0) \tag{1.20}$$

During the initial period $t = 0$, an individual selects α_0, which yields utility $U(\alpha_0)$. At the start of the next period $t = 1$, a signal m is observed influencing the beliefs of the individual about the distribution of the states of nature s. Conditional on having received signal m, the individual chooses an EU maximising α. The dynamic of this model originates from the choice α in the initial period which then may impact the subsequent period in two ways. More specifically, the choice set for α in the upcoming period may be limited by the original choice of α_0 accordingly set forth in the model by the restriction $\alpha \in B(\alpha_0)$ as well as the dependence of U on α_0, whose initial choice at $t = 0$ may directly affect utility U in the second period. What is more, as the individual is only aware about the probability distribution $(q^1, ..., q^M)$ over the set $\{P_m\}$ but not which of the M probability distributions P_m is the true one in the initial period, this engenders probabilistic uncertainty referred to as parameter risk which denotes uncertainty about the parameters of distribution regarding the risk for which an initial decision has to be made prior to this uncertainty being resolved in the next period.

Many real-life decisions involve probabilistic uncertainty including online situations such as making online purchases or simply conducting browser searches whereby receiving the purchase or visiting a website, respectively, is required to resolve the uncertainty. On the other hand, since some uncertainty may have the property to evolve as well as resolve over time, the timing of the decision may be crucial. It follows that the value ingrained in information also alludes to the

costs that may be incurred in the form of lost opportunities when decisions are delayed under the present uncertainty until additional information is received. However, by drawing a comparison between the optimal choice of α_0 when the resolution of uncertainty is expected and when no early resolution of uncertainty is available, the impact of information may be evaluated and thus an appropriate choice made.

1.3.3.4 Real option value and irreversibility

Over time, some of the decisions may not simply lead to bad outcomes but also result in further unexpected costs and value losses due to being irreversible in nature once made. Essentially, as new information becomes available on a continual basis, the decision maker might come to regret the initial decision in the instance of irreversibility. Therefore, when beliefs are expected to evolve over time, preserving some flexibility in decision making may be desirable.

While irreversibility of choice may affect any specific individual, it may be a particularly significant concern for firms. Hence, the focus is currently cast to the decision making of firms as it particularly relates to investment. For example, online platforms may currently choose to invest in a new technology which, however, may in the next period appear to lead to online harms, resulting in potentially costly regulatory or reputational repercussions and a subsequent loss of users. Alternatively, online platforms may invest in reducing online harms to retain and increase their user base, but which instead may be found to be disregarded by the users upon implementation.

To gain perspective into the problematic nature of irreversibility, a scenario whereby a risk-neutral firm must decide whether and when to invest in a risky

project with the selection of α_0 which modifies the opportunity set $B(\alpha_0)$ in the future without a direct effect on future utility ($U(\alpha_0) \equiv 0$), implying the inability to reverse, is analysed. In this scenario, the investment is expected to generate a net cash flow of x_0 in the initial period 0 and x_1 in the next period 1, where $\mathbb{E}(x_1) > 0$ and the investment is irreversible in that if the investment is made in period 0 ($\alpha_0 = 1$), the firm is unable to divest in period 1. The firm is also assumed to discount future cash flows at rate r.

Supposing that α_t denotes the production capacity in period$t$, where $t = 0; 1$, then irreversibility implies that $B(\alpha_0 = 0) = \{0; 1\}$, whereas $B(\alpha_0 = 1) = \{1\}$, suggesting that refraining from investing in the project at $t = 0$ ($\alpha_0 = 0$) provides for more flexibility in the future. In this sense, not investing as opposed to investing is a reversible action.

To analyse the value of the decisions, two scenarios when no information about the distribution of x_1 before the end of period 1 is expected and when there is a complete early resolution of uncertainty by the end of the period 0 at which x_1 is revealed with certainty are compared. In the absence of any early resolution of uncertainty, it is optimal to invest immediately if and only if x_0 is positive. If the firm also ponders delaying, it should then compare the Net Present Value (NPV)$== x_0 + (1 + r)^{-1}\mathbb{E}x_1$ of the immediate investment versus the decision to delay investment which, in this instance, results in a lower NVP at $(1 + r)^{-1}\mathbb{E}x_1$ by the amount of x_0.

When an early resolution is expected, however, it is only optimal to invest immediately if:

$$x_0 + \frac{\mathbb{E}x_1}{1 + r} \geqslant \frac{\mathbb{E}\max(0, x_1)}{1 + r} \tag{1.21}$$

On the right-hand side of inequality is the NPV denoting the decision to delay and invest in period 1 only if it is optimal and only if $x_1 > 0$. It follows that it may be also optimal to refrain from investing in the project altogether if $x_1 < 0$, allowing to evade loss in period 1. The NVPs of the two decisions may be rearranged as

$$x_0 \geqslant \frac{\mathbb{E}\max(-x_1, 0)}{1+r} \tag{1.22}$$

The right-hand side measures the benefit of waiting, enabling to avoid the loss $-x_1$ when x_1 is negative. The cost associated with the delay, however, is the opportunity cost of losing x_0. One of the general implications of the analysis which holds for all decision problems as in model 1.20, where $U(\alpha_0) \equiv 0$, is that the minimum value of x_0 leading firms to make an immediate investment will be higher when the uncertainty evolves over time, suggesting that accounting for the resolution of uncertainty induces the decision maker to value flexibility in the future. For example, social media platforms and apps more generally may delay launching new technological features and updates, often by developing and releasing early stage demo versions only available to some users to gather feedback as well as to decide whether the development of the full version is worth pursuing.

To summarise, firms that must decide at $t = 0$ whether to invest may usually decide to invest if and only if the NVP $= x_0 + \frac{\mathbb{E}x_1}{1+r} > 0$. In case delaying the decision is a viable option and some information is expected in the future, the correct cost-benefit analysis is to employ the criterion $x_0 \geqslant \frac{\mathbb{E}\max(-x_1,)}{1+r}$, which accounts for the premium $\frac{\mathbb{E}\max(0,x_1)}{1+r}$ in literature referred to as the real option value inherent in the ability to delay the decision.

1.3.3.5 Savings and the Early Resolution of Uncertainty

Irreversibility is not the only feature influencing the optimal early decision under the circumstances of evolving uncertainty. By way of illustration, assume a consumer who lives for three periods, $t = 0; 1; 2$, has an initial wealth of w_0 and at the final period of consumption at $t = 2$ earns an uncertain income $\tilde{x}$. When no early resolution of uncertainty is available and realisation of $\tilde{x}$ is observed only at the beginning of the period $t = 2$, the decision problem can be written as:

$$\max_{\alpha_0} U(w_0 - \alpha_0) + \max_{\alpha}[U(\alpha_0 - \alpha) + \mathbb{E}(\alpha + \tilde{x})] \tag{1.23}$$

where α_0 and α are the decision variables embodying savings at the end of periods $t = 0$ and $t = 1$, respectively. As the individual is also assumed to be averse to consumption fluctuations ($U'' < 0$), smoothening consumption over the first two periods $t = 0, 1$ by $w_0 - a_0 = a_0 - a$ is optimal which implies $\alpha = 2\alpha_0 - w_0$, following which, the decision problem can be expressed by a function $H(\alpha_0)$ as

$$\max_{\alpha_0} H(\alpha_0) = 2U(w_0 - \alpha_0) + \mathbb{E}[U(2\alpha_0 - w_0 + \tilde{x})] \tag{1.24}$$

The objective is to determine the impact the informational signals have on the optimal savings decision which is made prior to acquiring information. If the uncertainty is fully resolved at the end of the period $t = 0$, it is optimal to perfectly smoothen consumption over the remaining periods $t = 1, 2$:

$$\max_{\alpha_0} U(w_0 - \alpha_0) + 2\mathbb{E}\left[U\left(\frac{\alpha_0 + \tilde{x}}{2}\right)\right], \tag{1.25}$$

then taking its first-order condition (FOC)

$$U'(w_0 - \alpha_0^i) = \mathbb{E}\left[U'\left(\frac{\alpha_0^i + \tilde{x}}{2}\right)\right], \tag{1.26}$$

the optimal saving of the ex-ante informed agent α_0^i can be derived. In comparison, given that the function H is strictly concave in α_0, the optimal level of savings under late resolution of uncertainty is higher than the early resolution result α_0^i if and only if $H'(\alpha_0^i)$ is positive, underlying the following condition:

$$\mathbb{E}[U'(2\alpha_0^i - w_0 + \tilde{x})] \geqslant U'(w_0 - \alpha_0^i) \tag{1.27}$$

Considering prudent individuals whom Kimball [1989] shows to have a precautionary saving motive, the reduction in risk implicit in the complete early resolution of income uncertainty at the end of period $t = 0$ incentivises all prudent individuals to reduce the level of their precautionary savings, whilst the opposite is true for individuals who are imprudent. Effectively, the intuition behind earlier resolution of uncertainty prompting prudent individuals to lower their precautionary savings may be explained in terms of sooner information providing time to diversify and thereby reduce more of the future risk.

1.3.3.6 The Hirshleifer Effect

It has been stipulated that information has nonnegative value for the decision maker, however, this appears to hold only if this information has no bearing on the other elements within the decision-making environment, and, in that respect, is private. This, however, may not be the case when information is public. Suppose all risk averse agents face a risk of loss L. If a new technology is introduced that allows both parties to obtain information on a risk of loss at zero cost, exposing

who will suffer damage, together with the size of the damage, then there is nothing to insure anymore as risk is realised.

Supposing a scenario where all risk-averse agents face idiosyncratic risk of a loss L and there is a competitive insurance market with no transaction costs and no asymmetric information, then, in equilibrium all individuals are fully insured at actuarially fair premium $\mathbb{E}(L)$. However, if a new technology is introduced allowing at zero costs for all parties to obtain perfect information on the damages specifying who will suffer the damages and their magnitude, there is no longer anything to insure because under perfect information the risk becomes realised. Viewed ex ante, the value of this information is negative as everyone is made worse off at the eliminated possibility to insure at fair price. It follows that the cost of this information equals the risk premium associated with losses L, since individual now bear risk L rather than its mean $\mathbb{E}(L)$ reflecting the so-called Hirshleifer effect.

In contrast, when the probability that half of a population sustains loss L, in the absence of information, insurance can be bought against being an unfavourable loss type. This opportunity disappears when the individuals who will experience loss become known. In view of this, Hirshleifer [1978] argues against information being released too early which not only may prevent individuals to trade but also have adverse effects on risk sharing and lead to reduction in welfare. To mitigate the potential costly fallout from information in the sense of Hirshleifer, long-term contracts could insure against bad news, however with potential costs of parties becoming locked into a contract upon arrival of good news as well as being subjected to adverse selection problem if contract is signed under asymmetric information. Additionally, new technology accommodating information

sharing could be banned, or alternatively, access to private information may be prohibited, however, which then may also be exploited strategically creating the adverse selection problem of asymmetric information. Essentially, all long-term contracts wherein individual risks evolve over time in a Markovian way may be analysed in terms of the Hirshleifer argument.

In the online context, concerns for privacy are mounting against the incessant and ever growing data collection. Whilst individuals actively engage in willing sharing of personal data online through participation in forums, social media and other forms of media, it is usually confined to the space of individual privacy preferences and settings. In parallel, however, online users are often fully unaware about the degree and subsequent consequences of personal data collected in the background of their online activities. In particular, as the current and constantly evolving technologies allow tracking online behaviour and collecting information at the granularity of past purchases, browsing activities, search queries as well as the subsequent clicks made, not only insight into user interests and preferences may be gained, but also sold, traded and shared among different parties. Under these circumstances, the Hirshleifer effect may very well materialise in that the user may be first deprived of the ability to act on their personal data when the data is collected and becomes known by other parties. In turn, the commercial value inherent in personal data may be deployed against the user who may be subjected to a variety of personally costly practices, including price discrimination in retail markets, quantity discrimination in insurance and credit markets, spam, or risk of identity theft. Thereby, a reduction in private utility and, in aggregate, a decrease in social welfare may be imminent.

Viewed together, the analyses of real option value, early resolution of uncer-

tainty, and the Hirshleifer effect extend the cost–benefit logic of decision-making under uncertainty to settings where information acquisition itself has strategic value. In the context of online behaviour, they describe how users weigh the benefits of waiting for clearer signals, the preference for early certainty about information accuracy, and the potential disutility of overexposure to information. These insights collectively inform the modelling of verification effort developed later in the thesis, where user strategies for mitigating online harms are examined as optimisation problems under uncertainty.

1.3.4 Asymmetric Information

Within the online ecosystem, the principal-agent relationship can be interpreted through interactions between platforms (principals) and users (agents), or alternatively between content producers and consumers. Platforms partially delegate the moderation and verification of information to users, who act under asymmetric information about the reliability or potential harm of content. From a cost-benefit perspective, this asymmetry introduces moral hazard wherein users may underinvest in verification effort when its benefits are uncertain or externalised, thereby generating collective harms. Framing these interactions as principal-agent problems helps to align the theoretical treatment of incentives with the behavioural realities of online decision-making, even as platforms retain significant liability and control over algorithmic curation.

Utilizing comparative statics, the investigation into the implications of information for decision-making concerning welfare and behaviour itself as a form of risk management revealed information to not only be non-negative in value ir-

respective of the particular decision problem or information structure but also, under evolving uncertainty, enhancing the value of flexibility allowing to reverse choice in the future. However, it is noteworthy that the analysis maintained the assumption of complete information whereby knowledge about each individual is perfectly symmetrical and available to everyone. Under many circumstances, information on which decisions are based regarding any specific matter at any given time tends to differ across people. What is more, decision making process often transcends the sole means and perspective of a single individual. It may entail a complex multifaceted network of considerations drawn on resources, solutions and delivery mechanisms tendered by a variety of other individuals whose precise obligations are typically determined by a contractual agreement.

From the perspective of complete information, all parties have the same salient information and are accordingly able to plan and make their decision strategies when entering a transaction in an informed and confident manner, maximising expected utility. Following the proneness of many situations to confine individuals to private as well as varying in amount and quality information, the consequent disparity in the degree of informativeness may instil the pursuit of opportunistic objectives among parties involved. Effectively, information asymmetry may generate imbalances and abuse of power in transactions to ensure favourable outcome, but at a probable expense of allocative inefficiencies and, more rarely, market failures. Seemingly strategically beneficial, in expectation of opportunistic behaviour, individuals may be disposed to discount the potential information asymmetries by detracting value from the associated transactions and, thereby, inflict costs which may be further compounded by the difficulty of reducing informational discrepancies.

Although these notions have been implicit throughout the existence of markets, in a foundational contribution transforming economic thinking about the functioning of markets, asymmetric information was formally conceptualised as a fundamental factor in the theory of markets developed by Akerlof [1978]; Spence [1973] and Rothschild and Stiglitz [1978]. More specifically, Akerlof [1978] was the first to formally demonstrate that informational asymmetries, especially in the form of hidden characteristics before the transaction is settled, may engender the problem of adverse selection in the markets. To put into perspective, adverse selection refers to a situation wherein either sellers or buyers have private information about some aspect of the transaction being entered, as exemplified by Akerlof [1978] in the context of a market for used cars. In this market, the sellers know the quality of the car better than the buyer, which although may be of good quality, they may also be defective and colloquialised as 'lemons', a now well-known metaphor for economic problems regarding the value of a transaction under asymmetric information. Since buyers are unable to easily distinguish lemons from high-quality cars until after purchase, the suspicion of the car being defective results in sellers having to either accept a lower price than they would in markets where quality is transparent, or alternatively, withhold the car altogether if the market price is lower than its reservation price. In this fashion, Akerlof [1978] illustrates how the presence of asymmetric information may elicit adverse effects such as additional expenses, a decline in available quality for all individuals or even a market collapse when individuals on one side of a market are only aware of the distribution of transaction quality, rather than the quality of each individual transaction [Riley, 2001].

Another classic example of adverse selection first proposed byRothschild and

Stiglitz [1978] relates to the insurance industry with asymmetric information where insurance companies base their menu of prices on the risk types inferred from certain observable features of the applicant, whilst individuals are privy to private knowledge of being more prone to make a claim against a highly-likely loss and may self-select into buying insurance at the expense of insurers having to sustain the losses. In other words, each individual knows their own loss probability, but the insurers cannot observe individual and often hidden risk characteristics, being merely acquainted with the distribution of inferable risk classifications across the population.

As an option, in refusal to bear the losses, the insurance companies may raise the coverage price to everyone, however, potentially resulting in the eligible and desirable applicants being priced out of the market. For instance, in the context of cybersecurity insurance, following a major data breach in the industry, insurance companies may increase premiums across all clients to account for the heightened risk of cyberattacks. This could lead to small businesses being priced out of the market, as they may no longer be able to afford the higher premiums or the costs may be shifted onto customers. As another option, insurers may manufacture contracts at different prices on the basis of known risk distribution in a manner inducing applicants to reveal their risk characteristics. In the latter scenario, the problem of adverse selection may be mitigated but not entirely avoided, as although high risk types may experience no losses in welfare, individuals with low-risk characteristics may have to accept the additional cost of signalling their risk type. Viewed in this perspective, the car and insurance markets serve an illustration how under many different specific sets of circumstances, adverse selection may instigate costs which may be balanced by the benefits to achieve the

expected utility maximising outcome. A pertinent illustration of this market imbalance phenomenon outside traditional insurance markets may be observed in the tiered pricing strategies widely implemented by Software as a Service (SaaS) providers. For instance, Salesforce offers a range of Sales Cloud pricing editions, where varying access to features, data storage limits, and support levels aims to segment users based on their anticipated usage intensity and technical proficiency, thereby inducing self-selection and mitigating the risk of uniform pricing that would otherwise disproportionately burden low-demand, low-cost users

On account of these arguments, the far-reaching nature of implications associated with adverse selection appear to constitute a significant consideration for the cost-benefit analysis when it is performed in the presence of asymmetric information to inform the decision-making process in a manner which prevents allocative distortions leading to suboptimal expected utility. A pertinent illustration of this dynamic in the context of online harms is the challenge faced by social media platforms in mitigating the proliferation of misinformation. The underlying information asymmetry arises from the fact that content creators, particularly those disseminating false or misleading information, possess superior knowledge regarding the veracity of the content compared to the platform itself. This asymmetry exacerbates adverse selection, as platform algorithms, often optimised for engagement, may inadvertently amplify harmful content. This is due to the fact that such content typically garners more attention such as clicks and shares than reliable information. As a result, users who prioritise accuracy and factual content may disengage from the platform, recognising its diminishing informational quality. The long-term societal costs, ranging from public health deterioration to the erosion of trust in institutions and increasing political polarisation, are often

inadequately internalised in the platform's operational decisions. This creates a market failure wherein the platform fails to bear the full social cost of its decisions

That said, once the transaction is placed, the implications of asymmetric information may subsequently take the shape of hidden or unobservable actions which may give rise to moral hazard embodying a disincentive to guard against the risks associated with the transaction either induced by the lack of information about its implementation or inability to be held accountable upon discovery of risky behaviour and failure to act in good faith since it is the other party that bears the consequent economic impact. In other words, whilst adverse selection appears to occur 'ex-ante' as a transaction is being entered, moral hazard occurs 'ex-post' when the transaction is being executed. Considering a related context of used cars, once paid for, the seller no longer has an economic incentive to ensure careful delivery of the car to the buyer. In respect of insurance market, as explored by Stiglitz [1983], the insured party not only becomes less inclined to exert effort to prevent a state of loss, but knowing that the insurer is the party incurring the costs further incentivises to increase exposure to risk.

Alternatively, moral hazard may manifest in the arrangement between the principal and the agent as a problem wherein a conflict of interests emerges when an agent is legally and contractually authorised to act on the behalf of principal, but the actions of the agent are hidden or difficult to observe [Ross, 1973]. In particular, interests between the two parties may clash or be misaligned since the actions favoured by the principal may be costly for the agent who, in turn, may exploit being in the advantageous position of information asymmetry and act in accord with own best personal interests, not only contrary to those of the principal, but imposing negative externalities on the principal, or at the very least

shirking responsibilities. This may be linked back to the car example if instead the seller delivers the car to the buyer by hiring another individual, who however, has no incentive to exert careful behaviour and drive safe.

If it were possible to observe actions of the agent, especially in relation to productivity and respective effort made, directly, a clause may be inserted into principal-agent relationship defining contract, denying the agent a payment in the event of detrimental risky behaviour and lack of effort. Alternatively, the principal may attempt to influence the behaviour of its agent in the form of a contract instituting a compensation contingent on either the actions of the agent or their consequences. However, in many instances such as that of a hard-working lawyer losing the case due to bad luck, or a consulting company, despite little effort, attracting capital due to luck, the principal cannot compensate its agent based on the level of effort exerted as it constitutes information that is ex ante and often ex post unobservable and unverifiable to the principal but remains at the disposal of the agent, thereby allowing furthering personal agendas.

In essence, moral hazard arises when a party has limited responsibility for the risk and actions they take and, as a result, may occur in a variety of spheres. Thereby, moral hazard can present itself in online space where it may manifest in several different ways and consequences. Within the principal-agent framework, numerous problems may emerge between an online user as the principal providing the data in order to receive some form of benefits (e.g. personalised offers, advertisements) and data aggregators, advertising networks, and website operators as the agents controlling the collected data on the behalf of the principal to allow provision of such benefits in return for the revenue which may be generated on these data.

Having said that, these online agents may also have to be appropriately incentivised to not simply monetise on the data, returning inadequate output to the user, but to make effort and use data in a manner beneficial to the principal. Exerting effort in this context, however, may induce costs of lost revenues and required technological advancement in addition to those of inherently online and reputational nature such as software vulnerabilities, inadequate data protection and data-handling policies, breaches of security which have been shown to lead to market value losses [Acquisti et al., 2006; Cavusoglu et al., 2004], or costs due to underdeveloped technologies such as repetitive ads, ads of already purchased rather than undiscovered products.

To exemplify such an online principal-agent relationship, by using targeted advertising, the agent may benefit principals with information about products of interest, thereby reducing their search costs and, in turn, improving their welfare if the agent exerts effort and improves match quality of targeted advertising, as well as generally refrains from offering inferior or even potentially damaging products. When the online agent fails to put effort and only pursues private interests, detrimental impacts on user welfare may range from receiving spam, higher prices, being steered and manipulated into unnecessary products or else how marketed by data brokers, namely leading to targeting individuals suffering from addictions such as alcoholism or gambling.

In an endeavour to minimise the various costs associated with asymmetric information primarily taking the shape of unobserved characteristics or actions resulting in the corresponding problems of adverse selection and moral hazard, several measures have been proposed. As a solution for one of the adverse selection problems, Spence [1973] demonstrated that under informational asymme-

tries, the better-informed individuals entering into an agreement may be able to improve their market outcome by signalling their private information to the less informed individuals on the other side of the transaction. For instance, in the lemons market, the sellers might provide a credible signal such as a limited warranty which would be too costly for a poor-quality car seller, however, imposing additional cost on the seller of the good car. Alternatively, Rothschild and Stiglitz [1978] and Stiglitz [1975] posited that the informationally disadvantaged party may capture private and unobservable information by establishing a screening mechanism whereby a menu of different contracts based on certain observable characteristics may be manufactured to incentivise the respective different types of individuals to reveal their private knowledge by self-selecting into the contract offering them the most attractive terms. Accordingly, insurance companies may classify their clients into risk categories by extending different policies, which, for instance, may tender a lower premium in exchange for a higher deductible, repelling individuals with higher likelihood of loss. To prevent or mitigate moral hazard, the causal asymmetric information has also been suggested to be dealt via incentivising contracts designed on the basis of state contingent wealth to instate a better balance in the relationships between different parties whether wherein one party is at an advantage of not bearing the full costs and consequences of their actions [Stiglitz, 1983], or when a party is delegated to act on the behalf of another [Holmström, 1979]. Essentially, by incorporating a means addressing the asymmetries of information, better decisions, potentially improving the quality and the outcomes they deliver, may be reached.

Taking these insights upon consideration of online harms it follows that informational asymmetries may be construed to be integral components of most

interactions online by means of which continual engagement and participation is promoted, especially through intense collection of information often encroaching on privacy, allowing to encourage and sustain online presence. In this regard, however, as economies become ever more digitalised, the process of making informed decisions is not being solely distorted in respect of privacy, but, following the resultant and often unobstrusive collection of data via ever advancing technological solutions such as cookies, device fingerprinting, location tracking via GPS, cross-site tracking, real-time bidding systems, and AI-based behavioural profiling, users are being placed into a position of imperfect or asymmetric information wherein the timing, amount, specificity, purpose and the consequences of data harvested is unknown. This may result in the targeting and potential infringement on the ability to make independent choices, particularly in terms of willingness to pay, need for certain items, information content, and the emergence of bubbles and echo chambers. To illuminate the construction of models which provide an appropriate account of online harms caused by asymmetries in information, it is first important to gain perspective into the mechanics of the simple models, addressing the problems of adverse selection and moral hazard wherein optimal trade-off between costs and benefits under a given set of circumstances in the sense of EUT may be achieved.

Chapter 2

A Theory of Online Harm

2.1 Information Consumption on Social Media

Having profound implications for understanding economic and social phenomena, the economics of information challenged the prevalent paradigm of economic analysis [Stiglitz, 2000]. Although eighteenth and nineteenth century economists alluded to the problems of imperfect information, noting their ramifications and importance, no pursuit to attribute the logical implications or the source of the observed phenomena to the information imperfections were made. It was not before the second half of twentieth century that modern economists brought a revolution in economics, upsetting long-held assumptions that brought key changes spanning the entire field. Effectively, information, being a valuable resource, appears to exert its impact on pricing dynamics and resource allocation within market as well as mould decision-making [Allen, 1990]. More specifically, the economics of information has shaken the traditional assumptions of perfect information and rational decision-making in economics, paving the way for the

development of more realistic and nuanced models of economic behaviour and choice. Decision making under certainty may seem easy, but the real world is fraught with risks and uncertainties that can significantly impact the outcomes of any decisions individuals face. Among many others, these decisions span financial investments, business strategies, travel plans, and even online behaviour, where concerns about privacy, information security, and the pervasive threat of misinformation and disinformation have become a paramount issue in the current technological landscape, whereby opportunistic and strategic exploitation and manipulation is a real possibility.

Placing individuals into an unprecedented state of interconnectedness and instant access to a wealth of information, the Internet has become more than an integral part of modern lifestyle. Notably, the Internet has introduced and fuelled a vast range of ever evolving technologies which have fundamentally transformed and enhanced nearly every aspect of human experience, reshaping and introducing new digital alternatives and methods to the conventional channels of communication, work, education, entertainment, media and commerce, among others. Constituting a crucial leap forward, the Internet may be said to have democratised access to information and alleviated information inequality with anyone connected to the internet being empowered to access a vast wealth of knowledge, news, research, and educational resources. Complementarily, each connected individual is now provided a means to not only express approval or disapproval but also voice their opinions, ideas, and concerns and even generate and disseminate their own content, contributing to a diverse and dynamic ever expanding online ecosystem, and fostering a more inclusive and participatory digital society.

Effectively, the continual advances in online technologies have lowered the

costs and increased access to information production and dissemination, expanding the sources of information traditionally concentrated within a small group of gatekeeping media outlets, to a worldwide community of contributors and consumers, revolutionising the way knowledge is created and shared. On the other hand, as nearly anyone can now author information disseminated online, it has also raised concerns about the reliability and accuracy of the content available. Absent of universal publication standards, Information online which besides news articles may very well be posted and consumed on multiple online platforms and in many different forms such as blog entries, social media posts, videos, podcasts, and forums, and even mere comments, to name but a few [Hassoun et al., 2023; Kim et al., 2014], and which may also be readily modified, manipulated, or anonymously generated with deceptive intentions [Fritch and Cromwell, 2001, 2002; Johnson and Kaye, 2000; Metzger et al., 2003; Rieh, 2002].

Moreover, in recent years, technological advancements have further convoluted decision-making processes by introducing unprecedented levels of complexity. The rapid evolution of technology, big data, and interconnected digital ecosystems has not only expanded the volume of information available but has also increased the velocity at which this information is generated and disseminated. Consequently, decision-makers are faced with the challenge of sifting through vast amounts of data, dealing with cybersecurity threats, and adapting to the dynamic nature of the digital landscape. In this intricate environment, the peril of misinformation and disinformation adds an additional layer of complexity, demanding that decision-makers discern truth from falsehood while harnessing the power of information effectively to navigate the intricacies of the modern world. In particular, the expanding literature on online information consumption demonstrates the

ease of accessibility to misinformation [Allcott and Gentzkow, 2017; Del Vicario et al., 2016] and its rapid dissemination across digital social networks [Vosoughi et al., 2018]. While misinformation characterised as information that contradicts established facts [Ecker et al., 2021; Vraga and Bode, 2020] is usually disseminated inadvertently, disinformation is a subset of misinformation propagated intentionally with an aim to deceive [Starbird, 2019]. It is notable that despite establishing the direct causal impact of online misinformation being convoluted [Enders et al., 2022; Uscinski et al., 2022], a wealth of research has revealed that being exposed to misinformation is linked to the adoption of false beliefs [Bor and Petersen, 2022], endorsement of conspiracy theories [Xiao et al., 2021], and engagement in nonnormative behaviours like vaccine refusal [Romer and Jamieson, 2021].

In this manner, the democratisation of information has not just simply brought profound multifaceted benefits extending beyond the convenience and efficiency to empowerment and societal progress, but also given rise to a landscape ridden with issues concerning misinformation and disinformation. To further exacerbate it, the dissemination of false information online has been shown to be more rapid and widespread than the dissemination of true information [Vosoughi et al., 2018]. Therefore, whilst effectuating an indispensable and unfettered space for information exchange of varying forms, the internet, in its entirety, has transferred the responsibility of assessing credibility and ensuring quality from traditional gatekeepers to individual information consumers, necessitating critical evaluation and fact-checking to navigate the ever-expanding repository of digital knowledge. Given the incessantly increasing amounts of information and the varying degrees of relevance it bears to any individual compounded by various personal constraints of physical resources and intellectual capacities, many

internet users appear to resort to adopting practical rather than sophisticated decision-making processes which help them efficiently filter and extract valuable insights from the overwhelming data landscape [Briggle et al., 2008; Hilligoss and Rieh, 2008; Metzger et al., 2010; Sundar, 2008]. Captured in economic terms, users may be viewed as continually making choices how to allocate their limited resources against the potential outcomes. This allocation may involve trade-offs wherein users have to assess the costs and benefits associated with the different available courses of action to establish one with the most satisfying payoff. In the instance of content online, users may have to weight the cost of time against the significance of acquiring additional information, which, under many circumstances may be opting to consume an article of information without confirming its veracity as they scroll to yet make a decision on another item of information.

In essence, this prompts a fundamental inquiry into the intrinsic worth of information. On a daily basis, online users consume copious amounts of information, covering a wide spectrum of subjects. This information may range from profound scientific knowledge to fleeting gossip, which, to varying degrees, different individuals may regard as valuable in its own right, irrespective of any immediate practical applications. The associated intrinsic value lies in how this information contributes to the enrichment of human understanding, culture, and knowledge overall. In contrast, in modern economic research the primary emphasis revolves around the extrinsic value of information which materialises in the form of augmented decision-making. This valuation is rooted in the observation that information may enable individuals to make choices resulting in superior expected payoffs or utility in comparison to those made in the absence of such information.

Some prevalent exemplifications of extrinsic information value entail the reduction of uncertainty or improved risk management whereby risks may be identified, assessed and even mitigated, thereby allowing to make better choices. That is, by subscribing to a newspaper, an individual will be apprised of the most recent geopolitical developments, environmental forecasts, cultural trends, and scientific breakthroughs, which may lead to revisiting their decisions, ranging from investment portfolio adjustments and travel plans to lifestyle choices and daily routines. Offering a more specific decision problem example in the context online information, Pirolli [2005] applies the so called information foraging theory, which elucidates the observed user behaviour of gathering information for some purpose, such as informing a medical decision, selecting a restaurant, or purchasing real estate to the degree that it maximises the value the knowledge obtained from the web generates by improving ill-structured decision-making and problem solving relative to the cost of interaction, particularly the opportunity cost of the time invested in these online interactions. Premised on the conventional approach to information as bearing extrinsic value, the majority of academic research centres on individuals proactively seeking out information to complement their understanding of a predefined and existing decision problem, thereby making more informed choices.

Against the backdrop of wide online accessibility and the proliferation of digital technologies, deviating from the standard structure of decision problems, by and large social media emerges as one of the key online environments where extensive information sharing and consumption occurs without a predetermined immediate objective. Not only have several recent surveys revealed that more than half of the respondents regularly accessed news via social media and partic-

ularly via their Facebook feeds [Shearer and Mitchell, 2021], a third appeared to have initially believed fake news, an umbrella term for the various forms of misleading information which might include satirical news stories, large-scale hoaxes with an intent to deceive about a news story, news fabrications as well as deliberately sensationalised events, circulated on the social media platform to be true [Flintham et al., 2018]. In an experimental study, employing data on behavioural and neurophysiological responses of participants to displayed news headlines designed to be possibly true or false as well as either congruent or misaligned with their political views Moravec et al. [2019] examine how effective social media users with different political beliefs are at detecting false information as well as the changes in their cognition. In this effort, Moravec et al. [2019] found that online users were inept at distinguishing fake from true news with only 17% of the participants being better than chance, also manifesting strong confirmation bias wherein users appeared to believe in headlines that supported their priori opinions, disregarding the actual underlying truth or a fake news flag.

To encapsulate these notions succinctly, firstly, despite Internet users actively searching for information online, their pursuit of accuracy, as per the motivating principles of Chen and Chaiken [1999] applied to the online environment, may differ depending on the situation, resulting in varying levels of motivation to make accurate judgments across various contexts, leading to varied accuracy goals from search to search. Additionally, Internet information seeking may oscillate between casual and purposeful, contingent upon the context. Undoubtedly, certain online browsing activities may be motivated by the pursuit of accurate information, a significant portion of users' online information-seeking behaviour may lack a clear purpose. An individual may utilise the internet for leisurely en-

tertainment, embark on a search for a specific topic, but be subsequently steered to other content through hyperlinks, or come across unintended content while exploring their feed on social media. In the contexts characterised by a less well defined purpose of interaction with online information, it remains unjustified to presuppose that these users disregard the credibility of the information encountered online. Likewise, it is conceivable that these users may display less concern regarding credibility, bearing reduced willingness to dedicate their full cognitive resources to assess information online.

Moreover, in the conventional forms of media supplanting, nearly all encompassing social media, individuals engage with a multitude of content, often influenced by algorithms that customarily curate information for each user, drawing from the data collected on their behaviour online. This data encompasses past online searches, browsing history, content interactions, purchase history, likes, shares, comments, followed profiles, community engagements, demographics (e.g., location and age), and more to encapsulate the distinct preferences each user has for the content they interact with. By continuously assessing this data, as commercial entities in an attempt to maximise user satisfaction, social media platforms such as Facebook create tailored user profiles that inform content curation, delivering a highly personalised online experience. While providing benefits such as increased user engagement and satisfaction, by increasingly personalising content, the machine-learning models may be fostering filter bubbles, wherein algorithms automatically suggest content that is anticipated to resonate with the preferences and attitudes of a user [Hannak et al., 2013; Pariser, 2011], thereby inadvertently placing individuals in echo chambers, where they may be exposed primarily to information that aligns with their existing beliefs and preferences,

reinforcing confirmation and belief biases [Metzger and Flanagin, 2013], compounding the general tendency of online users to opt for content aligning with their attitudes and regard it more positively than information contradicting their views [Fischer et al., 2005; Garrett, 2009; Iyengar and Hahn, 2009]. Notably, cognitive biases such as the confirmation bias characterise the human tendency to selectively perceive and accord greater significance to information that aligns with existing beliefs, often resulting in the underestimation or neglect of contradictory information [Klayman and Ha, 1987]. The belief bias represents our inclination to support conclusions in harmony with our preconceived notions, irrespective of their logical validity [Evans et al., 1983]. These psychological phenomena among others may limit the diversity of perspectives, potentially isolating users from different opinions and deepening their preexisting biases.

Aside from delivering customised experience, in another significant divergence from the traditional media, the content disseminated on social media may be produced by a broad variety of different stakeholders, such as advertisers, the user community, and other individuals or organisations who may bear specific intentions or agendas, namely, celebrities, politicians, influencers or even news outlets. What also sets social media apart is that these various stakeholders not only have the means to make information and narratives accessible online, but, facilitated by technologies, they also possess the capability to actively influence, shape, and promote their agendas by appealing to specific individuals or groups, targeted and not, along with their peers and associates. In practice, the influence is exerted via generation of content that resonates with the intended audience, pandering to cognitive biases, and triggering sharing and dissemination among and beyond the initially targeted groups. To further their influence, stakeholders may employ

various strategies, such as deploying bots, internet trolls, or tampering with social media algorithms and systems, among other tactics.

For instance, in efforts to advance the understanding of the dynamics underlying the spread of falsehoods, extensive research has delved into the propagation of online political false content as a phenomenon primarily originating from political figures [Berlinski et al., 2023; Garrett, 2017; Lasser et al., 2022; Mosleh and Rand, 2022], untrustworthy online sources [Guess et al., 2020], and adversarial foreign governments [Bail et al., 2020] channelled via social media and other interconnected networks [Johnson et al., 2022]. Along with misinformation, a substantial proportion of online political material have been discovered to be generated by a relatively limited number of accounts [Grinberg et al., 2019; Hughes, 2019]. Besides political entities, many individuals are argued to purposefully disseminate incorrect information with the intention of misleading others in order to advance particular agendas [Buchanan and Benson, 2019; Littrell et al., 2021; MacKenzie and Bhatt, 2020; Metzger et al., 2021]. Frequently, individuals producing and circulating misinformation are motivated by the prospect of becoming viral and widely shared on the internet, potentially accumulating public interest and generating a consistent stream of advertising revenue [Guess and Lyons, 2020; Pennycook and Rand, 2020; Tucker et al., 2018]. Alternatively, some people may engage in the creation and dissemination of false information not only to disparage political or ideological factions or promote their own or collective ideological objectives, but solely derive pleasure from fomenting conflict and disorder in online spaces [Garrett et al., 2019; Marwick and Lewis, 2017; Petersen et al., 2023]. In terms of sophisticated disinformation campaigns which often entail a range of adept actors, a considerable portion of unsuspecting participants may become

entangled into unknowingly propagating and enhancing the false or misleading narrative, without being fully aware of the broader impact or implications of their involvement.

While the aforementioned traditional use of information to bolster decision-making is well-defined, the conceptualisation of the value of continually served and consumed online information to individual users, particularly when much of news is increasingly consumed as incidental by-product of extended periods of social media use [Boczkowski et al., 2018; Fletcher et al., 2015], presents a more intricate challenge. From the vantage point of online content and information creators, the wider outreach and influence within social media may strategically support their agendas by moulding the beliefs and behaviours of the connected and engaged audiences, thereby heightening the likelihood of achieving their sought-after goals, be it broader brand recognition, increased sales, political influence, social impact, sparking trends or mobilising community, and ultimately enhancing their own well-being. However, concerning online users, the information presented is frequently tailored to their existing dispositions and recent online inquiries as well as may be posted in many different formats such as text, videos, audio, images or hyperlinks to articles on other sources.

Although this customisation enhances user experience, it may not inherently contribute to any distinct or deliberate decision-making process, rendering the valuation of information on social media a complex matter. The intricacy emerges from the dual role online information frequently plays as an illuminating wellspring of knowledge and, simultaneously, as a persuasive instrument wielded by diverse stakeholders to further their own objectives. Therefore, gauging the genuine value of online information to the user, influenced by the interplay of en-

lightening insights and persuasive agendas, stands as a complex academic inquiry, warranting meticulous exploration in the digital age.

As a method for the conceptualisation and valuation of information consumed on social media, it is tenable to postulate that each individual derives a sense of gratification from attaining what they perceive as greater knowledge of the world, which may also translate into social rewards such as approval and acceptance that may provide a means to build personal social networks, and establish close relationships [Fareri and Delgado, 2014]. In this regard, the act of interacting with online content, as exemplified by scrolling through algorithmically generated material on social media platforms, may be deemed to be a mechanism providing individuals with cognitive payoffs varying in accordance with the level of perceived knowledge. Possibly enhancing the subjective sense of knowledge acquisition, the utility derived from interacting with specific information depends on the perception of understanding and the overall knowledge base in relation to the content. Nonetheless, this seemingly affirmative dynamic is imbued with nuances. Given the potentially agenda-driven nature of certain online information, not all content consistently aligns with factual accuracy or serves as a reliable source of accurate knowledge. On certain occasions, it may have adverse implications for the users it reaches. This duality underscores one of the inherent complexities of evaluating information in the digital landscape, further compounded by individual user preferences, biases, and the intricate interplay of social factors and interactions.

In a broader sociopolitical context, individuals often exhibit a tendency to aspire to uphold a socially and politically esteemed status, which is frequently realised through the conscious propagation of what is commonly perceived as the 'correct opinion.' In the digital environment, social media is a potent con-

duit for propagating societal norms, cultural pressures, and prevailing ideologies [Naranjo-Zolotov et al., 2021]. Additionally, the social milieu significantly shapes the cognitive frameworks of individuals, influencing their belief systems, attitudes, and behavioural patterns [Bandura, 2001; Bargh et al., 1996, 2001]. Individuals commonly strive to align their beliefs and perceptions to achieve internal consistency, thus mitigating cognitive dissonance [Festinger, 1962]. Simultaneously, they seek to externally conform to the prevailing norms within their social circles, bolstering their sense of belonging and in-group identity [Stets and Burke, 2000; Tajfel, 1981]. For example, within online communities, individuals may actively participate in discussions or share content that aligns with the prevailing views of their social groups, thereby solidifying their social standing and reinforcing their shared identity as well as fortifies self-affirmation [Toma and Hancock, 2013].

Such behaviour may be underpinned by a desire to evade social shame while garnering approval and respect from peers, particularly in alignment with shared political beliefs and ideological inclinations. This phenomenon finds a compelling explanation through the lens of social identity theory (SIT) [Stets and Burke, 2000; Tajfel, 1981], which, in short, posits that self-concept and behaviour of individuals are strongly influenced by their social group memberships. In particular, SIT suggests that people are motivated to maintain a positive social identity, which they achieve by associating with and conforming to the norms of their in-groups. Such adherence to group norms often leads individuals to adopt the opinions and stances endorsed by their respective social circles, reinforcing their social identity and strengthening their perceived belonging within the group. In an exploration of the impact of perceived political orientation of social media peers and individual self-objectivity on biased credibility assessments and shar-

ing of fake news, Turel and Osatuyi [2021] discovered that the alignment of fake news with the political leanings of individuals heightened both credibility bias and sharing bias in conjunction with a positive association between credibility bias and sharing bias, suggesting a greater likelihood for individuals to believe and disseminate misleading information that aligns with their political views regardless of its credibility. Additionally, the study revealed that the perceived congruence between a political orientation of a user and that of their peers acted as a mitigating factor, dampening the influence of credibility bias on sharing bias. This implies that individuals are less inclined to share fake news that contradicts their political beliefs, even if they consider it credible, when they anticipate disapproval from their social circles [Briley and Wyer Jr, 2002].

The avoidance of misinformation and disinformation plays a pivotal role in this dynamic. Consuming misinformation or disinformation may be essentially regarded as an act risking diminishing social status and potentially leading to a loss of respect and support from peer group. For instance, findings indicate that the Generation Z demographic, composed of individuals born between 1997 and 2012, demonstrates a fear of social error and sounding misinformed, wherein the implications of being wrong are perceived to have significant social costs and pose a risk to their social inclusion, consequently fostering a proclivity to scrutinise comments to orient themselves socially and actively pursue indicators that are not simply corroborating the truthfulness of information, but also the acceptance and validation of their peers [Hassoun et al., 2023]. Effectively, Generation Z individuals tend to evaluate information within the context of their established social influences, underscoring the inherently social nature of information processing which more likely than not extends beyond this demographic group, as

evidenced by Asch [1951] classic study of conformity. It is also plausible that the satisfaction individuals derive from the perceived value of augmenting knowledge is outweighed by the potentially greater decrement in their social standing resulting from the inadvertent consumption of falsehoods.

Accordingly, individuals may feel incentivised to invest valuable cognitive resources in meticulously discerning the veracity of the information encountered. By opting to exert additional effort to scrutinise the accuracy of content consumed, individuals may be assumed to be gleaning supplementary information, which, in turn, contributes to a perception of enhanced knowledge, countering the negative consequences associated with the potential exposure to misleading content. In this conceptual framework, the pursuit of knowledge operates as multifaceted strategy, interwoven with the preservation of one's social standing and the maintenance of intellectual integrity. It presents individuals with a complex decision-making process wherein they are compelled to navigate a delicate trade-off: the preservation of their social status in opposition to the cognitive costs associated with requisite research.

To put into perspective, when confronted with the decision of whether to invest significant effort, individuals are tasked with evaluating the potential returns or losses concerning their social status linked to their engagement with specific online content. Contingent on the circumstances and personal preferences, this evaluation may involve a meticulous or heuristic consideration of the benefits derived from potential incremental knowledge acquisition associated with content encountered, weighed against the costs incurred in verifying the accuracy of the said information. Amidst the deluge of information inundating users and its varying degrees of relevance, however, not only the motivation but the available

mental and physical resources to conduct further investigations differ from person to person, with users facing a serious dilemma of information triage, wherein they must judiciously prioritise what to engage with, reflecting the concept of 'cognitive resource allocation'. This challenge is highlighted by ample evidence demonstrating that a significant number of users do not consistently exert the necessary effort to confirm the credibility of online content and its sources, a phenomenon often attributed to cognitive constraints, motivation, biases and information overload in the digital age.

Beyond individual cognition, a substantial body of research examines how systemic and structural forces shape the informational environment in which individual users operate. The analyses of surveillance capitalism Zuboff [2023] and data colonialism Couldry and Mejias [2019] show how platform architectures and data-extraction logics produce profound information asymmetries. Related accounts of platform concentration and degradation Doctorow [2024] together with studies of the networked public sphere Cropf [2008] illustrate how the design and governance of digital infrastructures influence the accessibility, credibility, and circulation of information. Complementing these macro-level perspectives, empirical research on misinformation and trust by Nyhan and Reifler [2010]; Pennycook and Rand [2019]; Vosoughi et al. [2018] explains how cognitive reflection, motivated reasoning, and platform virality jointly affect verification behaviour. Taken together, these literatures establish the wider context of informational asymmetry within which this thesis situates its micro-level analysis of the user cost-benefit decision-making in verifying online content.

2.2 Model Development

While the constantly growing and evolving technologies are fuelling the dynamics of easily created, distributed, and accessible online content has piqued a simultaneous increase in concerns over misinformation and disinformation, garnering substantial attention from researchers, the prevailing focus of existing theoretical and empirical works remains primarily centred on the conceptualisation, deconstruction and characterisation of the information evaluation processes users conduct to determine the credibility of the sources and information online.

To put in perspective, these endeavours have yielded valuable insights into the complexities of online credibility establishment which, however, may not only hinge on user evaluations of the information source, the message in isolation, particularly when source information is concealed, or a combination of the two [Flanagin and Metzger, 2011; Metzger and Flanagin, 2013]. In tandem with many other factors, including source attractiveness, dynamism and rating [Kim and Dennis, 2019; O'keefe, 2015], influencing web credibility decisions, critical evaluation of online sources may be traded off for the convenience of accessing information [Connaway et al., 2011] or influenced by the order with which the source or the content is presented [Tormala et al., 2006, 2007] . Borne out by the empirical evidence of users rarely applying rigorous methods to verify the accuracy of information obtained online [Eysenbach and Köhler, 2002; Flanagin and Metzger, 2000; Scholz-Crane, 1998; Wilder, 2005], the credibility assessment may also vary with user perceptions, which, in turn, may be shaped by a range of individual characteristics. These may encompass demographic traits [Robertson-Lang et al., 2011; Sbaffi and Rowley, 2017; Zulman et al., 2011], user engagement

levels [Arazy and Kopak, 2011; Fogg, 2003; Lucassen et al., 2013; Metzger, 2007], and technological proficiency [Ahmad et al., 2010; Kim, 2012; Zulman et al., 2011].

To attain a more profound comprehension of how users evaluate the credibility of online resources, a series of studies have incorporated the various demographic, cultural, and physiological factors to investigate their influence on user information decision-making. Accordingly, variables such as age, gender, motivation, ability, familiarity, levels of information literacy, reliance on media were examined and determined as factors impacting the perception of credibility and the broader evaluative process of online credibility [Choi and Stvilia, 2015]. For instance, studies reveal that individuals with differing levels of motivation and ability employ different criteria when assessing website credibility [Fogg, 2003; Lucassen et al., 2013; Lucassen and Schraagen, 2011; Metzger, 2007]. Non-experts and those perceiving information as less personally relevant appear to often resort to straightforward heuristics such as visual aesthetics, whereas experts and those with a vested interest in the information consider other factors more significantly [Fogg, 2003; Metzger, 2007]. Moreover, when individuals possess both the motivation and ability to assess web resources, they tend to employ a more rigorous and systematic approach to evaluate credibility. Absent motivation, credibility assessment does not occur. However, when motivation exists alongside a lack of ability, users often stoop to relying on surface characteristics, peripheral cues, or heuristics to determine information credibility [Metzger, 2007]

Although the existing research revolving around credibility online has laid the definitional and important groundwork about the observed patterns of online behaviours, highlighting the glaring user deficiencies in information evaluation,

accompanied by ongoing efforts to educate and empower internet users with essential evaluative skills, the current body of literature is limited to theoretical postulations. Essentially, a discernible gap persists within the current academic landscape, as there is no formal modelling of online information consumption establishing a comprehensive framework of multi-stakeholder decision-making and cost-benefit optimisation that accounts for the complex network of interactions among a diverse range of user online behaviours, the varying levels of individual diligence applied to information verification, and the intricate interplay of various actions by different stakeholders involved in information production and dissemination, ranging from veracious to both intentionally and unintentionally misleading content, all of which motivated by distinct objectives.

In response to this void, I build a comprehensive model designed to elucidate the nuanced dynamics of information consumption on social media. The primary aim of this model is to provide a more profound understanding of the intricate interplay of factors that shape and incentivise user behaviour online as it particularly relates to information engagement. Ultimately, in these modelling efforts, the essential insights necessary to amend existing or establish novel mechanisms and inform the development of policies and strategies, with the overarching goal of mitigating the burden and exposure to online harms, specifically in the form of misinformation and disinformation.

In this model, the benefits of information are interpreted as improvements in decision quality, enhancement of social standing associated with being accurately informed, and the intrinsic satisfaction derived from perceived understanding. These benefits are evaluated relative to the cognitive and opportunity costs incurred when users choose to verify online content.

Bearing an important consideration for the model is the observation that individuals seldom engage in comprehensive information assessments, often relying on factors such as visual website design and ease of navigation to shape their decisions. For instance, users typically appear to allocate only a limited amount of time on any given website, counting on peripheral cues [Fogg et al., 2003] and adopting methods of information verification that necessitate minimal effort [Metzger, 2007]. Reflecting on the emergence of information economics, while models with imperfect and asymmetric information successfully elucidated many previously unexplained phenomena, models assuming rational behaviour with imperfect information still encountered limitations, paving the way for the emergence of behavioural economics [Stiglitz, 2017].

Having close proximity to the dynamics of online information decisions, in a significant contribution to the development of the field of behavioral economics and understanding of how individuals make choice and assess risks, Kahneman et al. [1982]; Kahneman and Tversky [1979b]; Tversky and Kahneman [1981] identified a multitude of systematic violations of the axioms of rationality in decision-making, highlighting numerous cognitive biases and cognitive shortcuts known as heuristics that influence choice and judgment. These include the anchoring bias, which involves an over-reliance on the initial piece of information provided, confirmation bias, a tendency to seek out information that aligns with pre-existing beliefs while disregarding contradictory evidence, and the availability heuristic, where decisions are made based on the information that is most easily accessible, rather than the most accurate or comprehensive information.

To top it, in his landmark research, Kahneman [2013] expounds on the concept of two distinct cognitive systems, System 1 and System 2, which govern

human decision-making processes. In brief, System 1 operates automatically and quickly, relying on cognitive heuristics and mental shortcuts to navigate complex situations quickly, while System 2 is deliberate and analytical, requiring more effort and concentration. With the ever-increasing volume of online content for users to process within a short time space, individuals may be believed to default to System 1 thought process, employing to mental shortcuts and heuristics to to cope with the cognitive and physical constraints associated with information overload efficiently, however, at risk of imparting various detractive biases to decision making. Essentially, given the tremendous amounts of data, engaging in comprehensive and meticulous information evaluation may become an arduous and costly task, prompting individuals to rely heavily on rapid decision-making strategies and cognitive heuristics. This reliance however, while facilitating rapid decision-making, may also contribute to the prevalence of cognitive biases and errors in online information evaluation.

This line of thinking is tied to the idea of bounded rationality, originally conceptualised by Simon [1955], has long been recognised by cognitive scientists as a fundamental characteristic of human information processing capabilities. It stipulates that individuals are inherently unable to consistently act in perfect accordance with rational decision-making due to various constraints imposed by the limitations of the human mind, such as finite computational resources, as well as by external conditions, including time constraints. Bounded rationality operates on the principle of least effort, acknowledging the reality that decision-makers are compelled to reach their conclusions using realistic amounts of time, information, and computational resources [Gigerenzer and Todd, 1999]. This concept sheds light on the human necessity to make decisions within practical

constraints, marking a departure from the idealised notion of perfectly rational decision-making.

These insights into information processing may collectively explain the observed low levels of online information scrutiny, in that users likely tackle the challenges of information search and overload by adopting strategies that minimise the mental effort and time involved. The argument unfolds on two fronts: one pointing to the likelihood of biases or inaccuracies in information processing stemming from the use of heuristics [Tversky and Kahneman, 1974], while the other underscores the supportive function of heuristics in enabling individuals to effectively manage the daily influx of information, often leading to sound decision-making [Gigerenzer and Todd, 1999]. Nonetheless, these findings underscore the significant costs associated with rigorous evaluation of online information amidst overwhelming information overload that prompts the adoption of heuristics and minimal effort. Resorting to mental shortcuts as such, however, may result in the acceptance of misleading or inaccurate content, although this tendency may be counteracted by individual motivations related to relevance, self-interest, and ability in managing the information encountered.

Hence, it is of paramount importance to comprehend the intricate underpinnings of the decision-making process concerning information consumption on social media. This understanding is not only crucial for establishing the pivotal mechanisms that drive or impede information credibility evaluations but also for allowing effective mitigation of the detrimental consequences stemming from the proliferation of inaccurate or misleading information. Furthermore, it is imperative to gain a nuanced understanding of its supply dynamics and the multifaceted stakeholder involvement, which can significantly complement the development

and implementation of more effective corrective strategies, thereby safeguarding the integrity and reliability of online information dissemination.

Therefore, in light of the insights provided by online credibility research into user propensities, reflecting the behavioural perspective of bounded rationality that acknowledges the human propensity to concede to various cognitive shortcuts and biases when interacting with online content, I will draw on the work by Tirole [2009], whose examination of cognitive exertion in the context of ex ante design of incomplete contracts provides a solid foundation to elaborate the cognitive mechanics of online users within the social media landscape exposed to an abundant wealth of information. Incorporating cognitive limitations within the process of information processing, I will leverage the expansive insights into the expected utility framework to represent the subjective valuations of individual benefits and costs of additional perceived knowledge under uncertainty of being exposed to potentially misleading information when modelling the decision making on the level of cognitive effort a user will choose to exert to achieve the most individually desirable outcome.

2.2.0.1 Payoff Structure

Commencing with a user-centric perspective, I will delineate the payoff structure to describe the benefits and costs associated with different choices related to online information consumption on social media. In this framework, the quest for knowledge is assumed to operate as a multifaceted strategy, intertwined with the preservation of one's social status and the maintenance of intellectual integrity, imploring an individual to make a trade-off between status preservation and the cognitive costs of research. When deciding whether to invest costly effort into

engaging with specific online content in light of the associated gains or losses to social status, an individual weighs the benefits of potentially gaining incremental knowledge against the costs of ensuring the accuracy of the information consumed and the social costs of being misled, which may have direct implications to both their social status and intellectual credibility. Furthermore, while increasing perceived knowledge may enhance social status, the incremental gain from accurate knowledge may be less than the reputational loss incurred from being misinformed or disinformed. To capture the decision making dependent on highly subjective valuations of online information consumption varying for each individual on the basis of bounds imposed to rational choice by physical and cognitive resources, as well as motivation, ability and prior knowledge, the expected utility framework is used.

The model does not refer to product-quality or purchase decisions but to the evaluation of online content credibility. Information acquisition is therefore interpreted as learning about the truthfulness or deceptiveness of social-media content rather than assessing the quality of a tangible good. At the early phase of exploring the feed delivering user-customised content posted on social media, there is a level of uncertainty associated with the credibility each piece of content bears to a user upon encounter, prior to choosing to interact with it. The ex ante uncertainty surrounding the credibility of some online content in reference to some subject may hinge on the prior familiarity with the source or the message relayed by the content, which user may inspect from evident visual cues such as the name of the source or the headline signalling the nature or relevance of content embedded in the post on social media.

Contingent upon underlying motivation as inferred from the headline of the

content, quantifiable by some parameter a that captures the value of additional knowledge gained from truthful information about some subject, users may opt to open the content linked to external sources and form an ex ante belief regarding the likelihood of it being misleading on the grounds of various superficial features of the webpage, namely, the domain name, meta descriptions, design, colour schemes, and overall functionalities of the website. Conversely, when a user is unfamiliar with the source or the subject headlined, there is an equal prior probability that information conveyed is truthful or misleading, implying that the user is equally uncertain about its credibility.

To formalise this in mathematical terms, there is probability ρ that the information relayed in content about some subject, onwards denoted by z, is misleading, where the true value of z, however, is unknown to the user at the stage of feed exploration and initiation of engagement. When user lacks familiarity to establish ex ante credibility of the content, the probability ρ of content being misleading is equal to 0.5.

Following initial visual inspection, a user may choose to engage with the content (e.g., due to the clickbait headline or some other motivational factor captured by parameter a) pertaining to some event z. If so, the user consumes the provided information on event z which the content provider leverages to steer the user towards some perceived value z_g. The perceived value z_g is equal to z if the information is truthful, but deviates further and further from the true value of z as the information becomes more misleading or inaccurate, where $z_g, z \in [0, 1]$.

After consuming the information provided by a piece of content engaged by a user, given that it is truthful and we are in the good state of the world, parameter a measures individual value to the perceived new or additional knowledge about

an event z. The maximum value one can generate through the consumption of truthful information is bounded by parameter a.

The parameter a is treated as a user-specific payoff measure of the marginal value of truthful information, capturing heterogeneity in user objectives, motivations, prior knowledge and cognitive capacities rather than being modelled explicitly as a function of underlying characteristics. It reflects that individuals consume information without a predetermined decision problem, and that the significance or payoff derived from a given piece of information depends on informational needs of the user, context, and existing knowledge.

Therefore, to first quantify the gains in the good state of the world when a user consumes information reflecting the true value of event z without having taken effort, that is, when $z = z_g$, the gains function $W_{good}(a)$ is defined as follows:

$$W_{good}(a) = \frac{a}{(1 + (0 - 0)^2)^{-1}} \tag{2.1}$$

In the bad state of the world, when the content is deceiving, $W_{bad}(z, z_g, a)$ denoting the loss function:

$$W_{bad}(z, z_g, a) = \frac{a}{(z - z_g)^2 + 1} \tag{2.2}$$

models the incremental social loss as the information deviates from the true value, which is always greater than the incremental gain that would have been achieved had the information been correct. Having described the base payoff structure underpinning online information consumption when no effort is exerted, the subsequent section focuses on integrating cognitive effort and associated costs into the model to provide a more holistic framework for understanding user

decision-making processes involved in online information evaluation.

2.2.0.2 Cognitive Effort and Transaction Costs

In the previous section, the gains and losses associated with online information consumption were defined without considering cognitive effort. In this section, the cognitive mechanisms underlying the bounded and expensive nature of information processing are specified and tied back into the payoff structure, allowing to model a better representation of the observed user interactions with online content on social media wherein many users appear to rarely perform effortful information evaluation and more often resort to using the surface characteristics to establish content credibility.

Specifically, building on prior work, I will inform the model of information processing within the broad framework of information consumption on social media by drawing on the well-established perspectives of behavioural economics on bounded rationality and approach of [Tirole, 2009] to modelling choice on the level of cognition. Based on the bounded rationality view that contracting parties use heuristics and leave contracts incomplete due to the high costs and limited cognition associated with gathering and processing information to understand all possible contingencies of a complete contract, Tirole [2009] measures the cognitive costs of exerting effort to reduce uncertainty in the ex ante design of a contract and weighs these costs against the benefits of writing a more complete contract, as well as the potential costs of ex post contract adjustments if the contract is found to be inadequate and needs to be renegotiated.

As the core of the underlying mechanisms employed in the attempt to capture the effect of cognitive efforts in processing information, the Bayesian updating is

applied to reflect that individuals revise their beliefs or probabilities contingent on new evidence or information. From a technical perspective, Bayesian updating is a powerful mathematical framework based on Bayes' theorem, which allows to calculate the posterior probability of an event, given the prior probability and the likelihood of the event. The prior probability is one's belief about the event before the new evidence is observed. The likelihood of the event is the probability of the new evidence given that the event is true. The posterior probability is one's belief about the event after having seen the new evidence.

Within the context of potential harms of misleading information online, Bayesian updating process may be considered to be a close approximation of how individuals incorporate new information into their existing beliefs as new data becomes available. It provides a systematic way to revise and refine understanding or belief maintained by an individual about a particular situation or outcome, taking into account both prior knowledge and new evidence. Concerning online information evaluation, Bayesian updating may allow to capture how individuals inform and judge online content as they navigate a deluge of online information, which is often uncertain and misleading. Consequentially, Bayesian updating may assist in accounting for the uncertainty associated with the veracity of online content conveying information about a specific event by adjusting established beliefs about it based on the related unseen information users choose to expose themselves to.

In this setting, the exertion of cognitive effort does not modify the underlying state of nature but instead enhances the precision with which that state is inferred. Accordingly, cognitive effort functions to refine posterior beliefs regarding the credibility of information, rather than to alter the objective likelihood that the information corresponds to the true state of nature.

In the online information context, an individual may choose to engage in effortful information assessment to dispel uncertainty about the nature of knowledge supplied by a piece of online content regarding some object or event z as being misleading with probability ρ. Then, conditional on the fresh information obtained as quantified by the level of cognitive effort applied, denoted by x, using Bayesian updating, the posterior probability $\hat{\rho}(\rho, x)$ of incurring a social loss linked to misleading information is:

$$\hat{\rho} = \frac{\rho(1-x)}{1-\rho x} \tag{2.3}$$

The intuition behind is that by making the decision to put additional effort to reduce uncertainty whether the content is misinformation or disinformation, an individual may be assumed to have explored and sought out additional information, thereby increasing their perceived knowledge and thus feeling more knowledgeable while ameliorating the negative effects of potentially having been exposed to untruthful information via the content user engaged on social media. Put differently, equation 2.3 captures how cognitive effort can reduce the perceived probability of experiencing social loss due to misleading information. Illustrating this concept, Figure 2.1 demonstrates how the uncertainty surrounding the veracity of online information diminishes as more cognitive effort is exerted.

Effectively, this suggests that while individuals may choose to exert different levels of effort, the extent to which they will clarify the uncertainty pertaining to the credibility of information about some event z posted in some online content encountered will vary because the posterior probability $\hat{\rho}$, which reflects the decrease in uncertainty about information on z conveyed by the content being

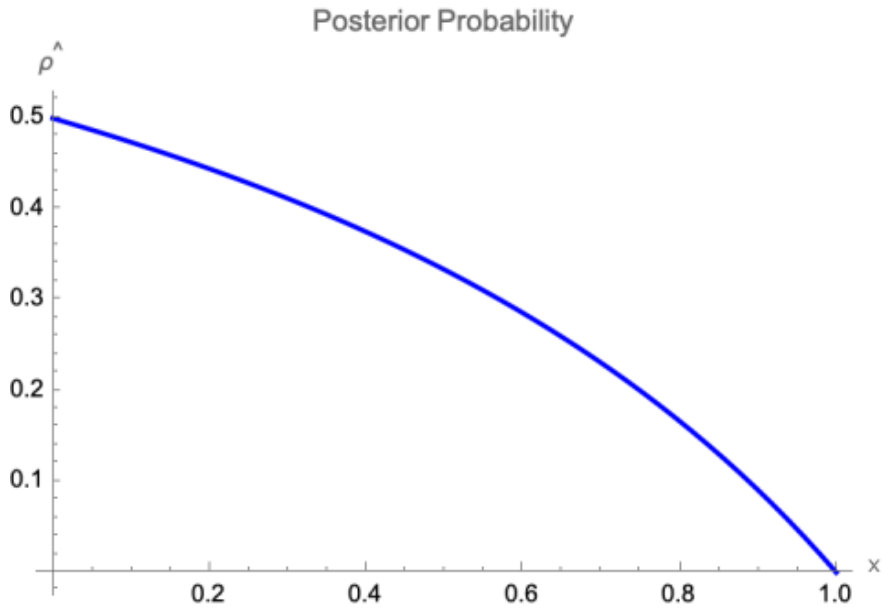

Figure 2.1: Cognitive Effort and Uncertainty

misleading, is conditional on cognitive effort b. However, cognitive effort allocated to each piece of online content is bounded and costly. Moreover, the impending consumption of the vast volumes of online content about other events as well as individual factors such as motivation, ability, and prior knowledge about the event conveyed by the online content at the time of encounter, may all influence the levels of effort exerted and the resulting cognitive costs.

One way to encapsulate this is to consider individual users as having a limited capacity for cognitive effort, which is depleted as they consume online content. The amount of effort required to process a given piece of content will depend on a number of factors, including the complexity of the content, prior knowledge of the topic, and the level of interest as well as the aforementioned individual factors such as motivation, ability, and prior knowledge. For example, users who are more motivated to learn about a particular event are likely to be willing to exert more

cognitive effort in processing content related to that topic. Similarly, users with higher levels of ability and prior knowledge may be able to process content more efficiently, requiring less cognitive effort. Therefore, despite the assumption that users are generally disposed to acquiring what they perceive as greater knowledge, the extent to which they will pursue to clarify the information about some event z will be constrained by the cognitive costs $C(a, x)$ they individually incur when putting effort x:

$$C(a, x) = \frac{x}{1 - x^{1/2}} * a, \tag{2.4}$$

where $C(x)$ is proportional to the subjective value of knowledge, a, information provides.

In this setup, the cognitive costs faced by users are effectively a function of their individual characteristics and the levels of effort they exert. To provide a frame of reference, these costs may manifest in a variety of ways, which besides experiencing negative cognitive states such as difficulty concentrating, fatigue, and stress, may be the opportunity costs of consuming content about other events, forfeiting gaining more perceived knowledge of the world lost in the time period devoted to investigating some event z. As a visualisation, Figure 2.2 displays the effect of increasing cognitive effort on the transaction costs associated with information processing.

In particular, in the good state of the world having already consumed true information about event z, the cognitive effort will detract the value from W_{good} by having wasted time which could have been spent engaging other content:

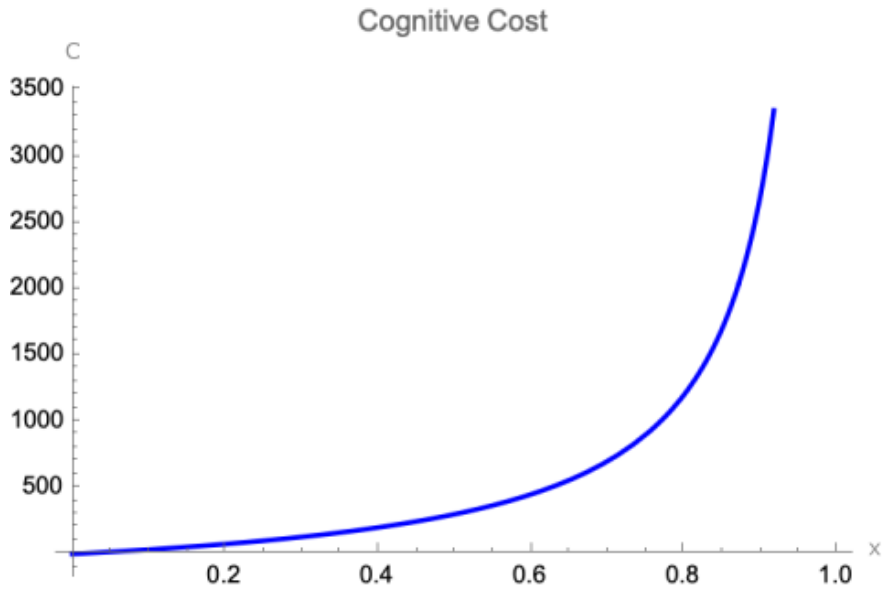

Figure 2.2: Cognitive Effort and Transaction Costs

$$W_{good}(a, x) = \frac{a}{(1 + (0 - 0)^2)^{-1}} - C(a, x) \tag{2.5}$$

Conversely, in the bad state of the world when information is misleading, the social loss will increase:

$$W_{bad}(z, z_g, a, x) = \frac{a}{(z - z_g)^2 + 1} - C(a, x) \tag{2.6}$$

Overall, the cognitive costs of processing information can be seen to be an inhibitor of the extent to which users will pursue information clarification, where gains diminish and losses increase with rising cognitive effort. In the next section, the expected utility framework is adopted to analyse how the choice of different levels of cognition employed in online information consumption on social media affects the satisfaction one derives.

2.2.0.3 Utility Function and Expected Utility Theory

Constrained by physical and cognitive limitations imposed by an excess of information created and propagated by multiple, diverse and in many cases questionable sources and stakeholders with questionable intentions, users are posited to optimise their online information consumption within the confines of bounded rationality. In this endeavour, users select a level of cognitive effort, x^*, that achieves the most optimally satisfactory trade-off between the subjective utility they can derive from information about event z, which provides them the gratification of acquiring an enhanced level of perceived knowledge, the associated cognitive costs, and the potential loss of perceived social status by having consumed misleading or inaccurate information.

To capture idiosyncratic preferences and valuations of online content shaped by the information acquired, a Constant Relative Risk Aversion (CRRA) utility function of the form (2.7) is employed to model the trade-offs between the various factors that influence user decision-making, graphically represented by the corresponding Figure 2.3.

$$U = \frac{W^{1-G}}{1-G} \tag{2.7}$$

where G is the coefficient of relative risk aversion, influencing the degree of risk aversion exhibited by the individual. Specifically, when $G < 0$, the individual demonstrates risk-seeking behaviour, favouring risky options over certain outcomes. When $G = 0$, utility is linear in consumption, corresponding to risk neutrality. The case where $G = 1$ corresponds to a logarithmic utility function, indicating a constant relative risk aversion (CRRA = 1), which implies that the

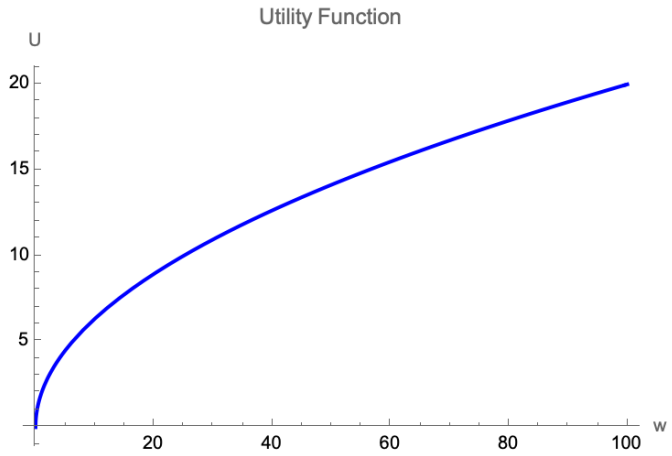

Figure 2.3: Constant Relative Risk Aversion (CRRA) Utility Function

individual is moderately risk-averse. In contrast, if $G > 1$, the individual exhibits stronger risk aversion, preferring certain outcomes over risky alternatives. Ultimately, the higher the value of G, the more risk-averse the individual is.

In the present context, greater risk aversion amplifies welfare losses arising from uncertainty about content credibility, as individuals with higher G experience sharper curvature in their utility function and therefore greater sensitivity to variations in expected outcomes. This relationship connects informational asymmetries and cognitive costs directly to welfare. While stronger risk aversion increases the incentive to verify information to mitigate exposure to misleading content, it may also lead to earlier disengagement when the marginal disutility of verification effort, reflected through the utility function, outweighs the anticipated informational benefit. Risk preferences therefore determine both the motivation to pursue accuracy and the point at which further effort ceases to be optimal.

Then, the expected utility of consuming online information about event z influenced by the exertion of cognitive effort$x$is:

$$E[U] = (1 - \hat{\rho}) * U[W_{good}(a, x)] - \hat{\rho} * U[W_{bad}(z, z_g, a, x)] \tag{2.8}$$

The outlined expected utility analysis facilitates the investigation of how the interaction with information about event z and the endogenous choice of cognitive effort$x$impact one's perceived subjective wellbeing which in this framework may be positive or negative. The potential implications of information-related decision-making on the expected utility are conveyed in Figure 2.4, providing a general overview of the impact cognitive effort may have on user welfare.

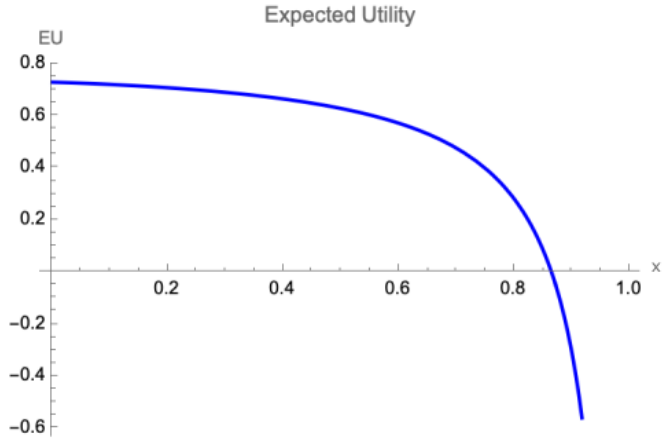

Figure 2.4: The Expected Utility and Cognitive Effort

2.2.1 Cognition in Information Decisions

To commence the construction of an in-depth model examining online information consumption, crucial theoretical frameworks and foundational principles are delineated, laying the groundwork for a comprehensive analysis and formulation. In particular, in combination with the conceptual and empirical insights into the cognitive processes furnished by the extensive literature on credibility judgment

of online content, the model of limited cognition posited by Tirole [2009] which, despite primarily addressing cognitive constraints within a contractual setting, presents a framework whose relevance is extendable to the broader sphere of online information consumption. More specifically, by highlighting the constraints imposed by limited cognition in decision-making processes concerning information necessary for writing efficient contracts, this framework and its insights may well generalise and shed light on how individuals navigate and process information in online environments, providing a lens to comprehend how cognitive limitations impact choice of online information to assimilate. Thereby, Tirole [2009] lends a valuable foundation to explore the nuanced dynamics involved in decision-making regarding online information consumption, against the backdrop of bounded and expensive nature of of processing unlimited amounts of available information.

To put into perspective, in contrast to the idealised assumptions of the mainstream contract theory, where contemplating contingencies, devising covenants, and discerning their ramifications incur no costs, the parties to a real-world contract grapple with incomplete information and lack of comprehensive awareness regarding all possible implications. While they may adhere to the available industry standards in contract design, they remain ignorant about the full scope of the consequences such contract bears, yet aware of potential uncertainties. As a result, parties often engage in cognitive efforts to anticipate potential risks and tailor contracts accordingly. However, the practicality of foreseeing certain contingencies is hindered by the prohibitively high costs associated with such foresight. Therefore, despite being foreseeable in principle, certain outcomes might not be actively anticipated or incorporated. As exemplified by Tirole [2009], while it is conceivable that oil prices might surge, warranting indexation of contracts, par-

ties might overlook this possibility in their deliberations, failing to adjust contract terms accordingly.

Acknowledging such costs associated with collecting and processing information which tend to lead to the utilisation of heuristics in contract design, Tirole [2009] merges the bounded rationality approach with that of mainstream economics. This amalgamation accounts for the cognitive limitations within the framework of rational decision-making by economic agents, recognising that contracting bodies are generally unaware, albeit cognisant of their bounded awareness regarding all facets of the contracting environment and the inherent uncertainty in future events. Two fundamental implications arise from this perspective, diverging from traditional contract theory. Firstly, it addresses the inherent transaction costs entangled in negotiation processes. Secondly, it challenges the conventional view favouring complete contracts, underscoring their potential inefficiencies. Prompted by individual interests, parties engage in contract refinement to avert post-contractual exploitation of being held up by the other party. To this end, the process of completing contracts often involves strategic manoeuvres aimed at securing advantageous outcomes.

In this light, Tirole [2009] defines an incomplete contract as initially describing a specific design, which in the event of being deemed inadequate, is permitted to be renegotiated. To that end, a scenario where a buyer and a seller enter an agreement for the delivery of a design A is constructed, allowing to determine the instances where contracts adhere to an available design initially and then undergo subsequent renegotiation. Effectively, this design A might not necessarily meet the requirements of the buyer, in which case an alternative undisclosed formulation of contract, A', may benefit the buyer more if the seller cooperates.

At the outset, both parties are knowledgable only about design A, acknowledging its potential mismatch. Prior to signing the contract, each party may dedicate cognitive effort to exploring alternatives to design A. Should one party discover that design A' rather than A is more suitable, they encounter the decision of whether to convey this alternative design option to the other party. The disclosure of A', however, exposes the state of nature and A is no longer a viable choice.

In terms of the incompleteness of a contract itself, Tirole [2009] stipulates it to be contingent upon the extent of resources allocated to delineate the suitable design. It follows that a greater incompleteness is exhibited if lower amount of resources is allocated to specify the correct design. Alternatively, contract incompleteness is gauged by the likelihood that the originally specified design necessitates subsequent ex post alterations. The incurrence of transaction costs may be deemed inefficient since every party has a vested interest in understanding their susceptibility to, or potential gains from, renegotiation. On the other hand, expenses arising from ex post contract adjustments may rationalise specific cognitive investments from a societal standpoint. In this fashion, two factors contribute to diverting the buyer from optimal cognition with one revolving around the desire to circumvent potential post-contract exploitation, while the other involving an initial upfront discount offered by the seller, who expects to capitalise through the standard but inappropriate contract, reducing the inclination of the buyer to deviate from it.

In this modelling framework, Tirole [2009] delves into the drivers behind equilibrium transaction costs, yielding pivotal insights. Broadly summarised, two critical observations unveil. In contractual agreements, adverse selection appears to be inherently introduced by cognition. What is more, contracts may poten-

tially suffer from over-completeness, implying an overflow rather than a scarcity of information influencing the design. In essence, this model proposes that individual stimulus for cognitive effort originates not solely from the endeavours to evade post-contract adjustments but also from rent-seeking intents.

While Tirole [2009] masterfully unifies various strands running through contract research, interlinking multiple perspectives and theoretical frameworks, without further elaboration on this synthesis, the focus is shifted directly on the detailed description of his model of limited cognition and its subsequent analysis.

The model is set up between a *Buyer* and a *Seller* who contract on the production and delivery of some item. At the initiation of contracting process, there is a universally-known contract design A which costs c for the seller to fulfil. While there is probability $1-\rho$ that A is suitable, providing utility $v > c$ for the buyer, with probability ρ a contract design A' unspecified during the contract stage furnishes buyer utility of v, in which case, A then only provides $v-\Delta$, where $\Delta > 0$. Once contract is signed, buyer needs seller to collaborate to transform A to A'. This leads seller to incur additional adjustment costs of $a \in [0, \Delta]$. As a result, by having to renegotiate, the gains will amount to $\Delta - a$. If design A' is originally delineated instead of A, the production cost totals c, devoid of any adjustment expenses.

Tirole [2009] suggests that these adjustment costs may pertain to physical designs such as computer codes or engine specifications wherein transitioning from A to A' may necessitate code modifications or aligning the existing engine with new specifications. If A and A' denote fundamental aspects of contract design such as covenants or indexation, the adjustment cost may signify tangible expenses arising from unforeseen but foreseeable liquidity positions or inefficient

risk distribution.

Within the context of contracting, the model has several intuitive interpretations. In a technology licensing scenario, the seller may license technology to the buyer, who, only post-contract, might discover the need for an additional license for another patent owned by the seller to be able to effectively operationalise the technology for a specific use. In procurement, the buyer may ascertain that the specified design falls short of the requirements, warranting supplementary requisitions from the seller. Ultimately, the model may be interpreted in terms of the ramifications of a specified contract itself. Regardless of a contract stipulating a specific course of action, the seller may be able to deploy more cost-effective approaches to meet the contractual obligations which, however, may be less appealing to the buyer. Should the delivery of suboptimal output lead the seller to cost-savings lower than the resultant surplus reduction born by the buyer, renegotiations to observe the essence, if not the explicit wording of the contract, may be initiated.

Resuming the description of the model, design A is presumed to engender trade benefits irrespective of cognitive input, $v - c - \rho a > 0$. To encapsulate the implications of renegotiation, a comparison between the following two instances are drawn. Essentially, under a contract originally stipulating design A' at a designated price p, the execution ensues seamlessly with the cost c born by the seller, who fulfils the according contractual obligations, while the buyer receives utility v. If, however, design A is selected, although the seller is still subject to the cost c, upon receipt of the good, the buyer assesses its appropriateness, which, in the event of inadequacy, results in contract being renegotiated to initiate the procurement of the adjustment to design A'.

The buyer and seller are attributed with bargaining powers β and σ, respectively, representing their respective capacities to claim a portion of the trade benefits during negotiations, where $\beta + \sigma = 1$. This allocation of bargaining power is maintained uniform for both parties before and after the contractual agreement as a means of simplification.

Tirole [2009] defines transaction costs, $T_B(b)$ and $T_S(s)$, that may be accrued by either the buyer and the seller, respectively, in the form of cognitive effort engaged in understanding the optimal design criteria. However, the outcomes of their inquiries may vary. Provided that A' is the fitting design, the buyer and the seller hold the respective probabilities of b and s of discovering it, otherwise facing the likelihood, expressed by probabilities of $1-b$ and $1-s$, of encountering inconclusive research outcomes. However, if A is identified as the suitable design, no additional information is gained. These cognitive cost functions T_i, for each party $i \in \{B, S\}$, exhibit properties of smoothness, monotonicity, and convexity, satisfying the following boundary conditions: $T_i(0) = 0$, $T_i'(0) = 0$, and $T_i(1) = \infty$. While Tirole [2009] does not explicitly define it, one such function depicting the cognition costs perceivable by the buyer may take the following form:

$$T_B(b) = \frac{e^{rb}}{1-b} \tag{2.9}$$

where r signifies cost elasticity. As the value of r increases, the buyer perceives exerting effort b to be more costly. In brief, cognitive costs may encompass a number of expenses, varying from managerial mental strain to the opportunity cost of wasted time as well as legal and consulting fees. The extent of these costs is observed indirectly through the widespread incompleteness of numerous contracts

and the subsequent financial implications stemming from this deficiency. On this account, the relative incompleteness of a contract is assumed to be inversely proportional to the ex ante probability that the design A is identified in the contract.

Focusing only on the analysis of the one sided cognition, the sole capability to investigate and determine the suitable design is assumed to rest with the buyer, who may be more aware of their needs and preferences. Tirole [2009] expresses the socially efficiently level of cognition as $\hat{b}$ at which the marginal cost of thinking is equal to its marginal benefit, $T'_B(\hat{b}) = \rho a$, where ρa signifies the evasion of the incurring adjustment cost a when A' is the fitting design. In the absence of adjustment costs ($a = 0$), any cognitive investment in this model becomes purely speculative, driven by rent-seeking motives and thus, exerting no effort is the most socially optimal solution.

Considering a deterministic cognition region where a pure-strategy equilibrium is obtained, b^* is indicated to represent the equilibrium probability at which the buyer discovers the factual inadequacy of A. On the other hand, the investigations of the buyer may prove fruitless and design A may be contracted. By applying Bayesian condition, Tirole [2009] arrives at:

$$\hat{\rho}(b) = \frac{\rho(1-b)}{1-\rho b} \tag{2.10}$$

where $\hat{\rho}(b)$ is the posterior probability contingent on being unaware and cognitive effort b that design A is inappropriate. Given that A' appears to be the suitable design, a portion σ of the renegotiation gain is seized by the seller, $h = \sigma(\Delta - a)$, representing a hold-up situation wherein a party, in this case the

seller, may extort surplus from a contract.

Effectively, when $b = b^*$, both parties anticipate a hold-up which manifests as a benefit to the seller and a cost to the buyer $\hat{\rho}(b^*)h$ and whose likelihood is thereby accounted for in the ex ante price $p(b^*)$ for design A:

$$\sigma[v - c - \hat{\rho}(b^*)\Delta] = p(b^*) - [c - \hat{\rho}(b^*)h], \tag{2.11}$$

or

$$p(b^*) = c + \sigma[v - c - \hat{\rho}(b^*)\Delta] \tag{2.12}$$

The expression on the left side of equation (2.11) represents the portion of the total surplus under the negotiations for contract delineating design A acquired by the seller, with the right-hand side denoting the profit seller earns. Essentially, while the seller may be able to hold up and exploit the buyer for an amount h with a conditional probability $\hat{\rho}(b^*)$, equivalently, $c-\hat{\rho}(b^*)h$ presents as an opportunity cost to the seller. The equation (2.12) determines the price $p(b^*)$ ensures that the seller secures a portion σ of the pre-contractually anticipated total surplus. The expression $\rho(b^*)h$ may be seen as representing a discount that the seller endows to dissuade the buyer from cognition, prompting agreement on design A due to the potential for post-contractual hold-up.

Within this set-up, the buyer is faced with deciding the optimal level of cognitive effort b^* to exert prior to signing the contract, a choice which is resolved through the process of solving:

$$\max_b \left\{ -T_B(b) + \rho b\beta(v-c) + \rho(1-b)[v-a-h-p(b^*)] + (1-\rho)[v-p(b^*)] \right\} \tag{2.13}$$

which, as per Tirole [2009], simplifies to:

$$\max_{b} \left\{ -T_B(b) + \beta(v-c) + (1-\rho b)\hat{\rho}(b^*)\sigma\Delta - \rho(1-b)(a+h) \right\} \tag{2.14}$$

To deduce equation (2.13), there is probability ρb that the buyer puts forward design A', securing a β portion of the combined surplus $v-c$. Conversely, with probability is $1-\rho b$, the design A is agreed at the price $p(b^*)$. There is also probability $\rho(1-b)$ of the suitable design remaining unknown, imposing the burden of the adjustment cost a compounded by the hold-up h on the buyer.

The differentiation of equation (2.13) with respect to the equilibrium condition $b = b^*$ leads to the first-order condition:

$$T'_B(b^*) = \rho a + \rho[h - \hat{\rho}(b^*)\sigma\Delta] \tag{2.15}$$

In brief, the left-hand side of equation (2.15) signifies marginal cost of cognitive effort. On the opposite side of (2.15) lies the marginal benefit associated with cognition which is constituted of ρa representing the social benefit, ρh depicting the incremental gain the buyer achieves by evading a potential extortionate hold-up, and $-\rho\hat{\rho}(b^*)\sigma\Delta$, denoting the decrease in the negotiated price when opting for design A' rather than A.

Applying the Bayesian updating condition (2.10) to (2.15) obtains:

$$T'_B(b^*) = \rho a + \rho\left[h - \frac{\rho(1-b^*)}{1-\rho b^*}\sigma\Delta\right] \tag{2.16}$$

which encapsulates both the key dynamics and the central results outlined in the model of limited cognition proposed by Tirole [2009] where individuals choose the optimal intensity with which to collect and process information considering the limited available cognitive resources in the context of contract negotiations. Having examined how bounded rationality is approached within a contractual environment, the Bayesian learning updating function emerges as a particularly compelling tool to integrate into a framework of online information consumption to illuminate the decision-making process regarding the efficient levels of cognitive effort in information acquisition, evaluation and consumption. To further reinforce its significance for understanding online content decisions, the analysis delves into an inquiry of how the cognitive mechanics involved in processing information have been addressed within existing research.

Throughout consumption and dissemination of misleading information, individuals may frequently appear to be motivated to maintain a positive social identity for which they are willing to adopt the opinions and stances endorsed by their respective social circles, leading to not only biased credibility assessments but sharing of false online content disregarding low credibility. Manifestation of akin behavioural tendencies is scrutinised by Bénabou and Tirole [2002], who delve into into the rationale behind the valuation of self-image and esteem.

Additionally, they explore the methods through which individuals strive to amplify or maintain these qualities via a spectrum of seemingly irrational behaviours, spanning from excessive pessimism to self-delusion. In their endeavours, Bénabou and Tirole [2002] build a model of self-deception utilising endogenous memory wherein the motivated and rational facets of cognitive processes are synthesised. Specifically, the analysis concentrates on the significance ratio-

nal individuals attribute to self-confidence as well as the tactics employed in its pursuit in order to elucidate their implications for information processing and decision-making.

As is the case with the observed online behaviour, self-confidence and positive rather than accurate self-beliefs appear to contradict the conventional view of rational human behaviour and cognition within the domain of economics. For an examination of it welfare consequences, the authors construct a simple formal framework of the demand and supply aspects of self-confidence that illuminates the economic implications of these psychological phenomena.

Regarding the demand side of self-confidence, Bénabou and Tirole [2002] delineate three primary reasons explaining why individuals might favour optimistic self-perceptions over accurate ones, namely, a consumption value, a signalling value, and a motivation value. However, the bulk of the analysis concentrates on the motivation value aspect. While all three are regarded as equally compatible with the analysis of the supply side, the motivation-based theory demonstrates broader explanatory power. Effectively, it generates an endogenous value of self-confidence that adapts to an individual circumstances and incentives, encompassing what Bénabou and Tirole [2002] refer to as "can-do" optimism and "defensive" pessimism.

In a broader context, self-confidence is noted to hold intrinsic value by bolstering motivation to initiate and sustain efforts directed at achieving personal goals in the course of which their resolve is tested. Extensive discussions within the realm of psychology literature, as evidenced by the works of Bandura [1997]; James [1890]; Seligman [2006], underscore the pervasive link between self-confidence and motivation.

As an illustration, individuals with heightened confidence tend to be more driven to intensify their efforts [Puri and Robinson, 2007]. Moreover, immediate emotional needs may also be catered by motivated beliefs with research indicating that individuals often experience better psychological well-being by embracing a more optimistic outlook [Alloy and Abramson, 1979; Korn et al., 2014]. Conversely, when individuals lean towards pessimism, understating accomplishments and deluding themselves about the imminence of supposedly greater difficulties may contribute to retaining their self-esteem.

Informed by the postulations on motivation in psychology [James, 1890; Nisbett and Wilson, 1977; Salancik et al., 1977], Bénabou and Tirole [2002] propose that while individuals often lack complete cognisance of their own abilities and the potential costs and benefits their actions may deliver. They suggest that ability commonly supplements effort, and their interplay significantly determines performance, wherein higher self-confidence fosters motivation for action. Bearing an important connection to the dynamics of influence online actors and content on associating users incentivised by being rewarded a sense of belonging, consolidating their social and self-identity, as per Bénabou and Tirole [2002], individuals personally interested in the performance of another individual are motivated to cultivate and uphold the self-esteem of that person to influence their commitment to tasks in various settings, including but not limited to educational or professional environments [Bénabou and Tirole, 2003]. This embodies a principal-agent relationship, wherein the informed party assumes the role of the principal, offering incentives to guide the actions of the agent.

Fundamentally, these conceptualisations may elucidate the phenomenon where online information choices are markedly characterised by confirmation and belief

biases. These cognitive predispositions lead individuals seeking information to validate their pre-existing beliefs and evaluate the strength of an argument based on the acceptability of its conclusion rather than its logical validity in order to conform to the norms of their in-groups, which, however, may be exposed and gravitated toward false information. In a self-fulling manner, these biases may emerge and strengthen during the process of individuals being motivated through manipulation from their social groups and prominent figures that may be regarded as the principals of online users.

As it relates the supply side of self-confidence, the Bénabou and Tirole [2002] impose constraints which restrict the degree to which individuals can exaggerate reality. Since rationality and the pursuit of information align with the overarching principles of Bayesian updating accentuated in classical literature [Festinger, 1954; Heider, 2013], Bénabou and Tirole [2002] assume the convention that individuals are Bayesian learners.

On the other hand, they also acknowledge the perspective upheld in contemporary cognitive literature which extensively showcases the non-rational, or at minimum, subjectively driven facets of human reasoning. Notably, a wealth of evidence highlights individuals to have a proclivity to recollect successful over futile outcomes, displaying memories and perception that are self-serving and biased. What is more, individuals often appear to misjudge their abilities and qualities in conjunction with their perceived control over outcomes. On top of that, and especially with the conviction of their ability to sway future events through their actions, the assessment of their personal probabilities is skewed to the right and left of the mean value for positive and negative future life events respectively [Alloy and Abramson, 1979; Gilbert and Cooper, 1985; Gilbert et al.,

1998; Taylor and Brown, 1988; Weinstein, 1980].

Delivering a significant contribution, Bénabou and Tirole [2002] capture the complexities of self-deception via a straightforward game-theoretic model of memory management. The framework bridges the gap between motivated and rational cognitive elements by leveraging the findings related to memory processes and its constraints and its potential utility extends to any context where motivated beliefs play a role. In essence, the model relies on the notion that individuals, within specific restrictions and with potential trade-offs, are capable to impact the likelihood of recalling specific information while retaining a level of logical reasoning at which they recognise having a capacity for selective memory.

The emergent framework delineates a dynamic interplay between the present and future selves of an individual conceptualised as a strategic communication game. When contemplating whether to suppress negative information, an individual balances the potential benefits of maintaining motivation with the risk of unwarranted optimism and exaggerated sense of confidence. Subsequently, the individual acknowledges the limited reliability of positive memories. To provide a succinct overview of the model to elucidate its relevance to understanding the dynamics of online information behaviour, it is set upon assuming a risk-neutral agent whose actions span a three-period horizon, $t = 0, 1, 2$.

At the initial decision point in period 0, the individual makes a choice that may influence both their flow payoff, u_0, and the information they possess at the subsequent stage in period 1. In period 1, individual deliberates whether to engage in an action which involves an exertion of effort incurring a cost $c > 0$, or to refrain from such effort. Representing the inherent ability of an individual, there is probability θ of the project succeeding and generating a benefit V at

the final stage in period 2, however, with probability $1 - \theta$ that failure will be experience resulting in no reward. Distribution functions $F(\theta)$ in period 0 and $F_1(\theta)$ at period 1 represent beliefs held by an individual about θ specifying their self-esteem.

During the interim period, an individual may encounter new information concerning their abilities, who, however, may prefer either to remain informed to avert overconfidence or to remain uninformed to preserve their confidence. In some instances, an individual may also choose self-handicapping strategies, deliberately impeding execution of their own actions. These tactics may manifest in the form of perfunctory effort, insufficient preparation, pre-task alcohol consumption or establishment of unattainably challenging goals, to name a few behavioural attempts at self-esteem preservation [Berglas and Jones, 1978; Fingarette, 1985; Gilovich, 2008]

To that effect, their intertemporal model provides a means to disentangle the cognisant and unaware states of self, synthesising the motivational and cognitive dimensions of self-deception. To illuminate the underlying concept of individuals who, to a certain extent, may influence the likelihood of recalling specific items of information, the motivational aspect of the mind is driven by the incentives, arising under the circumstance of temporal inconsistencies, for individual to remember information which supports long-term objectives and aspirations, while suppressing contradicting information. In terms of the cognitive side of the mind, the principle of rational inference which asserts that individuals are cognisant of their selective memory is upheld.

When memory is integrated into the model, individuals are assumed, albeit at a certain cost, to have the ability to modulate the probability of recalling or

accessing information received an earlier point in time. More specifically, this probability of recalling or accessing information in period 1 which, however, was acquired in period 0 is modelled as by $\lambda \in [0, 1]$. As a reference point, the natural rate of recollection, denoted as $\lambda_N \in [0, 1]$, is defined as the optimal level of memory that maximises the flow payoff, u_0 in period 0. Altering λ by increasing or decreasing it from its natural rate λ_N subjects an individual to a memory cost delineated by the function $M(\lambda)$, resulting in a decrease in u_0. This cost function adheres to the properties of $M(\lambda_N) = 0$, $M'(\lambda_N) \leq 0$ for $\lambda < \lambda_N$, and $M'(\lambda_N) \geq 0$ for $\lambda > \lambda_N$.

Numerous psychological studies and experiments demonstrate the malleability of memory, suggesting that individuals can exert a degree of control over what information they are more likely to consciously remember [Fazio and Zanna, 1981; Greenwald, 1980; Jones et al., 1981; Schacter, 1996]. This cognitive flexibility creates an avenue for motivated cognition, whereby individuals selectively retain and process information in a manner that aligns with their personal goals or beliefs. For instance, an individual who desires to maintain a positive self-image might focus on positive experiences, rehearse positive affirmations, and cultivate environments that reinforce their self-perception of success. Contrarily, they might actively avoid situations that evoke negative memories or threaten their self-esteem.

It is imperative to note the distinction between direct memory suppression and more indirect mechanisms that modulate memory accessibility. Specifically, Bénabou and Tirole [2002] maintain the model compatible with both Freudian and cognitive perspectives on memory dynamics. According to Freudian theory, memories may be repressed or relegated to the subconscious mind, with

the potential for later resurfacing and reappraisal. In contrast, while contending that control over memory itself is implausible, cognitive psychology underscores a spectrum of factors, such as the degree of attention during information collection and retrieval, the evasion or pursuit of cues, or rehearsal of specific information, which may influence consciousness and awareness.

Moreover, individuals who exhibit a consistent pattern of forgetting, distorting, or repressing specific information are likely to recognise this tendency. Consequently, they are less inclined to uncritically accept the apparent bias in favour of positive recollections regarding their past performances and received feedback. On this account, an individual employs a measure of rational reasoning to discern whether the information they may have suppressed is not merely random occurrences but rather purposeful omissions. This self-reflection, which highlights the fallibility of self-knowledge, Bénabou and Tirole [2002] formalise by applying Bayes' Rule, indicating that a person cannot systematically deceive themselves in the same manner.

To concisely summarise the incorporation of Bayesian learning in their model to succinctly outline how their model incorporates Bayesian learning, an individual may receive feedback σ regarding their abilities during period 0 which, with probabilities $1-q$ or q, may either be bad, $\sigma = L$, or no feedback may be received, $\sigma = \oslash$, respectively. Effectively, if this feedback is perceived as detrimental to self-perception, an individual may suppress it from their conscious awareness. However, when reflecting upon the recollected feedback in period 1, defined as $\hat{\sigma} \in [\oslash, L]$, if no negative feedback is remembered in period 1, $\hat{\sigma} = \oslash$, it prompts individual to conduct an internal inquiry. This inquiry involves examining whether there was genuinely no unfavourable feedback in period 0 or if

such information was potentially disregarded or obscured by the individual during period 0. Given that the individual perceives the likelihood of remembering negative information to be λ^*, Bayes' rule is utilised to calculate the reliability, referred to as r^*, of having no recollection in period 1 as:

$$r^* = Pr[\sigma = \oslash | \hat{\sigma} = \oslash; \lambda^*] = \frac{q}{q + (1-q)(1-\lambda^*)} \tag{2.17}$$

In essence, when confronted with self-esteem threatening feedback, $\sigma = L$, during period 0, an individual decides the probability of recalling this information, λ, in a way that optimises their overall well-being by solving:

$$\max_{\lambda} \{\lambda U_T(\theta_L) + (1_\lambda) U_C(\theta_L | r^*) - M(\lambda)\} \tag{2.18}$$

where $U_C(\theta_L|r^*)$ defines the expected utility in period 0 when an individual effectively forgets the occurrence of negative information, whereas $U_T(\theta_L)$ denotes the expected utility when information is remembered correctly, with the respective subscripts C and T representing censored and true recollections.

Owing to the Bayesian rationality demonstrated in period 1, the individual acknowledges that the decision made in period 0 involved a strategic selection of the recollection rate λ based on opportunistic reasoning. Therefore, this optimal λ is subsequently employed by the individual when evaluating the reliability of their own memories.

It is pertinent to note that the model also assumes individual preferences to be subject to time inconsistency attributable to quasi-hyperbolic discounting. This notion is corroborated by substantial evidence on the manifestation of a present-bias in intertemporal decision-making, wherein discount rates signifi-

cantly decrease over shorter time frames compared to longer ones [Ainslie, 1992, 2001; Laibson, 1997, 2001; Loewenstein and Prelec, 1992; O'Donoghue and Rabin, 1999; Phelps and Pollak, 1968; Strotz, 1973]. Bénabou and Tirole [2002] encapsulate these tendencies using the parameters δ to represent a conventional discount factor, and β to symbolise the present-biased nature of preferences. When $\beta < 1$, the individual in period 0 anticipates the preferences of their selves in period 1 to be excessively present-focused, which may consequently precipitate reduced effort, culminating in insufficient effort, a behaviour commonly associated with procrastination.

Whilst unveiling important implications self-esteem has for decision making, the finding that bears the most significance to the analysis of online information consumption relates to the dynamic of information processing. In particular, Bénabou and Tirole [2002] show the pivotal role of Bayesian-like introspection which embodies the ability to partially grasp motivations for self-esteem preservation rather than representing individuals as passively accepting all recollections. In essence, this approach seems to align seamlessly with the well-documented inclinations observed in online information behaviour, wherein the assessment of content credibility may be shaped by a multitude of diverse public and personal signals. These signals may stem from a variety of sources and target different individuals for multifaceted reasons. As individuals may be induced to cultivate a favourable social identity congruent with their social affiliations, exerting substantial influence on their self-perception and conduct, it becomes plausible to argue that some people may be motivated to engage in self-deception of endorsing and circulating false online news, while retaining a degree of self-awareness of the possible perpetuation of misinformation.

To underscore the importance and growing attention given to the motivational role of self-deception in decision-making, a growing body of research, extending beyond psychology to encompass economics, has delved into the concept of motivated false memory. Drawing upon the work of Bénabou and Tirole [2002], Chew et al. [2020] explore motivated false memory through an economic lens. Their research entails a large-scale experiment aimed at investigating the relationship between memory errors and individual preferences such as attitudes toward time, risk and ambiguity, and other psychological attributes or characteristics.

For a broader contextualisation, the prevalence of motivated false memory bears significant real-world implications, such as its role in augmenting self-image to improve employment prospects [Heckman and Rubinstein, 2001] or contributes to fostering collective delusions within organisational settings, aiming to improve corporate performance [Bénabou, 2013]. While the rationale behind individuals demanding motivated beliefs, thereby contributing to increased occurrences of self-fulfilling successes, appears evident, the dynamics of the production and supply sides entail a more nuanced consideration. Individuals encounter limitations in directly manipulating their beliefs induced by the responses from reality. As a result, individuals engage in motivated information processing, selectively accepting or rejecting information to arrive at conclusions that align with their preferences. This cognitive process may involve forgetfulness, false memory, and even memory illusion or delusion [Kunda, 1990; Pashler, 1998].

A robust body of research in psychology has consistently established the propensity of individuals to selectively concentrate on specific details, interpret information, and retain it in a manner that bolsters confidence in their capabilities [Dunning, 2001; Gilbert et al., 1998], describing a pattern of behaviour also

manifested throughout online information consumption. For instance, the research conducted by Mischel et al. [1976] discovers individuals to exhibit superior recollection of positive rather than negative information despite being exposed to an equitable amount of both positive and negative information concerning their personalities.

Given the vast and often unverified nature of online information, it becomes feasible to argue that individuals may exhibit biased memory retention, preferentially recalling information that aligns with their pre-existing beliefs, irrespective of its factual accuracy. This phenomenon is a manifestation of confirmation and belief biases which may be attributed to cognitive heuristics, prioritising and evaluating information in congruence with the preconceived notions. The underlying cognitive mechanism behind this well-evidenced inherent tendency to favour confirmatory information online, may be related to the findings of Eil and Rao [2011] in their examination of asymmetric information updating. In particular, Eil and Rao [2011] discover a tendency among individuals to prioritise favourable signals, employing Bayesian inference in their assessment, while tending to downplay or disregard negative signals. This highlights the relevance and applicability of Bayesian updating when processing online information which, in the instances of idealistically incongruent information, individuals may deem the cognitive effort required to reconcile this dissonance as too costly to execute, thereby constituting a plausible method to account for the observed phenomena of misinformation propagation.

In the same vein, Carrillo and Mariotti [2000], in conjunction with empirical validation provided by Brown et al. [2011], draw a connection between selective inattention and present bias. This relationship implies a tendency among indi-

viduals to suppress information that may erode their self-confidence in order to maintain their motivation. This behavioural pattern seems to similarly transcend into online contexts, indicating that individuals have an inclination to overlook information that may potentially weaken their self-assurance, thereby perpetuating their personal motivation [Briley and Wyer Jr, 2002; Hassoun et al., 2023].

Effectively, the conduct observed in online interactions is not confined to online environment; analogous tendencies have been extensively documented in psychological studies. For example, Fotopoulou et al. [2008] note instances where individuals exhibit false memory and a proclivity towards positive delusions. Furthermore, Howe and Derbish [2010] along Howe et al. [2011] propose that the fabrication of autobiographical memories holds an adaptive value by instigating an affirmative bias in personal history of oneself, thus contributing to self-enhancement. Additionally, individuals engaging in delusional behaviour may create fictitious evidence that aligns with their positive self-image. Noteworthy examples encompass personal adversities being ascribed to external conspiracies [Bortolotti, 2009] or individuals maintaining unfounded beliefs regarding the faithfulness of their partner as a coping measure aimed at protecting their self-esteem [Butler, 2000; McKay et al., 2005].

Hence, comprehending the underlying mechanisms underpinning false memory and delusion appears to hold a substantial relevance in contributing to the the modelling of decisions regarding online information. Providing a valuable point of reference, Chew et al. [2020] conducted theoretical and experimental analyses of the three types of positive memory errors, namely, positive amnesia (forgetting a negative event), positive delusion (creating a fictitious positive event), and positive confabulation (altering the memory of a negative event into a different

positive event). Specifically, Chew et al. [2020] investigate the associations of these memory errors with present bias, anticipatory emotions linked to self-image considerations, and attitudes towards risk and ambiguity.

Their conceptual framework extends the model presented by Bénabou and Tirole [2002] by introducing the prospect of delusion, which refers to the recollection of a positive signal when none has occurred. To put in perspective, Bénabou and Tirole [2002] established that imperfect recollection, combined with present bias, serves as a conduit for individuals to develop motivated beliefs and engage in intertemporal and intrapersonal management of their memories. Notably, this conceptualisation implies that self-delusion arises from the suppression of certain memories. In contrast, the perspective of Chew et al. [2020] diverges in that delusion and amnesia serve alternative motivational roles for individuals in their framework.

To test the implications of their model, Chew et al. [2020] conduct an experiment in Singapore with 701 subjects followed by a replication experiment conducted in Beijing with 445 subjects. The findings consistently demonstrate the systematic occurrence of three memory biases – delusion, amnesia, and confabulation – aligning with the equilibrium behaviour predicted by their model. The experiments revealed a systematic presence of false memory endorsing positive events and the phenomenon of positive amnesia, reflecting the tendency to overlook negative past events. The results also show a significant association between positive delusion and positive confabulation and the extent of present bias, albeit no such relationship was identified concerning positive amnesia. On the whole, their two-step model incorporating the potential confabulation, wherein a bad signal is forgotten and subsequently replaced by the creation of a positive

signal from no signal, contrary to the one-step model of positive amnesia, which transforms a negative signal into a positive signal, is found to be the model in line with their experimental findings. In essence, the presence of positive false memory rather than selective amnesia emerges as a reinforcer of self-esteem in equilibrium and thus contributes to explaining the observed outcomes in the experiments.

While a detailed discussion of the framework has been spared, substantiated by empirical findings, its demonstrated applicability and relevance in explaining biases in human behaviour and information assimilation indicates the significance the approach adopted by Chew et al. [2020] has for modelling online information consumption in light of the similarities in observed user behaviour. More specifically, their foundational model posits that individuals possess complete self-awareness, leading to an equilibrium characterised by a state of complete self-doubt. In line with Bénabou and Tirole [2002], they also postulate that individuals are cognisant of memory manipulation, however, underestimating its magnitude through imperfect Bayesian updating. Ultimately, this lends further support that integrating Bayesian Updating condition within the model of online information consumption stands as a reliable and robust approach. Individuals seemingly learn about online news and information in alignment with the foundational principles of Bayesian inference, indicating a self-awareness in decisions regarding information assimilation which may sometimes serve to preserve not only confidence, but self-identity often attested by a sense of belonging to their social and ideological circles.

Chapter 3

The Model in Action: Dissecting the Dynamics of Manipulative Information

3.1 Model Implementation

Having delineated the overarching framework of the model, it becomes imperative to critically examine its applications in light of particular misinformation scenarios. This analysis is not only instrumental but also fundamental to understanding the practical relevance and effectiveness of the model. Moreover, by exploring targeted scenarios, valuable insights may be gained into the operational dynamics, contributing to a more comprehensive understanding of the broader implications of the model in the context of misinformation, signifying its strengths and potential areas for refinement, paving the way for future inquiry.

To ensure continuity with the theoretical model developed in Chapter 2,

the following applications are interpreted through the same cost–benefit and expected-utility logic, where verification effort, information precision, and exposure to manipulation jointly determine welfare. Each context is therefore treated as a targeted illustration rather than an exhaustive account of the domain.

Although the potential for misleading information frequently arises from inadequate verification protocols, substandard lapses in journalism, and inadvertent errors, the dissemination of misinformation on digital platforms may be driven by a range of diverse factors. These may include the desire to shock or captivate audiences, to provoke emotional responses, to bolster user engagement and subsequent interaction, attract readership or enhance user retention, but also to advance and gain a following for ideological, political, commercial or other agendas. Recognising the underlying intentions of information production and circulation is vital for constructing a robust framework, as it enables the capture of the mechanisms behind the malevolent motivations, facilitating a deeper understanding of the intricacies of misinformation in the contemporary digital landscape.

At this juncture, it is imperative to make a terminological distinction concerning the terms utilised across the diverse contexts of misleading information. While misinformation serves as an umbrella term for various forms of falsehoods, irrespective of intent, the deliberate propagation of such information with the intent to deceive is more precisely termed disinformation. As a specific branch of misinformation, disinformation is characterised by the calculated and strategic propagation of false or misleading content, designed to manipulate perceptions, influence public discourse, and further specific political, ideological, or commercial interests. It serves as a tool for malevolent actors, aimed at exploiting informa-

tional asymmetries and psychological biases to achieve premeditated objectives.

In the political arena, disinformation campaigns have become a potent tool for election interference. Foreign actors frequently deploy disinformation to sow discord, undermine trust in democratic institutions, and influence voter behaviour. These efforts are designed not only to distort the electoral process but also to exacerbate partisan polarisation by reinforcing existing biases and further deepening political divisions.

Health-related disinformation is another critical domain where malevolent intent can have far-reaching consequences. Pseudoscientific claims and medical hoaxes, particularly around vaccines and health crises, can erode public trust in legitimate health authorities, fostering confusion and dangerous public health outcomes. Conspiracy theories often accompany such disinformation, targeting vulnerable populations and undermining confidence in established sources of knowledge. Similarly, extremist groups exploit disinformation to recruit followers and propagate their ideologies, spreading harmful narratives that support divisive "deep state" beliefs.

Moreover, the proliferation of fake news, clickbait, and sensationalised content is amplified by digital platforms. Advanced technologies, such as deepfakes and manipulated media, further complicate efforts to distinguish truth from falsehood. These tools enable disinformation campaigns to create content that appears credible, making it increasingly difficult for users to identify malicious intent. Thus, distinguishing between misinformation and disinformation is essential for developing an effective framework that captures the complex decision-making processes concerning the effort exerted by users to dispel uncertainty about the veracity of information, particularly in the face of increasingly sophisticated technological

manipulations that may overwhelm users beyond their cognitive capacities and complicate the already costly process of information verification.

Effectively, to delve into real life misinformation scenarios, a further review of the original framework is warranted. As in the initial formulation of the model, users interacting with online information may perform effortful assessments to mitigate the uncertainty surrounding the knowledge presented in digital content. To this end, Bayesian updating resumes as the pivotal mechanism through which users assimilate new information into their existing beliefs as new data emerges. Essentially, it provides a structured approach to capturing the revision and recalibration of individual understanding and interpretation of online information, integrating prior knowledge with newly obtained insights and thereby well encapsulates how users contend traverse uncertainty and misinformation online. As before, the posterior probability $\hat{\rho}(\rho, x)$ is revised applying Bayesian updating contingent on the chosen level of effort x:

$$\hat{\rho} = \frac{\rho(1-x)}{1-\rho x} \tag{3.1}$$

However, as misinformation may be deliberately disseminated with malevolent intents, whether the user encounters it initially or during the subsequent informational revision process, the posterior probability of exposure to such content may be affected irrespective of the efforts exerted by the individual user. Accordingly, malevolent actors may actively seek to mislead (unaware) users by feeding them false content, although such users may be anticipating such manipulation when deciding on the optimal level of effort x. This suggests that, in their attempts to deceive online users, these actors must also exert some level of effort s and incur

the associated transaction costs $C(s)$. These costs $C(s)$ may take various forms, including financial expenditures, cognitive burdens, and other resource commitments such as time. For instance, financial costs may encompass the monetary resources invested by these actors to spread falsehoods, cognitive costs may pertain to the intellectual labour required to design, refine, and distribute deceptive content, whereas time costs may represent the hours or days spent curating and propagating misleading information.

To account for the impact of the actions and efforts by malevolent actors on user beliefs and perceptions, particularly through the dissemination of misleading information, the posterior probability of being influenced by such misinformation is first expressed in a simplified nested form that highlights its dependence on the baseline posterior $\hat{\rho}$:

$$\hat{\rho}_s = \frac{\hat{\rho}(1-s)}{1-\hat{\rho}s}, \tag{3.2}$$

where $\hat{\rho}$ is given by equation 3.1, and $s \in [-1, 0]$ represents the degree of manipulation or bias introduced into the informational environment. This formulation shows that $\hat{\rho}_s$ is a nested function of $\hat{\rho}$, implying that manipulation operates by distorting the probability of inaccuracy perceived by the user. Conceptually, $\hat{\rho}_s(\hat{\rho}) \to \hat{\rho}$ as $s \to 0$, while larger $|s|$ values represent stronger manipulation and a greater divergence between the perceived and the true posterior probability.

The expanded version, which incorporates both user verification effort x and manipulative effort s, is given by

$$\hat{\rho}_s = \frac{\hat{\rho}(1-x)(1-s)}{\Big(1-\hat{\rho}x\Big)\Big(1-\dfrac{(1-x)s\hat{\rho}}{1-x\hat{\rho}}\Big)}. \tag{3.3}$$

In this updated formulation 3.3 of the posterior probability $\hat{\rho}_s$, Figure 3.1 graphically depicts how the effort s exerted by a malevolent actor destabilises the Bayesian updating process associated with user effort x by introducing noise into their beliefs as gauged by $\hat{\rho}_s$ through the spread of false information. Consequently, s serves as a negating factor that undermines the reduction of uncertainty regarding user knowledge. To delineate this relationship, s is confined to the range $[-1, 0]$, where, as s approaches -1, the posterior probability of exposure to misinformation increases. This structure formalises how manipulative signals degrade learning efficiency by embedding bias into the belief-updating process itself.

Conversely, it is also conceivable that some actors may aid users in dispelling misinformation when s resides within the interval $[0, 1]$, such that as s approaches 1, uncertainty diminishes and the accuracy of user beliefs improves. However, this aspect falls outside the scope of the present analysis.

While in this model setting, the malevolent actors may regress the posterior probability of the user experiencing social loss due to misinformation for any intensity of the effort exerted, consistent with the original framework, the user still incurs transaction costs induced by their choice of cognitive effort. To recapitulate, the capacity for cognitive effort is inherently limited and costly for each piece of online content consumed. Essentially, the encountering of substantial volumes of online information, coupled with individual factors, such as motivation, personal interest, cognitive ability, prior knowledge, and familiarity, may significantly modulate the level of effort and thus the resulting cognitive costs allocated

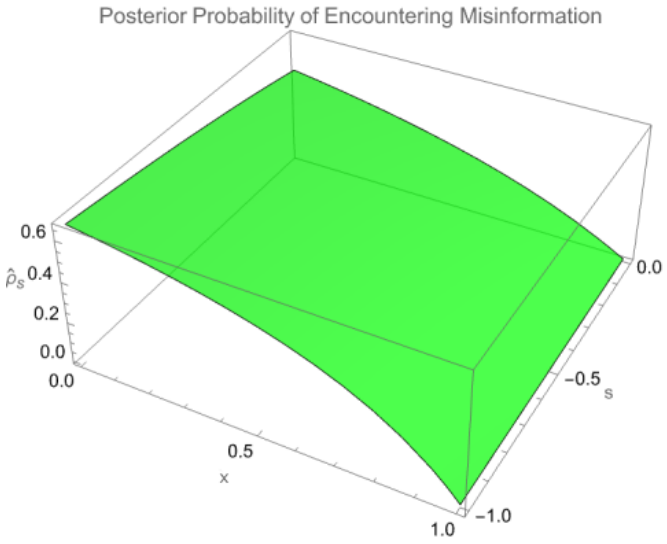

Figure 3.1: Posterior Probability of Encountering Misinformation

to analysing each piece of content. For instance, higher motivation to comprehend a specific subject may prompt users to utilise greater cognitive resources, while individuals with superior abilities or knowledge may process information more efficiently at lower cognitive costs.

Ultimately, despite that users are generally predisposed to seek clarification and certainty, their efforts to acquire knowledge about online content are typically constrained by cognitive costs which may also be further exacerbated by the presence of malevolent actors, requiring users to expend increasingly higher cognitive effort and incur even greater cognitive costs to achieve the same level of certainty as would be attainable in the absence of such interference. Deviating from the broader framework for the sake of notational simplicity and traceability, the transaction costs associated with user effort x are modelled as

$$C(x) = \frac{x}{1 - x^{1/2}}, \tag{3.4}$$

thereby simplifying the analysis without compromising accuracy.

While the influence of the malevolent actor is acknowledged, the analysis retains a user-centric perspective, focusing on the choice of effort by the user, while the actions of the malevolent actor, along with their associated costs are treated as arbitrary within the model. More specifically, the decisions of the malevolent actor are considered exogenous rather than endogenous. That is, the model does not examine how the malevolent actor weighs the benefits of deception against the incurred costs. For the purposes of this study, it is posited that the malevolent actor derives some form of satisfaction from their deceptive activities when exerting effort s, however, this dynamic is not elaborated upon within the present framework.

In another deliberate departure, the payoff structure associated with online information consumption delineated in the general framework is also simplified to enhance the notational traceability of the analysis, expressing the (net) benefits of being informed as $V - C(x) - \hat{\rho}_s L > 0$. Here, V denotes the net benefits derived from being informed, while $C(x)$ represents the transaction costs incurred by the user. The term $\hat{\rho}_s$ reflects the posterior probability of experiencing losses due to exposure to potential misinformation, with L representing the losses incurred from such exposure. Within this simplified framework, the condition for maximisation is achieved when the optimal level of effort, $\hat{b}$, equates the marginal cost of cognitive effort with the marginal benefit of mitigating expected losses from misinformation, formalised by:

$$C'(x) = \hat{\rho}_s L \tag{3.5}$$

The relationship in the equation 3.5 encapsulates the circumstances under which $V \geq 0$, indicating that users may derive net benefits from their information-seeking behaviours even in environments where misinformation is prevalent.

The presence of manipulation alters the marginal-benefit condition by effectively lowering the perceived value of information accuracy. As a result, users facing higher $|s|$ require greater expected utility gains to justify additional verification effort, implying that optimal search intensity declines when misinformation is pervasive or cognitively taxing to detect.

Given the subjective nature of individual perceptions regarding the benefits and costs associated with any given item of online information, the expected utility framework is utilised to account for varying preferences and valuations shaped by the content encountered. Specifically, a Constant Relative Risk Aversion (CRRA) utility function is employed to capture these idiosyncratic preferences. To this effect in this framework, the expected utility is modelled as a function of both the potential utility from being informed and the potential losses from misinformation, while also factoring in the transaction costs associated with exerting effort:

$$E[U] = (1 - \hat{\rho}_s(x, s))U[V - C(x)] + \hat{\rho}_s(b, s)U[V - C(x) - L] \tag{3.6}$$

where $1 - \hat{\rho}_s$ represents the probability of avoiding misinformation and $\hat{\rho}_s$ is the probability of suffering misinformation losses.

In order to examine how the expected utility changes with respect to changes in cognitive effort, the expected utility is differentiated with respect to effort x:

$$\frac{dE[U(x)]}{dx} = (1 - \hat{\rho}_s(x,s))\frac{dU[V - C(x)]}{dx} + \hat{\rho}_s(x,s)\frac{dU[V - C(x) - L]}{dx} \tag{3.7}$$

By setting 3.7 to 0, the optimal effort level b which maximises the expected utility can be determined.

Substituting back into the equation 3.6 can be expanded into:

$$E[U(x)] = \frac{(1-x)(1-s)\left(-\frac{x}{1-\sqrt{x}} - L + V\right)^{1-G}\hat{\rho}}{(1-G)(1-x\hat{\rho})\left(1 - \frac{(1-x)s\hat{\rho}}{1-x\hat{\rho}}\right)} + \frac{\left(-\frac{x}{1-\sqrt{x}} + V\right)^{1-G}\left(1 - \frac{(1-x)(1-s)\hat{\rho}}{(1-x\hat{\rho})\left(1-\frac{(1-x)s\hat{\rho}}{1-x\hat{\rho}}\right)}\right)}{1-G} \tag{3.8}$$

Differentiating the Expected Utility function 3.8 with respect to x yields:

$$
\begin{aligned}
\frac{dE[U(x)]}{dx} = & \frac{(1-x)\hat{\rho}^2(1-s)\left(-\frac{x}{1-\sqrt{x}}-L+V\right)^{1-G}}{(1-G)(1-x\hat{\rho})^2\left(1-\frac{(1-x)\hat{\rho}s}{1-x\hat{\rho}}\right)} \\
& - \frac{(1-x)\hat{\rho}(1-s)\left(\frac{\hat{\rho}s}{1-x\hat{\rho}}-\frac{(1-x)\hat{\rho}^2 s}{(1-x\hat{\rho})^2}\right)\left(-\frac{x}{1-\sqrt{x}}-L+V\right)^{1-G}}{(1-G)(1-x\hat{\rho})\left(1-\frac{(1-x)\hat{\rho}s}{1-x\hat{\rho}}\right)^2} \\
& - \frac{\hat{\rho}(1-s)\left(-\frac{x}{1-\sqrt{x}}-L+V\right)^{1-G}}{(1-G)(1-x\hat{\rho})\left(1-\frac{(1-x)\hat{\rho}s}{1-x\hat{\rho}}\right)} \\
& + \frac{(1-x)\left(-\frac{\sqrt{x}}{2(1-\sqrt{x})^2}-\frac{1}{1-\sqrt{x}}\right)\hat{\rho}(1-s)\left(-\frac{x}{1-\sqrt{x}}-L+V\right)^{-G}}{(1-x\hat{\rho})\left(1-\frac{(1-x)\hat{\rho}s}{1-x\hat{\rho}}\right)} \\
& + \frac{\left(V-\frac{x}{1-\sqrt{x}}\right)^{1-G}\left(-\frac{(1-x)\hat{\rho}^2(1-s)}{(1-x\hat{\rho})^2\left(1-\frac{(1-x)\hat{\rho}s}{1-x\hat{\rho}}\right)}\right)}{1-G} \\
& + \frac{\left(V-\frac{x}{1-\sqrt{x}}\right)^{1-G}\left(\frac{(1-x)\hat{\rho}(1-s)\left(\frac{\hat{\rho}s}{1-x\hat{\rho}}-\frac{(1-x)\hat{\rho}^2 s}{(1-x\hat{\rho})^2}\right)}{(1-x\hat{\rho})\left(1-\frac{(1-x)\hat{\rho}s}{1-x\hat{\rho}}\right)^2}\right)}{1-G} \\
& + \frac{\left(V-\frac{x}{1-\sqrt{x}}\right)^{1-G}\left(\frac{\hat{\rho}(1-s)}{(1-x\hat{\rho})\left(1-\frac{(1-x)\hat{\rho}s}{1-x\hat{\rho}}\right)}\right)}{1-G} \\
& + \left(-\frac{\sqrt{x}}{2(1-\sqrt{x})^2}-\frac{1}{1-\sqrt{x}}\right)\left(V-\frac{x}{1-\sqrt{x}}\right)^{-G} \\
& \times \left(1-\frac{(1-x)\hat{\rho}(1-s)}{(1-x\hat{\rho})\left(1-\frac{(1-x)\hat{\rho}s}{1-x\hat{\rho}}\right)}\right) \qquad (3.9)
\end{aligned}
$$

3.2 Disinformation Dynamics in Political Contexts

Having elucidated the model involving malevolent intents and actors, to operationalise and contextualise the model, the focus is cast on simulating a specific scenario in which online users encounter disinformation. It is essential to emphasise that while the dynamics of the model remain analogous across various forms of disinformation, including health misinformation, conspiracy theories, and the dissemination of misleading narratives through fabricated news and sensationalist media, the application of the model is examined within the context of political disinformation. Relative to the general framework in Section 3.1, the political-disinformation setting maintains the same optimisation structure but reinterprets the cognitive cost $C(x)$ as the mental effort required to evaluate the credibility of politically charged narratives. The information gathered represents cues about source reliability, ideological bias, and factual consistency that update the posterior user beliefs about content accuracy.

The emphasis is placed on the political context due to the significant global prevalence of disinformation campaigns that seek to influence public perception, electoral outcomes, and other democratic processes. More specifically, political disinformation may encompass a diverse array of strategies that are designed to manipulate public opinion, erode trust in institutions, and shape societal discourse on critical issues. A salient example of this phenomenon is election interference, wherein foreign actors frequently target democratic processes with disinformation strategies designed to sow discord, undermine trust in institutions, and manipulate voter behaviour. This was starkly evidenced during the 2016

U.S. presidential election, where Russian operatives employed social media platforms to disseminate divisive content and false information, profoundly impacting public sentiment and electoral dynamics.

In a similar vein, comparable tactics have been observed in more recent elections, such as the 2020 U.S. presidential election, where misinformation regarding mail-in voting and voter fraud was rampant, resulting in widespread confusion and distrust among voters. Beyond the United States, disinformation campaigns have also been documented in Brazil during the 2018 presidential election, where false narratives circulated about candidates and electoral processes, and in the United Kingdom, particularly during the Brexit referendum, where misleading information influenced public opinion on critical national issues.

By the same token, the ongoing Ukraine war has emerged as a significant focal point in global geopolitical tensions, prompting various state and non-state actors to engage in information warfare. In this context, Russian state-sponsored campaigns have been particularly pronounced, with allegations of disinformation surrounding the conflict, including false narratives about Ukrainian atrocities, NATO expansionism, and the origins of the war. Organisations such as the Atlantic Council, the Digital Forensic Research Lab, and the East Stratcom Task Force have documented these campaigns, highlighting the strategic use of disinformation to shape perceptions and narratives related to the ongoing conflict.

In light of these examples, it becomes evident that the ramifications of political disinformation are far-reaching, necessitating a robust understanding of its mechanisms and effects within the context of user engagement in digital environments, and particularly their information consumption-related decision making. In the political scenario, effort influences exposure through selective engagement and

verification intensity. Higher effort reduces susceptibility to deceptive partisan cues, whereas lower effort increases reliance on heuristics and amplifies the influence of manipulative content. Thus, disinformation intensity interacts inversely with verification effort in shaping user outcomes.

To commence the analysis of the scenario-specific application of the model, users engage with social media platforms to interact with and obtain information. For the purposes of this model, a user whose interests include current affairs, particularly political events in their country is considered. As part of their information-seeking behaviour, this user comes across an online piece of content pertaining to a significant political event. It is assumed that the gains V from obtaining accurate information regarding this event amount to 20 for the user, while the potential losses L due to this content being disinformation are set at 10. Subsequently, the user has to decide on the optimal level of effort x to exert in order to dispel uncertainty regarding the truthfulness of this content. Although the user may stand to benefit from this effort, it also leads them to incurring a corresponding cognitive cost $C(x)$ as defined in 3.4. Initially, the probability ρ that this information is false is set at 0.5, reflecting a state of uncertainty regarding its veracity.

Unbeknownst to the user, it is posited that the encountered information is disseminated with malevolent intent by an actor whose effort s is quantified at -0.6. Consequently, the Bayesian learning process of the user is disrupted, resulting in the posterior probability $\hat{\rho}_s$:

$$\hat{\rho}_s = \frac{\hat{\rho}(1-x)(1-s)}{\Big(1-\hat{\rho}x\Big)\Big(1-\dfrac{(1-x)s\hat{\rho}}{1-x\hat{\rho}}\Big)}. \tag{3.10}$$

Assuming that the CRRA utility function 2.7 adequately represents the preferences of the user, the coefficient G delineates the degree of risk aversion exhibited by the individual. Within the context of information consumption, G captures the propensity of the user to expend costly effort in verifying the accuracy of information encountered online, reflecting their preference for certainty over misinformation. Although G operates on the utility derived from information accuracy rather than direct monetary pay-offs, the object of risk aversion remains the overall expected utility of outcomes. In this setting, information accuracy functions as a proxy for the quality of pay-offs, so risk-averse individuals experience greater disutility from uncertainty in credibility and therefore adjust their effort decisions accordingly. A higher G value signifies a stronger preference for certainty regarding the correctness of information, as more risk-averse individuals demonstrate a lower tolerance for ambiguity. Table 3.1 provides a comprehensive overview of the various ranges of G values and their corresponding levels of risk aversion. In this particular scenario, the user is characterised as moderately risk averse, with G assigned a value of 1.7

Table 3.1: Ranges of Risk Aversion

Category	**Coefficient** G
Low Risk Aversion (Risk-Seeking)	$G < 0$
Neutral Risk Aversion	$G = 0$
Moderate Risk Aversion	$0 < G < 2$
High Risk Aversion	$2 < G < 4$
Very High Risk Aversion	$G \geq 4$

With all pertinent parameters established and the model fully parameterised, the expected utility function 3.6 within the current context of political disinformation can now be expressed and differentiated with respect to x. By setting the resulting derivative equal to zero, the optimal value of x can be derived, yielding the optimal level of effort the user is inclined to exert to mitigate their exposure to disinformation.

The framework developed here can be extended by endogenising the manipulator choice of s, thereby enabling the joint optimisation of user and adversarial strategies within a unified strategic setting. Such an extension would permit a comparative statics analysis of equilibrium effort levels under varying misinformation intensities, signal structures, and cognitive cost parameters. This would provide a more comprehensive characterisation of the strategic interaction between information producers and consumers, yielding deeper insights into the mechanisms through which policy interventions or platform design choices may influence the equilibrium dynamics of misinformation and verification effort.

Chapter 4

Validation of the Model Using Semi-Structured Interviews

Departing from the conventional economic assumption of perfect information and fully rational, optimally informed decision-making, real-world environments are inherently characterised by risk and uncertainty. Despite an unparalleled degree of global interconnectedness and the continuous evolution of technologies that have fundamentally reshaped nearly all aspects of human existence, the instantaneous accessibility of vast information repositories has not necessarily alleviated the cognitive burdens associated with decision-making. Instead of simplification, the expansion of digital information ecosystems has introduced greater complexity into the information consumption process. While the internet has empowered individuals not only to access news and knowledge in real time but also to generate, publish, and disseminate content in a variety of formats, including blog entries, social media posts, videos, podcasts, and forums, this expansion and decentralisation of the information space has significantly altered the mechanisms through

which credibility is assessed. The proliferation of digital communication channels has not only increased the sheer volume of information but has also diversified the mechanisms through which it is framed, interpreted, and disseminated, further complicating the distinction between credible sources and misleading content.

Having facilitated the transformation of human interactions, not only have the rapid technological advancements amplified the volume of information available but also accelerated the speed at which content is produced and disseminated. Historically, editorial oversight, peer review, and institutional gatekeeping constituted the means by which accuracy before information reached the public domain was ensured. However, the velocity with which information is generated, recombined, and circulated has significantly outpaced these traditional mechanisms of verification, often creating an environment in which false or misleading narratives can spread rapidly before they can be adequately scrutinised. Consequently, individuals are confronted with the dual challenge of sifting through an overwhelming influx of information while simultaneously contending with cybersecurity threats, algorithmic content curation, and the rapidly shifting digital landscape. In such a cognitively demanding and information-saturated environment, the prevalence of misinformation and disinformation introduces an additional layer of complexity, requiring that individuals exercise greater discernment in evaluating informational reliability. As information ecosystems become progressively intricate, decision-makers must navigate an ever-expanding web of truth and falsehood while seeking to deploy the informational resources at their disposal in a manner conducive to effective decision-making. However, the asymmetry between the vast quantities of available information and the cognitive and temporal constraints imposed on individuals creates vulnerabilities that may be strategically

exploited by malign actors seeking to manipulate perceptions and influence public discourse.

The immediate and unrestricted accessibility of information has thus yielded not only substantial and multifaceted benefits but also engendered a landscape fraught with challenges associated with misleading or deceptive content. While the internet functions as an expansive platform for knowledge exchange, it has also transferred the responsibility of evaluating the credibility and quality of information onto individual users, necessitating the application of critical reasoning skills to effectively navigate the digital sphere. However, the burden of verification is unevenly distributed, as individuals differ in their ability, willingness, and resources to engage in rigorous fact-checking. Given the sheer scale of information available, combined with the highly variable degrees of personal relevance, users have to grapple with constraints stemming from both cognitive limitations and the associated costs of information verification. The reliance on algorithmically curated content further exacerbates this issues by structuring exposure to information based on engagement patterns rather than informational accuracy, reinforcing biases and shaping the dynamics of digital discourse. As a result, decision-making in digital contexts often relies on pragmatic rather than methodologically rigorous heuristics, leading individuals to adopt practical yet imperfect strategies for assessing the veracity of encountered information. These mechanisms, while efficient in navigating large volumes of data, may inadvertently reinforce cognitive biases and selective exposure, affecting not only individual beliefs but also broader societal discourse.

Against this backdrop, Chapter 1 conceptualised the intricate cognitive processes underlying information consumption and formulated a foundational frame-

work for understanding the optimisation of cognition in digital environments. This framework was developed in the broader context of misinformation, providing a generalised structure for examining the cognitive and behavioural dimensions of information processing. Misinformation serves as an overarching term encompassing various forms of false or inaccurate information, irrespective of intent, often emerging as a byproduct of deficiencies in confirmation protocols, lapses in journalistic standards, or inadvertent human error. While misinformation can, at times, exert significant influence, particularly when inaccuracies persist uncorrected or when they gain traction in high-stakes domains, such as public health or financial decision-making, its stochastic and uncoordinated nature generally constrains its capacity to systematically shape public perceptions. Unlike disinformation, which is deliberately crafted to mislead, misinformation arises in a dispersed and unstructured manner, meaning that its impact, although potentially substantial for particular individuals or groups, does not follow a strategic or sustained trajectory. Nevertheless, its cumulative effect across digital environments may contribute to broader epistemic uncertainty, complicating the ability of information consumers to distinguish between credible and misleading content, thereby increasing the cognitive burden associated with information assessment.

In contrast, disinformation represents a distinct and more insidious category of false information, distinguished by its deliberate and strategic propagation with the explicit intent to manipulate perceptions, influence public discourse, and serve political, ideological, or commercial interests. Unlike misinformation, which emerges as a consequence of flawed informational processes, disinformation is systematically produced and disseminated to exploit informational asymmetries, psychological biases, and cognitive limitations, with the objective of

actively disrupting belief formation. The distinction lies not only in intent but also in execution since disinformation is characterised by its calculated and often methodically orchestrated deployment, designed not merely to mislead but to shape perceptions, construct narratives, and restructure decision environments in ways that yield tangible strategic advantages. Given its far-reaching implications, disinformation campaigns tend to be concentrated in contexts where influencing public opinion has significant consequences, namely, political elections, public health crises, and geopolitical conflicts, whereas misinformation, despite its potential to mislead, lacks the same degree of systemic intent and coordination.

Among these contexts, conflict environments represent a particularly salient case in which disinformation is not merely incidental but constitutes a core component of broader strategic operations. Unlike other domains where disinformation campaigns may be episodic or event-driven, warfare necessitates their continuous and deliberate deployment as an extension of military, psychological, and political tactics. The manipulation of informational flows in such environments is central to shaping battlefield realities, influencing civilian sentiments, and destabilising adversarial decision-making structures. Unlike transient misinformation, which may occasionally affect public discourse, disinformation in conflict settings assumes a persistent and adaptive form, responding dynamically to shifts in military, political, and diplomatic conditions. This sustained and strategically motivated distortion extends beyond immediate tactical objectives, contributing to long-term consequences such as the erosion of institutional trust, the reshaping of collective memory, and the reinforcement of polarising narratives. Given these dynamics, Chapter 2 extended the foundational framework developed in the preceding chapter by introducing an adversarial component to account for

the presence of a strategic actor exerting effort to deliberately obstruct verification processes. By incorporating this dimension, the framework encapsulates the destabilising effects of disinformation on the learning process and, consequently, belief formation. This, in turn, enables a more nuanced analysis of the cognitive processes governing the efforts individuals exert to mitigate uncertainty regarding the credibility of information, with the model capturing the intricacies of environments where manipulative strategies are designed to exploit and overwhelm cognitive capacity, thereby exacerbating the already costly process of information verification.

With the theoretical foundations of the general misinformation model and its disinformation-focused counterpart having been established in Chapters 1 and 2 respectively, the present chapter advances the analysis through a rigorous empirical calibration of the model. This process extends beyond simple parameter adjustment, encompassing a comprehensive evaluation and refinement of the model assumptions to not only ensure their congruence with real-world observations but also to deepen the understanding of behavioural dynamics through the lens of the model. The ongoing conflict in Ukraine provides a pertinent and illustrative setting for this empirical investigation, offering not only a vivid example of the pervasive nature of misinformation resulting from routine informational mismanagement but also serving as a salient case study for the deliberate deployment of disinformation as a strategic instrument in warfare. Given the inherently qualitative nature of the available data, which precludes traditional quantitative point-estimate analyses, semi-structured interviews are employed to establish parameter bounds, offering insights into the variable ranges within which key model inputs are situated. Despite the data constraints, this methodological approach

effectively integrates both quantitative and qualitative elements, facilitating a nuanced exploration of the cognitive and strategic factors at play in the context of disinformation. This approach allows for a refined adjustment of the model to reflect the empirical realities of information verification within adversarial settings, ensuring that the model remains not only theoretically robust but also empirically grounded and practically applicable. By incorporating qualitative insights, it becomes possible to account for factors that quantitative data alone may overlook, with the iterative calibration enabling the model to more accurately reflect the complexities of decision-making under information uncertainty, particularly in environments where strategic disinformation significantly influences cognitive processes and information verification efforts.

To this end, the current chapter begins by reviewing and subsequently outlining the methodological framework, integrating quantitative and qualitative approaches, that underpins the empirical analysis, before proceeding to a thematic examination of the qualitative data. The analysis then delineates plausible parameter ranges and contextual factors that shape model dynamics, ultimately feeding into the computational assessment of the trade-offs between the benefits and costs associated with information engagement. Specifically, the empirical insights are integrated into the model to assess the extent to which its theoretical assumptions hold when juxtaposed with the lived experiences of individuals navigating a disinformation-laden environment. Throughout this process, the parameter estimates are determined, concurrently testing the robustness and validity of the overall model. Given the computational complexity inherent in solving the model, graphical analysis is employed to visualise the behavioural patterns emerging from the parameter configurations. By systematically analysing the

semi structured interview data, in conjunction with defining parameter bounds and validating the trade-offs identified within decision-making, the empirical calibration strengthens the capacity of the model to capture the interdependencies between cognitive effort, information verification costs, and susceptibility to manipulation. Essentially, through the synthesis of theoretical constructs with empirical insights, it is ensured that the model remains rigorously validated while accommodating the multifaceted realities of information consumption in digital environments. In this manner, the recursive analysis process establishes a robust analytical foundation, enhancing the explanatory and predictive power of the proposed framework in understanding the behavioural intricacies of information consumption.

At the same time, the interview material also reveals behaviours that are only partially captured by the present model, such as collective verification practices, emotional fatigue, and the role of social norms in sustaining or discouraging effort. These features are documented explicitly in the analysis to delineate the limits of the current individual-level framework and to indicate directions in which future extensions could relax the assumptions of static choice and purely individual optimisation.

4.1 Model Validation Strategies

To validate the trade-offs between cognition and information accuracy posited by the theoretical model of online information consumption presented in previous chapters, it is essential to ground the theoretical framework in real-world data. While quantitative data is often preferred for such validation, its availability may

be limited, particularly in complex cognitive contexts. In these circumstances, adopting a mixed-methods approach that integrates both quantitative and qualitative methodologies is indispensable. Not only does this approach enable analysis under data constraints, but it also facilitates a deeper and more comprehensive investigation of the research questions, yielding insights that would otherwise remain inaccessible in the absence of such methodological integration.

In fact, a substantial body of literature demonstrates the value of integrating qualitative and quantitative methodologies to enhance the explanatory depth of theoretical frameworks, particularly in situations when data constraints limit the applicability of purely quantitative approaches. Although these studies may not always explicitly identify as mixed-methods research, they exemplify the principles articulated by Yin [2011], who advocates for the iterative synthesis of qualitative and quantitative methodologies. Examples of such methodological integration include de Gramatica et al. [2017], which validates a quantitative economic model of agency costs in aviation security through qualitative insights derived from interviews with key stakeholders, such as airport managers and security personnel. Likewise, De Gramatica et al. [2015] enriches a cybersecurity policy-economic model with qualitative data from interviews, exploring the unintended regulatory impacts on smaller airports. While not grounded in quantitative models, studies such as Elliott et al. [2019] and Haugstvedt and Tuastad [2023] employ qualitative data to inform conceptual frameworks, illustrating how these methods can deepen and contextualise theoretical constructs. Specifically, Elliott et al. [2019] uses qualitative insights to refine a conceptual framework for organisational knowledge protection strategies, leveraging interviews with industry experts to better understand the complexities surrounding intellectual property in techno-

logical firms, while Haugstvedt and Tuastad [2023] employs qualitative data from interviews with counter-terrorism experts to explore the unintended consequences of multi-agency collaborations in addressing violent extremism, offering critical insights into the interactions and challenges of cross-agency cooperation. These studies, alongside the methodological principles advanced by Yin [2011], will be comprehensively reviewed in subsequent sections, forming a robust foundation for calibrating and validating the theoretical model of online information consumption, particularly in the context of cognitive trade-offs involved in disinformation verification.

In methodological terms, this chapter most closely follows the logic exemplified by De Gramatica et al. [2015]; de Gramatica et al. [2017], in which qualitative material is used iteratively to refine parameter bounds and to interrogate the plausibility of key mechanisms in a formal model, rather than to generate an entirely inductive grounded theory. The work of Elliott et al. [2019] and related studies is drawn upon primarily as an illustration of how interview-based insights can sharpen conceptual categories such as protection strategies and cognitive constraints, while Yin [2011] provides the overarching case-study framework that structures the integration of qualitative and quantitative evidence. Within this mixed-method design, the present chapter uses the interviews both to benchmark the behavioural patterns implied by the model and to highlight residual dynamics, such as emotional fatigue and institutional trust, that motivate the discussion of limitations.

Following from this premise, Yin [2011] provides a substantive discussion on the integration of qualitative and quantitative approaches within a single study, emphasising the critical importance of methodological rigour in mixed-methods

research. Many studies across diverse contexts, extending beyond the analysis of online phenomena, are guided by this framework to blend qualitative and quantitative approaches within their research design and methodological decisions such as Alvarado-Alvarez et al. [2021], which examines the interplay between shared vision, trust, and conflict resolution in family businesses to enhance constructive conflict management. Similarly, other studies have applied mixed-methods approaches in domains such as psychology, for instance, exploring the predictive role of perfectionism, anxiety, and procrastination on academic achievement [Yurtseven and Akpur, 2018], and in in healthcare, investigating factors influencing the use of mobile health interventions for managing chronic obstructive pulmonary disease [Alwashmi et al., 2020].

However, to put into perspective, while acknowledging that there is no universally applicable methodology, Yin [2011] dedicates substantial attention to the foundational elements of qualitative research, which are central to shaping the overall structure of the study and refining research questions. These are complemented by a set of principles and procedures regarding a logical, rather than merely logistical, research design, systematic data collection and analysis, all grounded in a reflective and adaptive methodological approach. Exemplified in Lin et al. [2023], the study employs these foundational guidelines to conceptualised the investigation into the function that artificial intelligence plays in advancing sustainable education. Conducting sound qualitative research, therefore, requires thoughtful decisions pertaining to the selection of appropriate methods, especially the implementation of coherent data collection strategies. These considerations are essential for developing reliable data collection instruments and ensuring that qualitative data is recorded, analysed and presented with precision

and integrity.

In tandem with comprehensive qualitative research procedures, which encompass the meticulous and evolving process of study design and instrument planning and construction, Yin [2011] sets forth the methodological principles for combining not only varied types of data but also the design and technical components of quantitative and qualitative research. These principles ensure the seamless integration of qualitative and quantitative elements, contributing to a cohesive and unified research framework. For instance, Özdemir et al. [2020] demonstrates the application of this such a consolidated approach in investigating the influence of social intelligence on leadership behaviours, illustrating the power of a mixed-methods strategy in generating a more nuanced and multidimensional understanding. As such, the laid-out principles have significant potential for enhancing the richness and depth of the findings in studies of cognitive behaviours. Viewed in this manner, mixed-methods research is not merely an additive process, but requires a profound understanding of how the two types of research may synergise and enhance each other, with the flexibility to adapt as new insights emerge.

Aligned with comprehensive qualitative research protocols, which encompass the meticulous and iterative process of study design and instrument development, Yin [2011] outlines the fundamental methodological principles for synthesising not only varied types of data but also the design and technical aspects of both quantitative and qualitative research. These principles facilitate the seamless integration of qualitative and quantitative components, contributing to the establishment of a cohesive and robust research framework. For instance, Özdemir et al. [2020] demonstrates the application of this integrated approach in investigating the in-

fluence of social intelligence on leadership behaviours, illustrating the power of a mixed-methods strategy in generating a more nuanced and multidimensional understanding. This perspective underscores that mixed-methods research is not a mere aggregation of disparate approaches but rather a sophisticated synthesis, requiring an in-depth understanding of how both paradigms can complement and augment one another, with the adaptability to refine the research trajectory as new insights emerge.

Although the focus is primarily cast on qualitative research, Yin [2011] stresses the importance of integrating qualitative insights into the overall research process, particularly in the development and refinement of quantitative models. For instance, the qualitative data collection methods, such as interviews, observations, and document analysis, are discussed extensively by Yin [2011] as instruments that may be constructed and designed to gather rich, narrative data that may offer the contextual depth often missing in purely quantitative approaches [Vaismoradi et al., 2016]. To illustrate this point, Lin and Mattila [2021], Dai et al. [2024], Yurtseven and Dulay [2022], Clifford Astbury et al. [2024] and McNamara et al. [2022] are all exemplars, among numerous others, of studies where interviews or other forms of qualitative data have been leveraged to augment the robustness of quantitative analyses across a wide range of research domains. Having noted this, it is imperative to acknowledge that qualitative paradigm diverge fundamentally from its quantitative counterpart, as it extends beyond a mere procedural task. As Yin [2011] emphasises, qualitative research is not a mechanical process but rather one that requires a mental framework, which guides the study throughout its progression and governs the construction of data collection instruments, the recording and analysis of qualitative data, and the interpretation of findings,

ensuring that the research maintains its rigour and consistency.

Expanding on the design of qualitative data collection instruments, Yin [2011] gives a detailed consideration of interviews as a versatile and widely employed tool, distinguishing between structured, semi-structured, and unstructured formats, each suited to different research objectives. Structured interviews, characterised by predetermined questions administered uniformly, allow for comparability across responses. Semi-structured interviews strike a balance between structure and flexibility, enabling researchers to delve deeper into topics based on the responses of participants.

Illustrating the versatile applicability of semi-structured interviews, studies in diverse research contexts demonstrate their value in integrating qualitative insights into the analysis of complex phenomena, the interpretation of quantitative findings, the integration of theoretical frameworks, and the development of more comprehensive models. For instance, in healthcare, studies often go beyond reliance on purely quantitative data, integrating qualitative approaches such as semi-structured interviews to capture both measurable trends and the underlying contextual mechanisms influencing healthcare outcomes [Buchbinder et al., 2023; Tase et al., 2022; van Poelgeest et al., 2021]. Providing the notion of its usage, in Buchbinder et al. [2023], which examines occupational stress in healthcare, the use of semi-structured interviews supplements quantitative data, offering contextual depth to explore systemic factors contributing to physician burnout.

Outside healthcare, semi-structured interviews have been extensively utilised in business research to explore a range of organisational dynamics, from innovation adoption and leadership strategies to digital transformation processes and market adaptation strategies, offering a nuanced understanding of complex busi-

ness phenomena that quantitative methods alone may not be equipped to address [Grama-Vigouroux et al., 2020; Rialti et al., 2022; Schilling and Seuring, 2023]. As another compelling example of the effective deployment of interviews is provided by Grama-Vigouroux et al. [2020], which employs semi-structured interviews to examine the determinants guiding firms in their transition from closed to open innovation models. Utilising a qualitative approach uncovers the factors that foster or hinder the adoption of open innovation in small and medium-sized enterprises (SMEs) and reveals key insights into the contextual and relational dynamics influencing the adoption process. Derived from interviews, these findings offer a depth of understanding that quantitative data alone may be unable to capture, illuminating the key elements to consider in business decisions when embracing open innovation.

Sparing a detailed discussion, semi-structured interviews have also been employed in other fields, including law [Alexopoulos et al., 2020; Gialdini et al., 2024; Moser-Plautz, 2024] and psychology [Firat and Bildiren, 2024; Henke et al., 2022; Mumtaz and Nadeem, 2022], where they contribute to understanding legal processes, informing policy implementation, and exploring cognitive development and social behaviours.

Whilst not employed in the current study, unstructured interviews offering the most flexible form, focus on exploring participant experiences and narratives without rigid constraints, and have also been frequently utilised in research [Chalhoub et al., 2021; Qureshi et al., 2021; Vermaak and de Klerk, 2017]. In addition to delineating semi-structured and unstructured interviews, Yin [2011] provides detailed guidance on formulating interview questions, emphasising the importance of clarity, neutrality, and alignment with research goals. Moreover, the role

of the interviewer is highlighted as crucial in creating a conducive environment that encourages participants to share rich, authentic insights.

Accordingly, interviews, as qualitative techniques, are particularly effective in uncovering the contextual and narrative depth that quantitative data alone may fail to capture. By employing a dynamic interview instrument, complex behaviours and contexts can be explored, allowing for a more nuanced understanding of phenomena [Taylor et al., 2015]. For example, interviews may reveal subtle cognitive processes or social dynamics that may remain obscured in quantitative surveys or experimental data. Substantially, when utilised in the study of online information consumption, interviews as a data collection instrument may capture parameter values related to uncertainty, risk aversion, learning, cognitive load, and welfare outcomes during engagement with information, thus providing critical data to test, inform, and refine the model, from which other data methods are devoid.

On the other hand, when melded with quantitative approaches, such as surveys or experiments, qualitative methods, including interviews, may contribute to findings that are not only statistically valid but also contextually rich and narratively robust, thus providing a more comprehensive understanding of the research problem [Blessing and Chakrabarti, 2009]. Resultantly, this integration of qualitative and quantitative methods may well enhance the depth and breadth of the analysis, offering a richer, more holistic perspective.

In discussing the concept of combining qualitative and quantitative data, Yin [2011] also suggests that such integration may amplify the explanatory power of the data itself, leading to an ever more comprehensive understanding of the research problem. For instance, qualitative insights gleaned from interviews or

observations can inform the design of quantitative instruments, such as surveys, allowing the refinement of the variables and questions that will be explored in a larger sample. In the study by Higgins and BuShell [2018], for instance, the qualitative data gleaned from teacher interviews and student surveys were instrumental in defining the focus group questions, thereby fostering a meaningful incorporation of both data types. This iterative process, where qualitative data informs and refines quantitative models, is central to the success of mixed-methods research. It ensures that both data types are not only integrated but that they contribute meaningfully to each other.

The integration of qualitative and quantitative data, as discussed by Yin [2011], requires that both data types be considered in tandem throughout the study. The analysis of both types of data should not treat them as isolated entities but as interconnected components of a larger research framework [Creswell and Creswell, 2017]. By synthesising qualitative insights with quantitative data, a more holistic understanding of the phenomena investigated may be produced, ensuring that the findings reflect both individual experiences and broader patterns.

Having established the need for a robust and adaptable framework for understanding complex cognitive behaviours, the principles outlined by Yin [2011] for integrating qualitative and quantitative methods provide a solid foundation for conducting research in the absence of adequate quantitative data. The integration of qualitative insights into the design of quantitative models, along with the maintenance of a coherent relationship between both data types throughout the study, facilitates the generation of more comprehensive and insightful findings [Merriam and Tisdell, 2015]. This approach not only enhances the validity and

relevance of the study but also ensures that the research process remains rigorous and flexible, allowing for iterative refinement and a deeper, more nuanced understanding of the research problem [Creswell and Creswell, 2017]. The concurrent use of both qualitative and quantitative methods offers substantial benefits in studies of cognitive behaviour, serving to bridge the gap in current research, particularly in addressing the complexities of cognitive behaviours involved in verifying information online. Therefore, the combination of these methodologies provides a powerful tool for model calibration and validation of its outcomes, as well as for the exploration of optimal effort levels within the framework of cognitive trade-offs when verifying information online.

In clarification how qualitative methodologies may refine the parameters of a cognitive effort model, the study by de Gramatica et al. [2017] serves as an illustrative example demonstrating the practical application of the methodological principles articulated by Yin [2011]. de Gramatica et al. [2017] address agency costs within the domain of aviation security by combining a quantitative economic model with qualitative insights derived from semi-structured interviews. This methodological integration offers a comprehensive framework for capturing the interplay between cognitive and physical effort, monetary and non-monetary incentives, and the value of transferable human capital. The qualitative component, based on extensive interviews with airport security personnel and key stakeholders in Turkey, not only serves to validate the trade-offs identified in the quantitative model but also adds crucial contextual depth to the understanding of these trade-offs in practice.

Through these interviews, the researchers identify key factors influencing the effectiveness of security training. These factors include the alignment between

emotional buy-in, transferable skills, and cognitive burden, as well as the critical tipping points where additional training aligns security personnel's incentives with those of the airport authorities. Importantly, this qualitative data clarifies the non-monetary incentives that are crucial in mitigating agency costs and emphasises the role of cognitive and emotional factors in motivating security personnel. The findings suggest that non-monetary incentives, particularly those fostering intrinsic motivation and long-term career development, are pivotal in shaping the impact of security training programs.

In line with methodological framework of Yin [2011], this study employs a mixed-methods approach that intertwines qualitative case study research with quantitative modelling. The use of semi-structured interviews, guided by purposive sampling, is fundamental in obtaining the most relevant data and ensuring that the empirical evidence aligns contextually with the theoretical predictions of the model. Moreover, the semi-structured interviews were designed to elicit open and detailed responses through a set of pre-circulated grand tour questions, supplemented by follow-up questions tailored to interviewee reactions. These qualitative data points, derived from semi-structured interviews conducted with key airport security stakeholders, were audio-recorded, transcribed, and augmented with hand notes capturing perceptions and reflections. The results of these interviews were used to explore agency problems, assess motivational factors such as transferable skills and responsibility, and validate the quantitative model predictions regarding effective security training and its impact on reducing moral hazard. Applied in this fashion, the mixed methods approach highlights the relevance of the principles of methodological rigour, particularly in how qualitative insights refine and validate the assumptions of the quantitative model. The inte-

gration of qualitative data does not merely complement the quantitative analysis but actively enhances it, ensuring that the model predictions reflect the nuances of real-world decision-making processes.

The qualitative interviews in this study also demonstrate the emphasis on the iterative process between qualitative and quantitative methods, where qualitative insights do not merely validate the model but also inform its design. In particular, qualitative data derived from stakeholder experiences of security training enables the researchers to adjust the assumptions of the model about the effectiveness of different training regimes. This iterative adjustment helps illuminate the complex interplay between intrinsic and extrinsic incentives, cognitive load, and the development of transferable skills. Through this process, the research exemplifies how the methodology established by Yin [2011] allows for the adaptation of theoretical models to better account for the diverse, context-specific factors that influence decision-making.

Moreover, the conclusions of the study align with the broader implications of the work by Yin [2011], where qualitative data enhances the interpretation of quantitative results. In this case, the qualitative findings suggest that the traditional models of security training, based purely on fixed monetary rewards, fail to account for the intrinsic motivational factors and human capital considerations that shape security personnel's behaviour. By incorporating these qualitative insights into the model, the researchers are able to design a more effective security training portfolio that accounts for both the cognitive and emotional dynamics of the agents involved. The application of qualitative methods ensures that the model is not only validated but also enriched, offering a more comprehensive understanding of the factors driving security decisions.

The study further highlights that, although the mixed-methods approach does not allow for precise point predictions, it provides a valuable framework for risk analysts to understand the broader trends and dynamics at play. Through the integration of expert insights and empirical evidence, the study offers a robust methodology for designing security interventions where empirical data is scarce, and controlled experiments are ethically or practically impossible. This reinforces the contention that qualitative research is indispensable in shaping the context and interpretation of quantitative models, particularly when addressing complex socio-technical systems where individual behaviour and decision-making are central.

Thus, by applying mixed-methods principles, de Gramatica et al. [2017] offer a compelling case for the use of qualitative data in enhancing the predictive validity of quantitative models, ensuring that these models reflect the complexity of real-world decision-making. The integration of qualitative and quantitative research methodologies allows for a more nuanced understanding of the cognitive and emotional factors involved in verifying information and making risk-related decisions, contributing to the development of more effective, evidence-based interventions in high-stakes environments such as airport security.

Another notable example of a study employing a mixed-methods approach is the research by De Gramatica et al. [2015], which examines the challenges of cybersecurity regulations in the aviation sector. The study explores how interdependencies between IT systems within airports and across sectors complicate the design and implementation of effective cybersecurity policies, particularly in light of how existing regulations may disadvantage smaller airports with fewer financial resources, despite the increasing threats posed by cyberattacks in an

interconnected environment.

The study incorporates both quantitative economic modelling and qualitative case study research to analyse the trade-offs involved in achieving optimal cybersecurity investments, particularly balancing the regulatory demands and economic limitations faced by stakeholders, such as smaller airports. The quantitative component involves an economic analysis of cybersecurity investments and the associated costs and benefits, particularly in relation to the incentives and behaviours of different size airports as well as mandated security requirements. By using quantitative techniques, such as cost-benefit analysis and economic game theoretic modelling, the researchers identify key variables that influence decisions regarding the level of investment in cybersecurity measures.

However, as is often the case in studies involving complex socio-technical systems, the quantitative analysis alone cannot fully capture the intricate social and organisational factors that shape airport decisions and attitudes toward cybersecurity regulations and subsequent investments. To address this limitation, De Gramatica et al. [2015] incorporate qualitative data collected through semi-structured interviews with cybersecurity experts, airport authorities, and government regulators. This qualitative data provides rich insights into the challenges airports face in implementing cybersecurity measures, as well as the non-monetary incentives that influence their decision-making processes.

The integration of qualitative data serves multiple purposes within the study. First, it provides contextual depth, elucidating the complex social, organisational, regulatory and individual factors that may drive or hinder cybersecurity investment decisions. For example, the airport stakeholders interviewed find cyberthreats challenging to quantify and often consider them as unpredictable

risks, which adds to the uncertainty in risk management and thus decision making in cybersecurity investments. Additionally, the qualitative data highlights unintended consequences of specific regulatory frameworks, including the risk of regulatory fatigue among smaller airports, which often face significant challenges in adhering to compliance requirements due to constrained financial resources. These insights inform the trade-offs in the financing mechanism for cybersecurity within the economic model of optimal expenditure presented in the study.

On top of offering crucial contextual insights into the emerging cybersecurity threats, the study by De Gramatica et al. [2015] utilises qualitative interviews to assess the efficacy of existing security regulations in mitigating these risks. The qualitative data collected through these interviews is pivotal in refining the economic model by providing clarification on the assumptions and parameter values that inform the quantitative analysis. Specifically, the insights gained from stakeholder perspectives serve to adjust and contextualise the theoretical underpinnings of the quantitative framework, ensuring that it accounts for both the economic and security dynamics at play in the aviation sector. This iterative exchange between qualitative insights and quantitative modelling is essential for enhancing the validity of the model and applicability to real-world scenarios.

Through the careful integration of these qualitative insights, the researchers fine tune their quantitative models, ensuring that they better reflect the real-world dynamics at play in the aviation sector. This iterative process of refinement, where qualitative data informs and enhances quantitative analysis, aligns with the emphasis Yin [2011] makes on the complementary roles of both research methods. Notably, the qualitative component not only validates the quantitative findings but also contributes to the theoretical development of the research, providing a

more nuanced and holistic understanding of the phenomena being studied.

In particular, the usage of semi-structured interviews as a qualitative tool demonstrates the utility of this method in unpacking the motivations and perceptions that drive decisions in the context of cybersecurity. The interviews, conducted with a purposive sample of experts and decision-makers, are carefully designed to explore the intricacies of the airport interests and the broader social dynamics influencing cybersecurity governance. The qualitative data, once transcribed and analysed, provides valuable insights that help refine the economic model, enhancing its predictive accuracy and relevance to the real-world context.

Moreover, the approach the study applies to synthesising both data types signifies and reinforces the argument mixed-methods research should not be treated as an additive process but rather as one where both quantitative and qualitative elements are integrally woven together to strengthen the overall research framework. The qualitative data enriches the interpretation of the quantitative results, offering a deeper understanding of how economic and regulatory considerations align with the investments in cybersecurity. The fairness analysis conducted in the study further emphasises the disproportionate cybersecurity cost burden faced by smaller airports compared to larger ones, accentuating the need for redistribution mechanisms within the regulatory framework to address this imbalance.

Ultimately, by combining quantitative economic models with qualitative case study data, the study exemplifies how mixed-methods research can enhance the explanatory power of the analysis and offer a more comprehensive understanding of complex, multi-faceted issues. The integration of these methods allows the researchers to account for both the objective, measurable variables and the subjective, context-dependent factors that shape decision-making, resulting in more

robust and applicable findings to real-world regulatory contexts.

To conclude, the study by De Gramatica et al. [2015] on cybersecurity regulations in civil aviation offers a compelling example of how the integration of qualitative and quantitative methods, as advocated by Yin [2011], can lead to more comprehensive and actionable insights. By grounding their economic models in the contextual realities manifested through qualitative interviews, the researchers enhance the validity and applicability of their findings, ensuring that the proposed regulatory frameworks are not only economically sound but also socially and organisationally feasible. This case further underscores the importance of adopting a mixed-methods approach in research involving complex systems and stakeholder interactions, where a deep understanding of both economic and human factors is essential for formulating effective solutions.

While the aforementioned studies provide direct insights into the integration of qualitative data into quantitative modelling, it is also worth acknowledging that qualitative data have been extensively utilised to inform not only quantitative models but also theoretical frameworks, conceptual structures, and broader theories. For instance, Elliott et al. [2019] construct a conceptual framework to examine organisational methods of knowledge protection. The study underscores the trade-off between promoting innovation via open communication and mitigating security risks through the imposition of information flow restrictions. This equilibrium is shaped by variables such as the sensitivity of the information, the trustworthiness of employees, and the legal protective measures. By leveraging qualitative data, the framework posited by Elliott et al. [2019] bridges conceptual understanding and practical applications in organisational settings, demonstrating the broader utility of qualitative evidence beyond model calibration.

Building on this conceptual framework, the study draws on evidence from HP Labs to assess the its predictions and applicability. The qualitative data support the emphasis the model places on employee trustworthiness and the significance of informal codes of behaviour in maintaining open communication while simultaneously managing security risks. For example, interviews conducted at HP Labs underscore that although open communication is essential for innovation, there are circumstances, such as with highly sensitive client information, that necessitate restrictions to prevent potential breaches. This dual focus on encouraging innovation while safeguarding sensitive data highlights the utility of qualitative insights in capturing the multifaceted dynamics of organisational behaviour, which quantitative models alone cannot fully address.

Furthermore, the framework emphasises that information security requirements are not uniform across all organisational tasks. While some data, such as client confidentiality, particularly for sensitive accounts, demand higher security measures, other types of information such as technical data may be subject to less stringent controls. These variations underscore the need for tailored security strategies that align with specific organisational priorities. The study also exposes that informal mechanisms, including trust, training, and relational contracts, play a critical role in supporting security practices. These findings exceed the scope of the model validation to also provide deeper as well as practical insights into how organisations may craft strategies balancing innovation needs with robust information protection.

While the primary focus of the HP Labs case study is on knowledge protection strategies, the methodological approach itself highlights the broader potential of integrating qualitative data into research efforts. Semi-structured interviews, doc-

umentary analysis, and process coding enabled a rich exploration of knowledge-sharing practices, offering actionable insights into balancing innovation with security. The study thus exemplifies how qualitative evidence can inform both conceptual frameworks and real-world applications, providing a comprehensive analysis that complements and extends quantitative methods.

Another noteworthy research, which rather than employing a quantitative model uses a theoretical framework, is presented in Haugstvedt and Tuastad [2023], where qualitative data is shown to enhance and refine theoretical structures, particularly in the context of multi-agency collaborations. In their investigation of the role of social workers cooperating with the police and security services within frameworks organised to prevent and counter violent extremism, the authors utilise qualitative methods such as in-depth interviews and focus group discussions to examine the unintended consequences of these collaborations. The qualitative data collected is instrumental in revealing the nuances of professional dynamics and jurisdictional disputes between social workers and law enforcement, which could not be captured through quantitative analysis alone.

The theoretical framework guiding Haugstvedt and Tuastad [2023] draws on the theory of jurisdiction proposed by Abbott [2014], which seeks to explain the conflicting professional interests and territoriality that arises when multiple agencies with overlapping responsibilities engage in collaborative efforts. By incorporating qualitative data, the study enhances the understanding of these jurisdictional tensions and their implications for the professional roles of social workers. Interviews with practitioners provide concrete examples of how these tensions manifest in practice, helping to test and refine the theoretical framework proposed by Abbott. For example, the study uncovers the ethical dilemmas faced

by social workers in Norway as they navigate the blurred boundaries between welfare provision and security surveillance, a dynamic that quantitative data alone could not fully capture.

The integration of qualitative data Haugstvedt and Tuastad [2023] perform thus serves a crucial function in exploring and validating the theoretical assumptions underlying the collaborative framework. It demonstrates how qualitative insights can provide deeper contextual understanding, shedding light on the complexities of professional identity and ethics in multi-agency settings. These insights not only support the theoretical framework but also inform the broader applicability of the findings to policy and practice.

In conclusion, the study underscores the significant role qualitative data plays in refining theoretical models and frameworks, particularly in complex, real-world contexts. By integrating qualitative evidence, the authors are able to provide a richer, more comprehensive analysis that would be difficult to achieve through quantitative methods alone. This approach marks the value of qualitative data not only in model calibration but also in testing and enriching theoretical assumptions, thereby expanding the utility of qualitative research in bridging theoretical gaps and enhancing the robustness of quantitative models.

Drawing on the methodological principles demonstrated in these studies, the methods outlined by Yin [2011] offer a solid framework for testing the validity of theoretical models and calibrating their parameters to enhance the relevance of the model outcomes. In this process, qualitative data may be used to inform the assumptions and parameters of the quantitative model, ensuring they are grounded in the lived experiences and real-world contexts of the data sources. Thereby, this adjustment of parameters not only allows to strengthen the robust-

ness of the model but also align it with the complexities of investigate cognitive behaviours, reflecting the nuanced interplay between qualitative and quantitative insights, as observed in online information consumption decisions.

4.2 Methodology: Calibrating the Model Using the Case Study of Ukraine War

With the general structure of the model for online information consumption concerning misinformation defined in Chapter 3, and further refined in Chapter 4 to address the deliberate disinformation strategies characteristic of political scenarios, this chapter shifts toward a more focused application of the model. Specifically, to investigate its implications, the model will be embedded within the ongoing war in Ukraine, utilising it as a case study to explore the dynamics of online information consumption, which, in turn, allows to evaluate the validity and robustness of the model in a real-world context while fine-tuning its parameters to better capture the distinctive characteristics of disinformation observed during the Ukraine conflict.

As a contextual input for the model, this war offers a compelling instance of how disinformation thrives under conditions of heightened stakes and uncertainty. The conflict serves as an invaluable lens through which individual and societal interactions with misleading content can be examined. Moreover, it enables the application of the model to explore the intricate dynamics of cognitive effort, transaction and cognitive costs, and the consequent trade-offs between the benefits and costs of information consumption under uncertainty. By delving

into this environment saturated with disinformation, the optimal strategies for information verification may be assessed.

To reiterate, political disinformation campaigns, as articulated in the game-theoretic model of information consumption in Chapter 4, leverage diverse strategies to manipulate public opinion, erode institutional trust, and shape the discourse surrounding critical issues. While misinformation is pervasive across various domains, including health, politics, and science, its impact is particularly pronounced in conflict settings. War, by nature, fosters an environment conducive to disinformation due to the inherent uncertainty, the proliferation of competing narratives, and the high stakes surrounding public perception. Such contexts provide a critical juncture for examining the effects of disinformation on both individuals and society.

In this respect, the ongoing war in Ukraine offers a rich case study of disinformation phenomenon in a conflict setting. The invasion initiated by Russia in 2022 reoriented global focus dramatically, moving attention away from the COVID-19 pandemic to the geopolitical shock of war. The pandemic itself had been rife with misinformation, from debates surrounding vaccine efficacy and public health measures to conspiracy theories regarding government control. These narratives contributed to a polarised information landscape, which, as the war unfolded, seamlessly morphed into a new set of geopolitical narratives.

To provide context, it is essential to recognise that the conflict in Ukraine did not emerge suddenly but can be traced back to 2014, when Russia annexed Crimea, setting in motion a series of events that escalated into a full-scale invasion of Ukraine in February 2022. In the years preceding this invasion, Russia orchestrated a comprehensive disinformation campaign aimed at undermining

the position and image of Ukraine on the global stage. These efforts sought to portray the Ukrainian leadership as aligned with extremist ideologies, thereby attempting to discredit its political authority and legitimacy, such as by framing the Ukrainian government as a fascist regime[1]. Additionally, Russia disseminated narratives accusing Ukraine of suppressing the rights of Russian-speaking populations, particularly those in Crimea and eastern Ukraine, using these claims as a pretext to justify its intervention and annexation under the guise of a supposed peacekeeping mission. Disinformation efforts further aimed to disrupt the aspirations of Ukraine for Euro-Atlantic integration, as expressed in the 2008 Bucharest NATO summit and enshrined in the 2019 constitutional amendment, which set the goal of full NATO membership, with Russia alleging closer ties with NATO to pose a direct threat to regional stability and portraying Ukraine as a pawn exploited by Western powers advancing their agendas. These strategies formed part of a broader disinformation apparatus designed to manipulate both domestic and international audiences by distorting facts and constructing a more favourable narrative to the geopolitical ambitions pursued by Russia.

Compounded by the the rapid advancement of technology, which has fundamentally transformed the ways in which individuals access and consume information, not only has the overall significance of disinformation soared but its impact, particularly in the context of conflict such as the Ukraine war, has also been amplified. In the past, people tended to rely on traditional media such as broadcast television, radio, and newspapers for news. These sources, while still influential today, have been increasingly supplanted by digital platforms that allow for

[1]See: Fascism Conquered Most of Ukraine, East Stratcom Task Force, https://euvsdisinfo.eu/report/fascism-conquered-most-of-ukraine/, Accessed: 30 November 2024

the near-instantaneous dissemination of information. With the proliferation of the internet and mobile technology, the information landscape has expanded exponentially, and individuals are now bombarded by an overwhelming variety of content. News related to economics, politics, sports, celebrities, music, and culture is now available in real-time, and can be accessed via social media platforms such as Facebook, Instagram, X (formerly Twitter), blogs, podcasts, and news aggregators.

This digital evolution has not only expanded the range of available topics but has also fostered an environment where both true and false information spread rapidly and may be created by anyone, regardless of expertise or intent. Historically, major events such as the moon landing in 1969 or the fall of the Berlin Wall in 1989 were moments of global significance, during which information flowed through slower, more traditional outlets and which also gave rise to conspiracy theories and false narratives of the matter in question. However, the speed at which information now travels is unparalleled. Events like the Arab Spring and the spread of misinformation during the COVID-19 pandemic demonstrate how both true and false stories can travel across the globe in the blink of an eye, reaching millions instantaneously. Such technological advancement has ushered in new challenges, particularly in the realm of misinformation and disinformation, where misleading content can be as potent as factual reporting and can be deliberately designed to polarise public opinion. As such, Ukraine offers a striking example of how social media platforms play a crucial role in disseminating both accurate and misleading content across its borders and in neighbouring regions.

The Ukrainian case thus serves as an extreme context in which the mechanisms of costly verification, strategic disinformation, and government signalling

are particularly salient and observable. While the empirical focus is geographically specific, the underlying mechanisms of online harm modelled in this thesis, including costly cognitive effort, exposure to manipulative content, and the use of institutional or social signals, are not unique to Ukraine and may also emerge in other polarised online ecosystems, albeit under different parameter configurations. Accordingly, the interviews are interpreted as an informative extreme-case study that sharpens the identification of these mechanisms, while the discussion throughout the chapter explicitly acknowledges that the numerical calibration and the intensity of harms cannot be generalised mechanically across countries or platforms.

The Ukrainian case thus serves as an extreme context in which the mechanisms of costly verification, strategic disinformation, and government signalling are particularly salient and observable. While the empirical focus is geographically specific, the underlying mechanisms of online harm modelled in this thesis costly cognitive effort, exposure to manipulative content, and the use of institutional or social signals are not unique to Ukraine and can arise in other polarised online ecosystems, albeit with different parameter configurations. Accordingly, the interviews are interpreted as an informative extreme-case study that sharpens identification of the mechanisms, while the discussion throughout the chapter explicitly acknowledges that the numerical calibration and the intensity of harms cannot be generalised mechanically to all countries or platforms.

The rapidly advancing online platforms and the increasing prevalence of disinformation campaigns have been further accelerated by the onset of the war in Ukraine. In this context, disinformation has become a pivotal tool for shaping both international and domestic perceptions. Social media platforms, particu-

larly Telegram, have become crucial vectors for the dissemination of both factual information and misleading content among the Ukrainian population. The decentralised and encrypted features of Telegram, combined with its widespread use across Ukraine, Russia, and neighbouring regions, have facilitated the rapid spread of disinformation. This underscores the importance of taking into account both contextual and platform-specific dynamics when studying how information is consumed, which is crucial for modelling the trade-offs individuals face in determining the optimal cognitive effort required to distinguish between true and false information. Fig. 4.1 presents a typical screenshot from Telegram on the war in Ukraine from August 2, 2025.

While the rationale for analysing the model of online information consumption under the uncertainty of disinformation is clearly exemplified by the Ukrainian conflict, as this represents the first attempt to formally model information verification in an online context, the methodological approach to calibrate the model parameters presents considerable challenges, particularly with regard to capturing values related to risk aversion, utility, effort, and costs through quantitative data and statistical analysis. Given the limitations in obtaining robust quantitative data measurements for these variables, integrating the quantitative model of online information consumption with qualitative approaches becomes essential for elucidating the complex human behaviours and decision-making processes that underlie these parameters. This combination may represent the most viable option available for analysis in the absence of reliable statistical data, prompting a transition towards mixed methods, where qualitative insights not only complement but also inform quantitative analyses, thereby offering a more holistic understanding of the model dynamics [de Gramatica et al., 2017]. Although qualitative

Figure 4.1: A typical screenshot from a Telegram channel on the Ukrainian War. This screenshot refers to a drone attack by Ukraine on a Russian Oil Depot, the picture is from August 2, 2025.

exploration of the theoretical framework does not produce precise numerical outcomes, it is instrumental in identifying trade-offs and defining potential solution domains, shedding light on the behavioural complexities of optimal information verification, wherein personal cognitive effort is balanced against the associated benefits and costs of distinguishing between true and false information.

In the present chapter, the semi-structured interviews are therefore used to validate three core elements of the theoretical model, namely the existence of a non-linear cognitive cost of verification effort, heterogeneous perceived losses from misinformation, and the presence of manipulative signals that introduce noise into belief updating. The qualitative narratives are interpreted as evidence on the shape and relative magnitude of these trade-offs, rather than as statistical estimates, and are used to assess whether the patterns implied by the model are consistent with the behaviours described by interviewees.

Therefore, as a means to bridge the gap between the theoretical assumptions and real-world dynamics of information consumption, the quantitative model is validated through qualitative insights derived from individuals directly engaged in information verification during the Ukrainian conflict. Specifically, the outputs of the quantitative model are synthesised with the experiences of these individuals through a two-phase qualitative data collection approach, involving semi-structured interviews and a follow-up survey. With the procedure guided by the methodological framework of Yin [2011], model parameters may be refined and calibrated, aligning them more effectively with the contextual conditions and unique challenges observed in the Ukrainian disinformation environment.

In operationalising this approach, the first phase involves conducting semi-structured interviews with purposively selected individuals in Ukraine, aiming to

capture the cognitive, emotional, and contextual factors that influence the decision to verify or disregard online information. To provide further clarification, the interview guide was developed in advance to address key themes pertinent to understanding the context in which the participants operate. As outlined by Merton et al. [2003], this approach enables a flexible yet structured exploration of the experiences of interviewees, ensuring that their perspectives may be examined within the unique circumstances of their environment, particularly the war zone in Ukraine. Fostering deep and reflective discussions, the interview questions are constructed to be open-ended and broad, facilitating the emergence of detailed narratives and personal accounts [Brenner, 2012]. These exploratory questions, properly designed to arouse broad subject matters [Maxwell, 2008], encourage participants to reflect on their encounters with disinformation, thereby capturing essential aspects of their lived experiences, and thus gather rich, contextually grounded data. The core questions are summarised in 4.2 To support a natural conversational flow, interviewees were encouraged to elaborate on their initial responses, with follow-up questions introduced as needed to ensure that all relevant aspects of their specific circumstances were thoroughly explored.

The interview protocol was designed to map directly onto the theoretical components of the model. Questions 1-3, which explore positions, responsibilities, and the challenges encountered by respondents in their professional roles, were intended to elicit information about baseline beliefs, exposure to informational constraints, and the implicit costs of verifying or sharing content within their institutional settings. Questions 4-5, which probe definitions and examples of misinformation and disinformation interviewees maintain, capture perceptions of informational accuracy, exposure to manipulative interference, and the types of

cognitive effort typically required to distinguish credible from misleading signals. Finally, Questions 6-7, which invite participants to identify under-examined issues and offer additional reflections, provide insight into how individuals prioritise verification relative to other demands and how these priorities reflect the broader cost-benefit logic of information engagement. The responses to these clusters were thematically coded and used to construct the stylised relationships depicted in Figures 4.2 and 4.3, which illustrate the implied trade-offs between verification effort, perceived loss from misinformation, and cognitive burden rather than representing a direct statistical fit to the small sample.

To proceed the interview process, a non-random sampling approach was adopted [Maxwell, 2008], with access to participants secured through a gatekeeper, as conceptualised by Yin [2011]. Leveraging their extensive expertise in cybersecurity, financial technology, and transnational research initiatives within the Ukrainian context, the gatekeeper identified and selected six individuals from the Ukrainian Cluster Alliance. These individuals were chosen based on their specialised knowledge, domain expertise, and direct engagement with issues pertaining to disinformation in Ukraine, ensuring the relevance and depth of insights obtained. In Table 4.1, details of the interviewees, including their respective roles and affiliations, are presented.

The relatively small sample size of $N = 6$ is appropriate in this context because the purpose of the interviews is not to estimate population parameters but to conduct an in-depth, theoretically informed validation of the mechanisms embedded in the model. Following Yin [2011], the cases were selected purposively as information-rich and heterogeneous instances of intensive engagement with disinformation in the Ukrainian context, which allows the analysis to probe

whether the predicted trade-offs between effort, perceived loss, and manipulation are observable across diverse roles and experiences. The diversity of backgrounds therefore enhances, rather than undermines, the validation exercise by testing the robustness of the mechanisms under varying informational and institutional conditions.

This configuration of interviewees ensures coverage of multiple vantage points across information production, dissemination, and policy engagement. It enables the mechanisms of the model, particularly the trade-off between verification effort and perceived loss, to be examined under heterogeneous cognitive and institutional constraints. The small but strategically varied sample is consistent with the principle of analytical rather than statistical generalisation Yin [2011] and helps to validate whether the predicted behavioural relationships hold across contrasting informational environments.

#ID	Role	Institution	Interview Time	Interview Date
1.	Head	Danish Ukrainian Resource Centre	49 min	28/10/2024
2.	PR and Engagement Director	Greencubator	57 min	30/10/2024
3.	President	Ukrainian Cluster Alliance	38 min	31/10/2024
4.	Head of Startup Accelerator	Kyiv National University of Architecture and Development	38 min	31/10/2024
5.	Head	Digital Innovation Hub Ukraine	37 min	07/11/2024
6.	Journalist, Researcher	Energy Think Tank Ukraine	58 min	22/11/2024

Table 4.1: Participant details of the Semi-Structured Interviews

These stakeholders were purposefully chosen for their direct engagement with disinformation phenomena in Ukraine. They represent individuals from organisations potentially targeted by disinformation campaigns, those directly affected by the Ukrainian conflict, and those actively involved in addressing potentially

misleading information [Bloom and Van Reenen, 2010]. To safeguard participant privacy, the names and identities of the interviewees were anonymised.

The interviews were conducted via Microsoft Teams, beginning with introductions and contextual background delivered by the gatekeeper. Each session lasted approximately 30–50 minutes and, with the consent of participants, was audio recorded. The recordings were transcribed automatically and subsequently amended to correct transcription inaccuracies, ensuring fidelity to the original recordings.

These interviews serve as the primary data source, providing rich insights into the decision-making processes of individuals within the specific circumstances of the disinformation environment in Ukraine and form the foundation for the second phase, which involves constructing and administering a survey tailored to the Ukrainian audience and the wartime context. This approach captures the nuances of their experiences and, most importantly, quantifies the cost-benefit trade-offs inherent in information verification.

4.3 The Design of Semi-Structured Interviews

The semi-structured interview questions outlined in table 4.2 were meticulously formulated to elicit nuanced insights into the cognitive, strategic, and behavioural dimensions of information engagement, both in professional and personal contexts. Given that the theoretical model investigates decision-making under uncertainty in environments rife with misinformation and disinformation, each question was designed to probe specific aspects of how individuals process, evaluate, and respond to information in both high-risk and routine settings. The questions

thus serve a dual function of providing empirical grounding for the assumptions of the model regarding the costs and constraints associated with information verification and concurrently refining its parameters by capturing the complexities of information consumption across different roles and responsibilities.

The opening question, which seeks an account of the position and key duties of the respondents, establishes an essential contextual foundation. By understanding the professional domain in which an individual operates, it becomes possible to discern whether their responsibilities necessitate active information engagement, verification, or dissemination. This distinction is critical, as those in roles requiring frequent interaction with information, particularly in policymaking, journalism, or security-sensitive professions, may exhibit distinct cognitive and strategic adaptations compared to individuals whose engagement with information is more circumstantial. Additionally, this question enables an assessment of whether institutional constraints or organisational priorities shape the ways in which information is processed and acted upon.

The second question explores whether their professional position requires public consensus or support. This is particularly relevant in professions where decision-making is contingent on external validation or stakeholder alignment, such as governance, policy advocacy, or public-facing roles. The necessity of consensus-building imposes additional layers of complexity on information verification, as individuals in such positions may be required to navigate conflicting narratives while ensuring credibility and maintaining public trust. The degree to which external pressures shape information-processing strategies is particularly significant, as it may influence not only the effort expended on verification but also the willingness to engage with contested or ambiguous information, thereby

affecting the broader informational ecosystem in which such individuals operate.

The third question seeks to identify the most challenging aspects of the role of the interviewee, thereby providing insight into whether information-related burdens constitute a significant strain in their professional lives. In roles where decision-making hinges on accurate information, the cognitive demands of filtering, verifying, and acting upon information may be magnified, leading to increased verification costs. The model accounts for such costs through its transaction cost function, and responses to this question help refine the parameters by identifying the specific sources of cognitive pressure and the extent to which they influence decision-making. Furthermore, sustained cognitive strain in professional settings may not remain confined to occupational contexts but may extend into personal domains, altering the manner in which individuals engage with information outside of work, shaping their verification habits, and potentially exacerbating information fatigue in their broader daily interactions.

The fourth and fifth questions probe the conceptual understandings of misinformation and disinformation, exploring whether these terms are perceived as distinct phenomena and how such distinctions inform their approach to information evaluation. Existing literature frequently distinguishes misinformation as the unintended propagation of inaccurate information, whereas disinformation is characterised by the deliberate dissemination of falsehoods with the intent to mislead. The extent to which individuals recognise and operationalise this distinction in their verification strategies is crucial, as it affects their decision-making under uncertainty. If respondents are able to provide examples from their own experience, these accounts offer empirical evidence of how theoretical classifications manifest in real-world contexts, thereby allowing to refine the theoretical assump-

tions about the means by which individuals interpret and categorise unreliable information.

The sixth question broadens the scope by inquiring whether respondents perceive any critical issues related to misinformation that have not been adequately explored in existing research. In particular, this exploration facilitates the recognition of emerging challenges that may not yet be formalised in academic discourse, thereby ensuring that the model remains attuned to contemporary realities of the evolving informational landscape. By prompting respondents to reflect on information-related difficulties in their day-to-day activities, this question also provides insight into potential structural or cognitive barriers to effective information evaluation that may not be readily discernible when approached as purely rational decision-making process.

Additionally, the follow-up sub-questions a. and b. encourage respondents to prioritise or categorise the sources of misinformation they encounter, offering a comparative perspective that may further refine the approach employed in the model to weighting different types of informational threats. Finally, the concluding question invites respondents to share any additional thoughts or reflections, ensuring that no relevant dimensions of information engagement are overlooked. Open-ended responses often reveal latent patterns in information-processing behaviour, particularly regarding the psychological and emotional toll of navigating volatile informational environment in Ukraine. This question also provides an opportunity to identify any unforeseen cognitive burdens or strategic adaptations that may not have been explicitly captured in the preceding inquiries.

By systematically addressing these dimensions, the interview framework ensures that the model is not only theoretically robust but also empirically cal-

ibrated to reflect the lived experiences of individuals operating in complex informational landscapes. The insights derived from these responses enable the refinement of key parameters, particularly regarding the cognitive costs of verification, the role of risk aversion in information-related decision-making, and the broader constraints individuals face when engaging with uncertain or potentially deceptive information.

	Question
1.	Could you tell me about your position and what your key responsibilities or duties are (you can provide more generic examples if the duties are confidential).
2.	Does your position require any kind of public consent or support for you to operate? For instance, in political decision-making, do you need to find consensus with either colleagues or members of the public?
3.	What are the most challenging aspects of your job?
4.	How would you define misinformation and disinformation? What differences, if any, do you see between them?
5.	If no examples are given for [4], ask about some examples from their experience, either in their current position or previous positions.
6.	Are there any specific issues that you feel are important but have not been examined carefully in research that you are aware of?
	a. Example: Are there any challenges you face in your day-to-day activities that you feel have documented approaches to solving them?
	b. If they have more than one example, can they categorize the importance of different misinformation sources?
7.	Are there any further comments or thoughts that you would like to share?

Table 4.2: Interview Questions

Table 4.3: Mapping interview questions to theoretical constructs of the model

Q No.	**Interview focus**	**Corresponding model construct / mechanism**
1-3	Roles, responsibilities, and main challenges in respondents' professional or organisational context	Baseline beliefs and opportunity costs influencing verification effort x; cognitive and temporal constraints shaping the cost function $C(x)$
4-5	Definitions and examples of misinformation and disinformation drawn from personal experience	Perceived accuracy of information and exposure to manipulative interference s; formation of the posterior probability $\hat{\rho}$ or $\hat{\rho_s}$ through noisy belief updating
6	Identification of issues insufficiently addressed in research or practice, and categorisation of misinformation sources	Perceived losses L from misinformation and prioritisation of threats; evaluation of trade-offs between cognitive cost and expected informational benefit
7	Open reflections and additional comments	Overall behavioural adaptation to uncertainty; corner solutions such as disengagement or high effort consistent with utility maximisation under the CRRA specification

Table 4.3 summarises how each interview question relates to the theoretical components of the model, clarifying how the qualitative material informs parameter interpretation and subsequent validation.

4.4 Results

This section presents the findings derived from the the analysis of a series of interviews conducted with professionals in Ukraine who engage with information in various capacities and are potentially exposed to and affected by misinformation and disinformation. These interviews offer critical insights into real-world

decision-making processes surrounding the consumption, verification, and dissemination of information, particularly as they are drawn from a high-stakes environment where the prevalence and ramifications of misinformation and disinformation are especially pronounced. Effectively, the analysis explores how participants navigate the complexities of information consumption, including the level of effort they invest, the risks they perceive, the cognitive costs they incur, and the strategies they employ to both minimise their exposure to false information and mitigate its dissemination.

The primary objective of this section is to critically analyse the responses of the participants through the lens of the game-theoretic model developed in this study. By incorporating elements such as cognitive costs and effort, the model provides a structured framework to assess how individuals weigh the potential gains derived from accessing truthful information against the losses incurred due to misinformation or disinformation. Anchoring the analysis in the key components of the model, this section aims to identify patterns in decision-making by the interviewees, quantify their behaviours, and extract insights that enhance the understanding of information verification in real-world contexts. This, in turn, facilitates the parameterisation, validation, and refinement of the theoretical model of information consumption and verification

The interviews were structured to elicit responses related to key themes, including risk aversion, cognition and the cost of information, the distinction between misinformation and disinformation, improvements in information, and the uncertainty that accompanies information consumption. Each of these themes will be examined in turn, with direct excerpts from the interviews illustrating the diverse perspectives and decision-making processes exhibited by participants.

In parallel, these findings will be systematically integrated into the theoretical framework, providing a nuanced understanding of how individuals operating in different professional environments navigate the challenges of information verification.

4.5 Risk Aversion and Preferences

In the context of online information consumption, risk aversion manifests in the strategies individuals employ to verify, avoid, or act upon information, and in risk-laden environments, such as the war situation in Ukraine, wherein misinformation or disinformation may have tangible consequences. Within the theoretical framework laid out in this study, risk aversion and individual preferences are captured by a utility function, which determines how individuals weigh potential losses resulting from false information against the potential gains associated with truthful information. This section draws on the semi-structured interviews to integrate the parametric insights into the model regarding individual preferences as quantified by a class of Constant Relative Risk Aversion (CRRA) utility functions, while also examining and refining the assumptions and underlying dynamics of information engagement from a practical, real-life standpoint. Given the proliferation of disinformation during conflicts, individuals may be expected to exhibit a heightened degree of risk aversion, as the stakes of misinformation transcend mere informational inaccuracies and can directly impact personal safety, professional responsibilities, and national security.

Given the high prevalence of disinformation in the context of war, individuals are likely to exhibit an increased preference for certainty regarding the situa-

tion in Ukraine, as access to accurate information is critical for both immediate decision-making and long-term considerations regarding the country's future. Consequently, individuals may exhibit a strong preference for information that provides a definitive and reliable knowledge outcome, indicating a heightened degree of risk aversion. Effectively, the interviews unveil dynamic patterns in risk preferences, with some participants initially engaging intensely with information in an effort to navigate uncertainty, while others, over time, adopt filtering strategies or disengage entirely to mitigate cognitive and emotional costs. These evolving responses highlight the adaptive nature of risk aversion in information consumption, reinforcing the model's predictions that individuals calibrate their engagement with information based on perceived risks, expected utility, and the broader contextual landscape of the conflict.

Given the high prevalence of disinformation in the context of war, individuals are likely to exhibit an increased preference for certainty regarding the situation in Ukraine, as access to accurate information is critical for both immediate decision-making and long-term considerations regarding the future of the country and personal welfare. Consequently, individuals may exhibit a strong preference for information that provides a definitive and reliable knowledge outcome, indicating a heightened degree of risk aversion. Effectively, the interviews provide empirical evidence of these tendencies, illustrating how preferences evolve over time and under varying conditions, exhibiting dynamic patterns in risk perception and information-seeking behaviour.

Some participants appear to have initially engaged intensely with information in an effort to navigate uncertainty, while others, over time, adopted filtering strategies or disengaged entirely to mitigate cognitive and emotional costs, im-

plying a shift in preferences. These evolving responses highlight the adaptive nature of risk aversion in information consumption, reinforcing the predictions of the model that individuals calibrate their engagement with information based on perceived risks, expected utility, and the broader contextual landscape of the conflict.

To firstly put the nature of information exposure into perspective, a clear discrepancy emerges between the impact of misinformation and disinformation in professional and personal domains. Within professional settings, individuals appear to be less affected by misleading information, which often arises as a consequence of mismanagement or a lack of analytical scrutiny, falling under the premise of misinformation, rather than deliberate attempts to deceive, as is typical in the cases of disinformation. One of the interviewees highlighted the issue of incorrect information circulating within professional networks:

> "So the 1st for profession. Is here is. Main. Impact. I think that if. Some. Project that you work give you some not. Enough or wrong information. That. Impacts that you at least spend them time for them? Yes and. This time spent not not efficiently. Yes, for for for me. Or for this project you spent. Time resource of my my cap. Yes. And it's not efficient so it's. Impact from the site, so the the suspended resources not not proper, right?"
>
> — Interviewee 5

This suggests that misinformation in professional settings often results in inefficiencies and resource misallocation, rather than direct harm. The impact is primarily economic, as misinformation can create misleading trends or misplaced priorities. Another participant elaborated on how such misinformation propagates within their industry, creating self-reinforcing cycles of misleading information:

> "Yeah, misinformation, typically. Very spread in my professional life because. Typically it relates to wrong priority of people. In my environment when for example, they try to react on many opportunity or threats, we actually we are we are not really important, so they just follow some public hype and some. Information. Yeah, feedbacks or information figures. And without any analysis, if it's important, not important, but they they they raise this hype, they try to follow some some modern trends and it creates still bigger hype of the bigger information bubble when other people are mislead and mislead and they just lose focus on on true priority."
>
> — Interviewee 3

The above illustrates how misinformation can create self-sustaining information bubbles, diverting attention from substantive issues and amplifying misleading narratives. Although these cases do not reflect intentional deception, they demonstrate how professional environments can inadvertently generate and perpetuate misinformation. On the other hand, one participant recounted an incident highlighting that, at times, false information is deliberately provided by others within their professional sphere, leading to wasted time and efforts:

> "And. In in. One example that. I said, said said the. Message. In in Facebook that from. A man that prides that he's cancer. Ready. Nice innovation. Centre yes and. I'll. It provides the website. And I spent some time for some communication and as well I. Sent to my team or this information to to whether we willing to collaborate and while checking this all information we discovered that. It's not. Really. Put. Not, not not. Any information in the. About a legal entity, no, no information about. The Centre No, no project. Real project. So. So it was just. Spend part time spent for this communication."
>
> — Interviewee 5

The example above reflects the cost of engagement with unverified sources, reinforcing the prediction of the model that individuals with higher sensitivity to losses will be more inclined to exert verification effort before acting on information.

While misinformation is prevalent in professional spheres, disinformation that is false information disseminated with deliberate intent appears to have a more significant impact on personal decision-making. A notable pattern that emerges from the interviews is the evolution of risk preferences over the course of the war. Initially, many participants engaged with information intensely, but as the conflict progressed, their approach changed, often towards filtering or avoidance strategies. One participant recalled how disinformation succeeded at targeting emotions in the early days of the invasion:

> "According like these specific topics, but in my life? Sure I was Influenced by that. The most. I would say like we have this example in general, then the war started. It was. We had a lot of disinformation. And also how to say? Informational attacks on Ukrainians, and I remember that we had an open use information that Russians made some marks on the tops of the buildings and everyone need to head to Take them off from the streets or from the top of the buildings, and it was during the first days of invasion and I was also Sharing this news because I was scared that they will hit like Russians will charge was going to target those buildings according to those marks. But this this wasn't true. So it was Disinformation also fake information because they Were targeting my Emotions and it was successfully. They was successful. They succeeded that."
>
> — Interviewee 2

Nevertheless, the interviewee indicated that as the war has progressed, it became increasingly necessary to detach from the information consumption:

> "For example our media. They had a lot of these News which can Influence on my emotions, for example, about [war prisoners?], or which? Killed on the front line, et cetera. So I knew that those news Will affect me. Or it can be a Russian propaganda, again about nuclear War or bomb, and we already had that in the past so. Now I just skip those news or don't pay attention to them because. Well, then, the nuclear war will happened. We will know. But right now I need to finish my work, sorry. So I had this kind of attitude now so."
>
> — Interviewee 2

This shift in behaviour reflects an adjustment in risk preferences under the relevant circumstances of informational manipulation, where the individual began to prioritise mental well-being and stability over the need to stay constantly updated to maintain a sense of normalcy. Initially, emotional responses to disinformation were heightened, as shown by the reaction of the interviewee to misleading news about Russian markings on buildings. However, over time, the emotional impact diminished, and the participant adopted a more measured approach to information. In a parallel vein, another interviewee also stated having actively engaged with news at the outset of the war:

> "For example, beginning of war in Ukraine. Yes, if I use it tomorrow will be a war. Yes. I have my daughter and go out from my place. Of living so, but I was. Was not. had not this information, so this was some problem. From the from. When Some fighting was around and it was not easy to escape."
>
> — Interviewee 5

but, as the conflict progressed, their reliance on information shifted from continuous news monitoring to more immediate and situational alerts necessary for safety and preparedness:

> "But, but. When in the court files of the war in the in this in. The Kyiv region where I live. This was opposite side because I check news every every hours because I have had to know situation about fighting and to be ready to escape or some. So so from from this point of view. Information. Needed but. But now we have, for example, alarm chatter for in Ukraine, where regional administration alarm, for example, when some dangerous rockets or so on so. Here. in Kyiv almost all services alarm and its use as this is not about, news more about. More about some alarm."
>
> — Interviewee 5

Essentially, this encapsulates an initial preference for engagement with information due to its perceived urgency and potential consequences. However, over

the course of time, complemented by improvements in war-time infrastructure such as the establishment of reliable alert systems, some individuals adjusted to the environment, exhibiting a gradual shift towards avoidance strategies, selectively filtering news to minimise emotional distress:

> "Maybe. I. Yes, I. filter filter information and I cannot. For example, I don't see some any news and don't read the news. So because we have a lot of different news and a lot of negative news because. A lot of news about fighting about rocket attacks. This so. I. Don't see it and don't read news any day. But. Sometime only. See for for some important look for example. Election in USA some. Situation on in war. But not not every day. It's some, maybe one or a few. A few day. I just check what happened in the in the world, in the Ukraine."
>
> — Interviewee 5

Such behavioural shift suggests that risk aversion is dynamic, adapting to the cognitive costs of continuous exposure to distressing information. Individuals reassess the trade-off between the perceived importance of information and the emotional or cognitive burden associated with engaging with it. However, not all individuals responded by disengaging. While disinformation has led some individuals to avoid information and news, some, instead, increased their information consumption in an effort to navigate uncertainty:

> "With war, we really have a lot of disinformation and If, so I'm trying not to read news. I can't handle them. But for example, my husband always read news. Listen to some YouTube video and I see how effect they have on him. Just today, he said I am so disappointed about Something he read about, I don't know, USA, they said. They said something. And we, you know, it's hard to understand How war is going? So I think we, I'm talking about me and my closest we don't have trust, we we don't trust no Ukrainian. News, no International because. They're always different and. The real Picture I don't know from the front line we can heard from other relatives who are there, Has a big difference from this, which I see in the news sometimes, so I would say I don't trust maybe information at all."
>
> — Interviewee 1

Nonetheless, these lived experiences may suggest that the Ukrainian population has developed a heightened scepticism towards information due to the war and the pervasive presence of disinformation. The evidence indicates that distrust in information sources, particularly institutional narratives, can manifest not only as increased scrutiny but also as a near-complete disengagement. This finding underscores the complexity of risk aversion, demonstrating that increased scepticism does not invariably translate into greater verification efforts. Instead, in some cases, it paradoxically may lead to a withdrawal from information consumption altogether as individuals perceive the cognitive and emotional costs of engagement to outweigh the potential benefits of staying informed.

Synthesising these insights, the interviews illustrate how risk aversion in information consumption is contingent upon individual preferences, evolving over time and adapting to both personal and contextual factors. The findings align with the model predictions that individuals adjust their verification effort x based on perceived costs and benefits, while their risk aversion remains subject to changes in perceived stakes. The ongoing war in Ukraine serves as an extreme test case, reinforcing the notion that individuals dynamically calibrate their engagement with information in response to its potential risks and rewards.

From a model-based perspective, as risk aversion G increases, the costs associated with engaging with information also rise. This results in individuals becoming increasingly disengaged, viewing the effort required to verify information as disproportionately burdensome. The growing aversion to risk, combined with the increasing perceived costs of engagement, thus leads to a reduction in the effort x that individuals are willing to invest in verifying information. In this manner, as G increases, the act of engaging with information itself becomes less

attractive, resulting in a reduction in both engagement and verification effort.

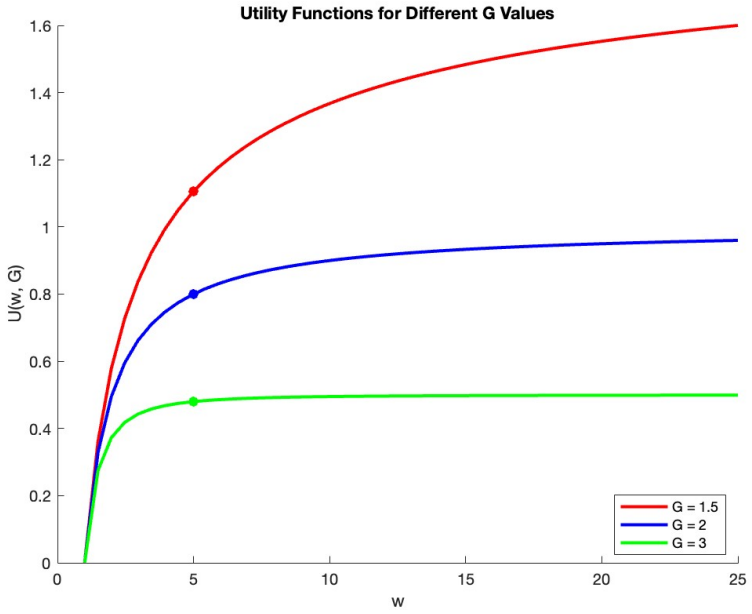

Figure 4.2: Utility Functions for Different G Values

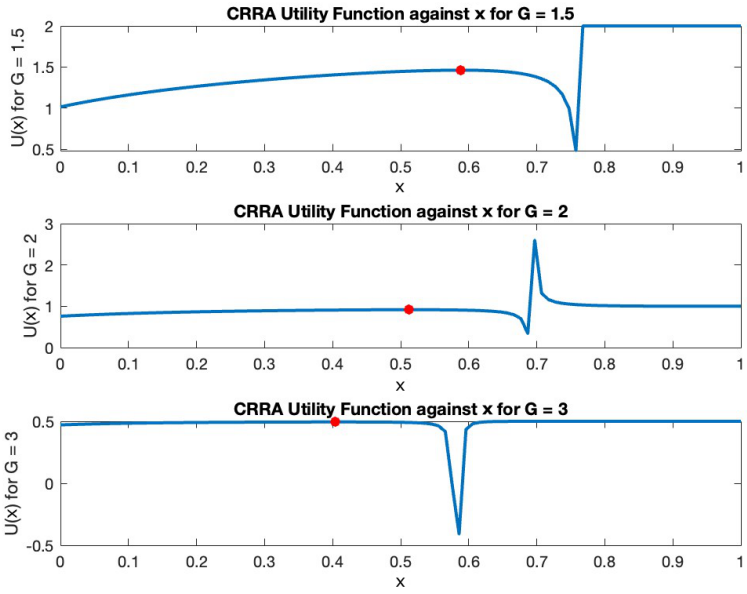

Figure 4.3: x Values for CRRA Utilility Functions for Different G Values

c *Note:* Figure 4.3 plots the Constant Relative Risk Aversion (CRRA) utility function $U(x, G)$ for three representative degrees of risk aversion, $G = \{1.5, 2, 3\}$. The function follows the CRRA specification implemented in the model:

$$U(b, G) = \begin{cases} \dfrac{w_1(x, \hat{\rho}_s, G)^{1-G} - 1}{1 - G}, & G \neq 1, \\ \ln\big(w_1(x, \hat{\rho}_s, G)\big), & G = 1, \end{cases}$$

where $w_1(x, \hat{\rho}_s, G) = (1 - \hat{\rho}_s(x, s))V - L\hat{\rho}_s(x, s) - C(x, G)$ represents the expected outcome net of informational losses and cognitive costs. The horizontal axis corresponds to the verification effort x, and the vertical axis depicts the associated expected utility. As risk aversion G increases, the utility curves flatten, indicating reduced marginal utility from additional verification effort and a lower optimal effort level x^*, shown by the red point on each curve.

In Figure 4.2, three CRRA utility functions are depicted, each representing individuals with varying degrees of risk aversion as captured by the parameter G. It is essential to recognise that comparing utility functions across individuals presents significant challenges, primarily due to the inherently subjective nature of utility. Since utility does not possess a standardised unit of measurement, direct inter-individual comparisons are not feasible. Nevertheless, a clear trend emerges where individuals with lower values of G, corresponding to lesser risk aversion, derive greater utility from engaging with information. As risk aversion increases, reflected by higher values of G, the utility derived from engagement diminishes, with the utility function becoming flatter. This flattening indicates that more risk-averse individuals perceive diminishing returns from information engagement, as the perceived costs of verification rise disproportionately to the potential benefits. Taking the analysis further, in Figure 4.3, the red dots on the CRRA utility functions represent the optimal values of x illustrating how individuals with differing risk preferences G not only derive different levels of utility but also identify distinct levels of effort x as optimal for their subjective assessments of information processing and verification. This divergence underscores the central role that risk preferences play in shaping individual decisions regarding information consumption, reflecting the critical intersection of economic theory and behavioural choice in environments characterised by uncertainty and the need for costly verification.

To distinguish between disengagement and unaffordable cognitive cost, each transcript was coded line by line. Statements in which participants explicitly described withdrawing from particular platforms, muting channels, or avoiding news altogether were coded as instances of disengagement, whereas statements

describing a desire to check information but an inability to do so due to time pressure, mental fatigue, or limited resources were coded as episodes of high marginal cognitive cost. Figures 4.3 and 4.2 plot stylised functional forms that are consistent with the relative frequency and sequencing of these codes across interviews, rather than providing an econometric fit. Subjective interpretations are stated explicitly in the accompanying text, and ambiguous responses are only used to illustrate possible mechanisms, not to anchor the shape of the curves.

Empirically grounded, the interview excerpts substantiate the notion that risk aversion directly influences the effort x individuals are willing to expend in verifying information. While it may intuitively follow that more risk-averse individuals would engage more extensively in verification, motivated by the potential risks of acting on incorrect information, the interviews reveal a more intricate reality. The cognitive and emotional burdens associated with verification may render it prohibitively expensive, prompting individuals to reconsider their engagement. In this sense, heightened risk aversion does not necessarily translate into greater verification efforts. On the contrary, it may lead individuals to disengage entirely from information, perceiving the cognitive and emotional costs as insurmountable to justify any further involvement. In the context of the Ukrainian war, this dynamic is particularly evident, as heightened exposure to misinformation and disinformation has driven individuals to recalibrate their engagement approaches, adjusting their verification efforts in response to evolving perceptions of risk, at times withdrawing altogether. The escalating complexity and cost of information verification, especially in conflict zones characterised by pervasive disinformation, underscore the complex relationship between risk aversion and the decision-making processes that guide information consumption.

Upon reflection, the interviews demonstrate that risk preferences in online information consumption are not static but evolve over time, driven by both emotional and cognitive costs, particularly in conflict-affected environments such as the war in Ukraine, wherein misinformation and disinformation are rampant and can significantly influence decision-making. Participants exhibited varying preferences, initially displaying a strong inclination towards engaging with information, driven by the urgency they attributed to it, but progressively adopting avoidance strategies to mitigate the cognitive and emotional burdens associated with constant exposure. These findings align with the theoretical model, which suggests that individuals dynamically adjust their information engagement according to perceived risks and rewards. In essence, as risk preferences influence the level of verification effort, elevated perceptions of risk prompt a lower cognitive investment in verifying information as individuals become more attuned to the potential costs of misinformation, tailoring their strategies accordingly.

4.6 Cognition and Cost of Information Verification

The verification of online information entails cognitive effort, which incurs costs that individuals must weigh against the potential benefits of obtaining accurate information. Within the theoretical model, these costs are captured by the function $C(x)$, which reflects the increasing cognitive costs associated with exerting higher levels of effort x. As the effort required to assess information increases, individuals must make trade-offs between allocating ever-increasing cognitive re-

sources for verification or other competing priorities. These transaction costs might be expected to be particularly high in demanding environments, such as the ongoing war in Ukraine, where individuals face both an abundance of information and the heightened risk of exposure to misinformation and disinformation. The insights gathered from the interviews provide valuable input for tuning the model, particularly with regard to the specific cognitive pressure and strategies individuals employ in response to information overload and emotional strain. By integrating the diverse verification strategies the participants apply with the model, the calibration process ensures that the theoretical framework accurately captures the nuances of informational decision-making, especially in environments characterised by high uncertainty and risk.

The interviews reveal that individuals engage in different strategies of information verification, implying different levels of effort, which may depend on factors such as their cognitive capacity, available resources, and the perceived necessity of obtaining reliable information. A case in point is one participant describing their verification strategies as dependent on pre-existing beliefs about the credibility of information providers. This approach entails disregarding sources deemed untrustworthy while scrutinising information from typically trusted providers only if it raises inconsistencies or conflicts with prior expectations:

> "So I put it through certain criteria and if to start with some sources of information I'm not even consider. So they I I don't even look at them because it's. By default, not trusted can't be trusted, and I and I don't, and I don't waste my time on it if I get the information. From the source, I usually tend to trust, right? And it does not correlate with something that I saw, but I expected to see from that sort of information. So it says something different. Then I'm trying to see the purposes why they have changed their mind, why what could be the reason? So I tried to get other information from different sources in respect to that fact to see what could be the purpose of changing

> the opinion of. So it's a kind of complex complex process from my side."
>
> — Interviewee 1

In conjunction with shedding light on the risk preferences, the explanation put forth by the interviewee corroborates the notion that the effort x depends on the valuation of the content with which the individual engages, with unreliable sources being bypassed entirely, as the value of information from these providers is perceived as too low to merit costly verification, resulting in no effort expenditure. Moreover, the selective engagement strategy of the participant aligns with the theoretical expectation that individuals optimise cognitive effort by minimising unnecessary verification when certainty about an information source is already established. Under conditions of substantial informational noise, the cost of verification may become prohibitively high, leading individuals to opt out of information engagement altogether. This reinforces the model prediction that $C(x)$ acts as a deterrent when the cognitive effort required exceeds the perceived benefits. The tendency to disengage from information also parallels the findings on risk aversion, where an overload of conflicting information may prompt individuals to rationally reduce verification efforts rather than attempt to process an unmanageable volume of data.

Echoing this sentiment, another interviewee described their engagement with information as increasingly selective and strategic. Rather than completely disengaging, they adopted a filtered approach, restricting their exposure to specific sources and limiting engagement to certain times of day to mitigate cognitive and emotional burdens:

> "I now developed a few rules for myself that I will I'm I can read and use just after work. I also clean My telegram channels and also also

> I'm very specific with the media which I'm reading. I also Choose the information which I consume. For example our media. They had a lot of these News which can Influence on my emotions, for example, about [war prisoners?], or which? Killed on the front line, et cetera. So I knew that those news Will affect me. Or it can be a Russian propaganda, again about nuclear War or bomb, and we already had that in the past so. Now I just skip those news or don't pay attention to them because."
>
> — Interviewee 2

In view of the intense recount of experiences, the respondent demonstrates an intentional adaptation of verification effort in response to the psychological strain imposed by excessive exposure to emotionally charged content. Such filtering mechanisms function as a self-imposed cognitive safeguard, limiting engagement with distressing information while preserving a degree of situational awareness. This selective filtering strategy highlights how individuals regulate their cognitive load while ensuring they remain informed about critical developments.

Furthermore, the structured approach to verification is reflected in strategies of the participants for guiding others in their immediate environment. One interviewee described actively educating family members about the risks of misinformation and disinformation, providing them with practical strategies to assess credibility:

> "I keep telling to my parents that they have to be also Aware about disinformation, about propaganda, about informational war, and I only like I wrote them the questions how they can Check If should they believe in this information or not. First of all, it is who pay for this information. Do you know who is owners of the media? Who said that? Or do they have links so you can go and check their first source now? Also, it is important to see what kind of feelings you have after the news. When you read it, so it's sure it's impossible to. Understand All the fake news but And disinformation and et cetera, but still. It's very relevant for everyone to Learn how we can protect ourselves and our families, loved ones, etcetera."
>
> — Interviewee 2

Put in the broader context of information consumption, this response underscores an externalisation of verification strategies, where individuals not only refine their own cognitive filtering mechanisms but also seek to instil similar habits in others. This suggests that awareness of misinformation is a collective issue, reinforcing the idea that verification efforts transcend the individual level to broader social networks. When viewed through the lens of the model, such externalisation, in line with the theoretical assumptions, highlights the dynamic adaptation of verification strategies, as individuals adjust their efforts based on cognitive load, the perceived costs of verification, and the external resources available to them. More specifically, cognitive effort x may be higher in some individuals, particularly those with limited resources or cognitive capacity, while others may benefit from external help, thereby reducing the perceived cost of verification. In this regard, the model predicts that when cognitive load becomes too great, individuals may choose to rely on social networks or trusted sources to ease the burden, thereby demonstrating lower levels of effort x. Ultimately, this reflects the interplay between individual cognitive constraints and socially distributed verification efforts in shaping information consumption behaviours.

Beyond filtering strategies, nonetheless, the overall common approach among interviewees remains the reliance on a set of trusted sources to streamline information consumption:

> "It's. It's good to have some trustful sources for information and not spent time because really. In Internet there are some. Some information that. Is that is that you? For example, you see that the source is some. Not. Well known sources so so you. Put. Understands that it's would be not not true, so it's important to have some sources. Of verified sources of information and so. That is mean for for you. You don't need. To spend the time to to check this information because if you for example for example if you. Have some news that is. Is not.

> Well, not usual. So. You need to check it from other sources anyway. If it's very important, use for example. News about related to some. Rocket attack, yes. You know that it's going to be some."
>
> — Interviewee 5

> "Disinformation from Russia's side, for example, or something. So you you check this information from different from Ukraine, news from European news and so on. And. Run information. More important, so you more more. You anytime I have to check this information. From different sources"
>
> — Interviewee 5

With other interviewees seconding this approach, yet, emphasising the necessity to explore more when content is not certain:

> "I choose medias whom I Believe. I believe that they made some, some they, some of those [part]. They there I believe that they verify information and they work in to find the first source, etcetera. So, like the economy, eastern New York Times, when we talk about these global media and in Ukraine I also read some media home Whom I know. And if I have others, Like another information, I need to check. Who said that? Why he said that. And as I told you already. What does it mean for me and should I verify it somewhere or I'm OK just with this piece of information? And sure what I should do if that."
>
> — Interviewee 2

> "I don't know. I OK. Obviously I don't believe in some publication from no name person from the Twitter or Facebook. When I see some hysterical Text, I won't believe it. There are some people who write about war about The situation which I think I can't believe like an informational source, but maybe like emotional I don't."
>
> — Interviewee 4

These responses signify that trusted sources function as cognitive shortcuts, reducing the verification burden. However, while reliance on such sources can be an effective strategy for efficiency, it may also increase the risk of confirmation biases, as individuals selectively engage with information that aligns with their prior beliefs.

Essentially, these examples highlight that there are significant transaction costs associated with information verification, where individuals must allocate time and cognitive resources to discern the legitimacy of claims. In many of the instances, participants appear to have developed a structured system for processing information, which simplifies the process, varying from relying solely on vetted sources to employing a set of rules to interrogate the information when doubt arises. This suggests different levels of effort, contingent on an individual's cognitive capacity, available resources, and the perceived necessity of obtaining reliable information:

> "I think I'm quite experienced information consumer and. I just don't have time enough. You know, there's so much to do."
>
> — Interviewee 1

As a consequence, the cost of verification effort is not trivial, as individuals may experience opportunity costs such as the time spent verifying information, which diverts attention from other critical tasks. This aligns with previous assertions that individuals, when faced with information overload, often weigh the cognitive effort against the perceived value of the information:

> "In general, as I told you earlier, we don't have always enough time. To verify everything so I I know that also I can be influenced influenced by. Something you can use. Well, so. I think like still the challenge here is To check everything by myself, sure, like and also I knew. I know that if I will, if I check information for example on on the website some official website still I don't know how this information was made and for what. So because I already mentioned to you that official person from my the head of the regions said. Misinformation. So in this case I have to think like with helicopter view a bit and see Where the waves of this information goes, what the purpose behind that and sure when you have a lot of work and your personal life, it is not always easy."
>
> — Interviewee 2

Compounding these challenges, interviewees generally indicate that, within their professional lives, the necessity of task prioritisation, coupled with limited available time, significantly constrains their capacity for both information consumption and production. Consequently, information that falls outside their professional responsibilities, especially when requiring closer assessment, may become increasingly difficult to manage:

> "A lot of tasks that you need to do during the day. And sometimes they are too much. For 8 hours, let's say. And a lot of information. So I become overwhelmed in the end of the day"
>
> — Interviewee 4

From the viewpoint of the model, these findings align with the prediction that cognitive costs $C(x)$ impose significant constraints on verification effort by introducing opportunity costs that individuals must weigh against competing demands. As professional obligations and daily responsibilities accumulate, the marginal cost of verification increases, making extensive scrutiny of information less viable. As expected, the interviews imply that, beyond a certain threshold, $C(x)$ becomes prohibitive, leading individuals to either disengage from verification efforts or adopt heuristic-based strategies to optimise cognitive resources. Thereby, the notion that information processing is not solely determined by individual intent but is dynamically shaped by cognitive limitations and external pressures is reinforced.

Moreover, the findings suggest that individuals selectively interrogate information sources when the content deviates from their expectations, while completely avoiding those deemed unreliable. This behavioural pattern is consistent with the model prediction illustrated in Figure 4.4 positing that as uncertainty ρ increases, individuals initially augment their verification effort x. However, once

uncertainty surpasses a critical threshold, the cognitive load intensifies to such a degree that verification effort x declines sharply, reflecting a rational disengagement in response to the overwhelming complexity of the information environment.

In addition to the substantial cost of verification under conditions characterised by significant informational noise and rising prior uncertainty ρ, the disengagement from interaction with content may also be influenced by varying levels of risk aversion. As the effort required for verification increases, particularly at higher levels of initial uncertainty ρ, the cost may become increasingly prohibitive. This, in turn, may lead some individuals, depending on their risk preference G, to lower their optimum effort threshold, ultimately opting to reduce or entirely forgo engagement with information. As depicted in Figure 4.5, verification costs may become excessive for highly risk-averse individuals to sustain their engagement with the information, with Figure 4.6 providing a closer view of the increases in G leading to relatively higher transaction costs $C(x^*)$ for individuals with greater risk aversion exerting optimal cognitive effort.

Further illustrating this dynamic, Figure 4.7 shows how increasing risk aversion G results in diminishing levels of optimal verification effort x. This trend suggests individuals exhibiting higher risk aversion perceive the utility of engaging with uncertain content as insufficient to warrant the cognitive effort required, thereby opting to disengage or reallocate their resources to alternative activities that are perceived as more rewarding or less cognitively taxing.

Effectively, drawing on the interview data, the findings indicate that information verification is not an all-or-nothing process but rather a dynamic trade-off between effort, cognitive burden, prior beliefs and perceived informational value. The theoretical framework predicts that individuals will adjust their verification

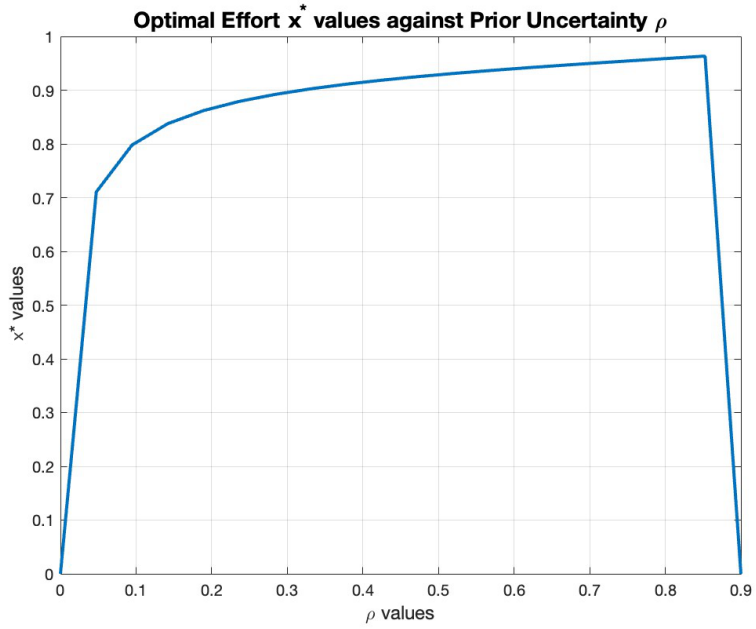

Figure 4.4: Optimal Effort against Prior Uncertainty ρ

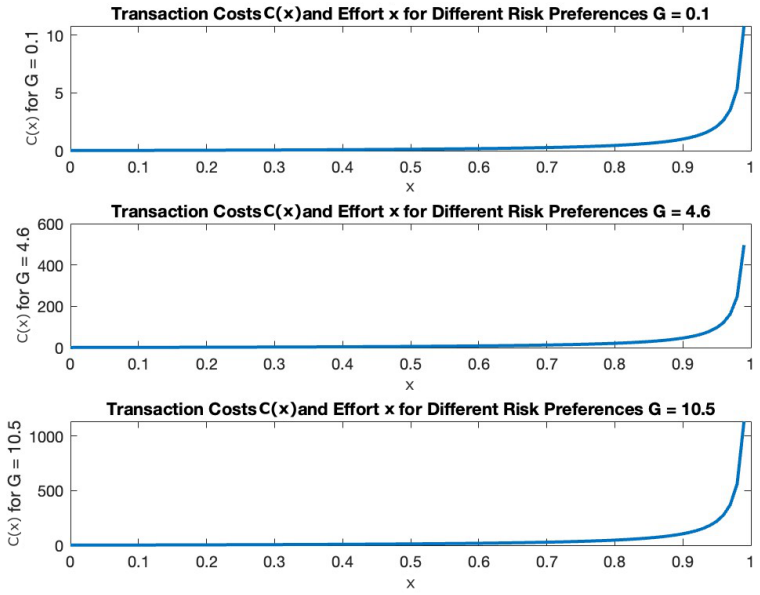

Figure 4.5: Transaction Costs and Effort for Different Risk Preferences

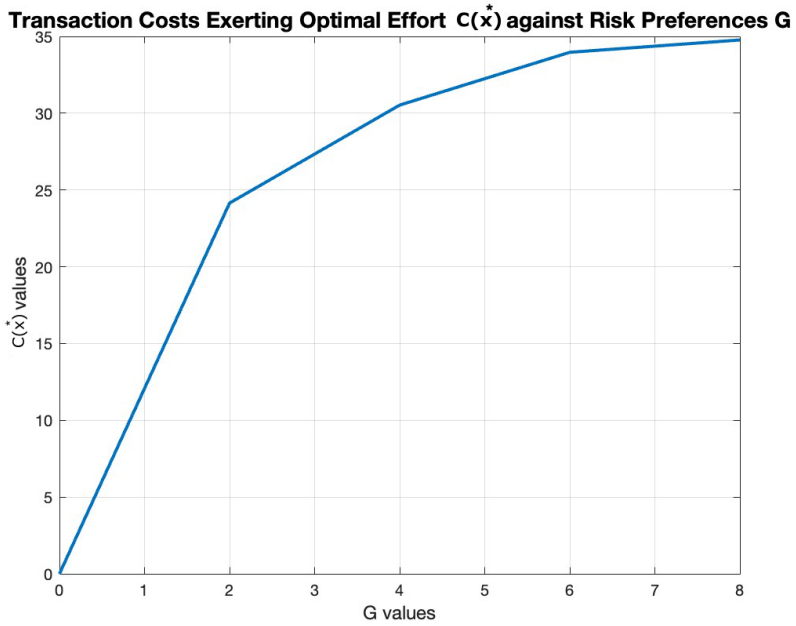

Figure 4.6: Transaction Costs Exerting Optimal Effort against Risk Preferences

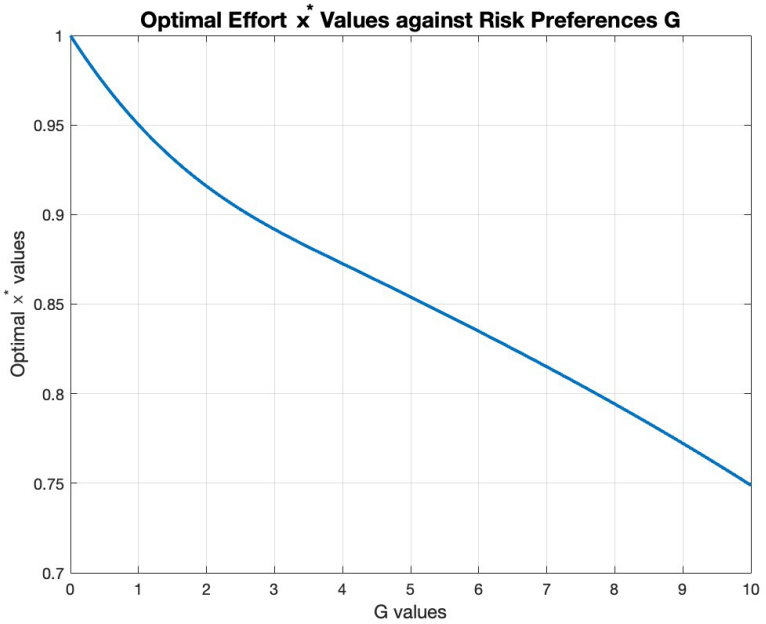

Figure 4.7: Optimal Effort against Risk Preferences

strategies according to the marginal costs of effort, aligning with the empirical evidence that some participants opt for disengagement, while others refine their filtering mechanisms. In substance, the interviews demonstrate that cognitive costs play a significant role in shaping verification behaviours or the lack thereof, further validating the postulation of the model that transaction costs $C(x)$ are key determinants in the decision to engage with or disregard information.

4.7 Distinction between Misinformation and Disinformation

Having established the role of risk preferences and cognitive effort in forming information consumption behaviours, the distinction between misinformation and disinformation becomes pivotal in understanding the complexities of verification strategies. Misinformation, broadly construed, encompasses all instances where incorrect or misleading information is propagated, irrespective of intent. Within this category, disinformation constitutes a distinct and deliberate subset, wherein falsehoods are strategically crafted and disseminated with the explicit objective of manipulating perceptions and shaping narratives. While misinformation may arise from errors, misinterpretations, or unverified reporting, disinformation is inherently linked to a malign actor who engages in intentional deception.

This differentiation is particularly crucial within the theoretical model, as the presence of an adversarial entity actively distorting informational landscapes introduces a strategic element that necessitates increased verification effort x on the part of the consumer. In this context, the model conceptualises disinformation

as an adversarial signal, where an actor exerts effort s to mislead the information consumer. The extent to which individuals allocate cognitive resources to counteract disinformation depends on their awareness and anticipation of the potential for manipulation, reflecting the strategic, game-theoretic nature of information verification. The interaction between the strategy of the malign actor and the response of the consumer creates an equilibrium, wherein the information consumer effort x is dynamically adjusted in response to expected intensity of distortions, resulting in varied verification behaviours depending on perceived risk and cognitive costs. Empirical insights from interviews provide a critical opportunity to examine these theoretical predictions in real-world contexts, exposing the extent to which verification strategies align with model expectations and offering a basis for further refinement of its underlying assumptions.

The interviews seem to affirm the conceptual distinction between misinformation and disinformation, with participants emphasising the role of intent as the key differentiator between the two phenomena. One participant underscored the deliberate nature of disinformation, describing it as a targeted effort to mislead, whilst misinformation was characterised as a more ambiguous concept, often arising from the mismanagement or miscommunication of information:

> "disinformation in some it's something. What deliberately has been launched campaign right? So there is a purpose in it. There is a agenda. There is a considerate target group, so it's a kind of campaign, disinformation campaign, while misinformation could be just mishandling of information or send through different wrong channels or presented in a wrong way. So it could be so it's a more vague concept, misinformation. While disinformation, it's something done on purpose"
>
> — Interviewee 1

By way of this explanation, the evidence reinforces the notion that disin-

formation requires active intent and coordination, distinguishing it from misinformation, which can stem from errors or misinterpretation. In support of this viewpoint, other participants echoed similar disparities, further clarifying how misinformation can sometimes contain elements of truth, whereas disinformation operates through fabrication and deception:

> "So, in my opinion, misinformation is when you don't have enough information about something which happen in the world or. Or. The story is real, but some facts. Was changed by accident or by or because someone want to change them and so we will consume. Just half of truth and half of false. And if we think about disinformation, in my opinion, it's like all fake information which were created by someone To confuse us."
>
> — Interviewee 2

> "Misinformation is like conscious information with some purposes. Yeah, to to, to lead you wrong way. Or, sorry, it's disinformation. Yeah, and misinformation. It may be some confusion. Yeah, in. I was giving you some fact I I would be sound like that, I don't know. It's it's my understanding it's like that."
>
> — Interviewee 3

Based on the evidence from these responses, the burdens posed by misinformation and disinformation are evidently distinct, aligning with the theoretical postulation that individuals must account for the presence of deceptive actors when evaluating information. In cases of misinformation, the verification effort x is often guided by perceived reliability or personal heuristics, and may require only a moderate level of scrutiny, as the information is frequently partially correct and, for instance, may be cross-checked with additional sources. Conversely, disinformation introduces an adversarial dimension, where the individual must consider the active effort s exerted by malign actors to mislead. Anticipating this strategic manipulation, individuals may rationally escalate their verification effort

x, engaging in a more extensive process to uncover the falsehood and mitigate the associated risks. This interplay reflects a strategic decision-making process, where consumers weigh the cognitive cost of verification against the potential dangers of falling victim to deceptive content. Consequently, in environments saturated with disinformation, the model predicts an increase in the required verification effort, as individuals seek to distinguish credible information from deliberate misinformation.

On the other hand, while the observed differentiation between the concepts of misinformation and disinformation might suggest that individuals adjust their effort levels x accordingly, this understanding does not necessarily translate into corresponding behavioural changes. As one interviewee noted, some individuals may completely disregard whether the information in question is classified as misinformation or disinformation, instead evaluating it based on the subjective value of the content, the perceived trustworthiness of the source, and their own assessment of its relevance:

> "So I could label that something which is so could be disinformation, could be misinformation. I try to really qualify that piece of information, whether it's relevant or trusted, could be trusted or not, and if I identify this information as something that can be trusted, then I basically don't care whether it's information, misinformation or disinformation. I just don't take it into account for my decision making."
>
> — Interviewee 1

In this view, the categorisation of the information does not significantly alter the decision-making process, as the primary concern of the information consumer is on determining whether the content is reliable. This perspective is further reinforced by the approach this interviewee recounted to evaluating sources, de-

scribing it as a process where some sources are automatically excluded from consideration based on pre-established trust criteria:

> "So I put it through certain criteria and if to start with some sources of information I'm not even consider. So they I I don't even look at them because it's. By default, not trusted can't be trusted, and I and I don't, and I don't waste my time on it if I get the information. From the source, I usually tend to trust, right? And it does not correlate with something that I saw, but I expected to see from that sort of information. So it says something different. Then I'm trying to see the purposes why they have changed their mind, why what could be the reason? So I tried to get other information from different sources in respect to that fact to see what could be the purpose of changing the opinion of. So it's a kind of complex complex process from my side. And you know if I. Identify whether it's disinformation or information. For me, the result would be the same. I'm not taking into account, so it's not very important whether it's because it could be camouflaged right under misinformation while dis information. So it's not that important if I not taking into account. So those people who are making this attempt to fool me, then they just can't succeed anyway"
>
> — Interviewee 1

On that note, for some individuals, the distinction between misinformation and disinformation may become largely irrelevant. Instead, their focus may shift to the consistency of the information with their prior expectations or knowledge. If information from a trusted source deviates from what is expected, the individual may engage in a more complex evaluation process, seeking additional sources to understand the discrepancy. In such instances, the label assigned to the content whether misinformation or disinformation holds little significance, as the primary concern remains the trustworthiness and purpose of the information. This approach accentuates a more nuanced method of information verification, where the emphasis is placed on the credibility of the source and the coherence of the information rather than the specific categorisation of the content. The findings

suggest that, for certain individuals, the effort required to assess information is not primarily influenced by whether the information is deemed misinformation or disinformation, but by a broader evaluation of its trustworthiness and alignment with existing knowledge.

From the perspective of the model, however, the effort s exerted by malicious actors still plays a pivotal role. The model posits that, when individuals are confronted with information that is potentially deceptive, the malicious effort to mislead by means of disinformation or other strategic manipulations triggers a corresponding increase in the verification effort x. This escalation in effort reflects the cognitive burden imposed by the recognition of potential deception and the need for more rigorous scrutiny as individuals attempt to discern the true nature of the content. It follows that when information from a trusted source deviates from established expectations or contains inconsistencies, the individual may infer a strategic manipulation effort or sloppiness s, necessitating an increased verification effort x to counteract its potential influence and safeguard against perceptual distortion. In such cases, the model predicts that even in the absence of explicit categorisation as misinformation or disinformation, the inherently adversarial nature of the information itself still drives an intensification of verification efforts, stressing the strategic dynamics at play in the consumption of information.

4.8 Effects of Misinformation and Disinformation

The interviews further reveal how misinformation and disinformation may not only interject into the decision making by individuals by affecting cognitive processes but also by extending to emotional responses, social trust, which may in turn exacerbate devision making. Several participants recounted how exposure to disinformation, particularly during times of heightened geopolitical tension, had profound effects on public perception and personal well-being with one participant remarking:

> "I saw how powerful can be information during the war. We. Had a lot of those informational attacks on On all the society. And it influenced me as well."
>
> — Interviewee 2

Others conveyed comparable concerns, underscoring the psychological toll of persistent exposure to misleading narratives, which not only intensified anxiety and distress but also reinforced a cycle of negativity, influencing perceptions of reality through a lens of fear and apprehension:

> "or example my Facebook stream, yeah, we can see man in use. Which are not faking, but they let's say. Grab your attention to some bad news, for example from frontline or something about what is what is going on in regional. Region of Ukraine or any way it's connected somehow to the War. Yeah. So this is up to me. It's really conscious, dedicated efforts to create some, some picture of the external world in your mind. Which. Lead you to some negative thinking. So in in my experience we have a lot of such kind of information and it is rather about disinformation."
>
> — Interviewee 3

In addition to its psychological impact, disinformation also appears to disproportionately affect vulnerable demographics, particularly the elderly, who may struggle to critically assess information sources. One participant highlighted the challenges posed by misinformation within their own family, describing the difficulties of countering misleading narratives consumed by older relatives:

> "Yeah, I noticed some misinformation from my Mother-in-law Yeah, because I see that she's reading different type of information from some sources. Like Russian propaganda. Yeah. And every time she comes, she started to tell about. We need to Surrender and something about church like she's sensitive. She she don't categorise and she doesn't have filter so we always have some fights about it but which I see it's from elder people. Yeah. And also my grandma. She is on Facebook...." "she reads a lot of information about. Oh, when we choose President. I forgot this word. Election. So now she has a lot of information about current presidents, about ex president, and she talks how bad they are. And so I think, yeah, she consume a lot of information about elections and I'm not sure they're right. But she won't believe me because I am young and she lived her life long life."
>
> — Interviewee 4

Such accounts from the participants underscore the profound emotional influence of disinformation, particularly in times of crisis when false narratives have the potential to amplify fear and confusion. These manipulative campaigns not only create confusion but also magnify emotional responses, often aggravating fear and anxiety, especially when individuals are confronted with deceptive content that aligns with their pre-existing beliefs or concerns. These occurrences highlight that the effects of disinformation extend far beyond individual cognition, shaping collective attitudes, perceptions, and societal dynamics.

Further illustrating the emotional toll, an interviewee also reflected on the impact of advances in technology, inducing a difficulty in distinguishing between disinformation and reliable information:

> "But not everyone understood that it is AI, and even my parents, they share to me this video and said how good that our politicians are saying that. But and I noticed that My parents, they couldn't understand and distinguish which video was created by AI and which was true"
>
> — Interviewee 2

Compounded by the increasing sophistication of technology, this challenge may be contributing to the erosion of ability to recognise when individuals are being exposed to disinformation. In this manner, their capacity to critically evaluate content, even from sources they perceive as trustworthy, including public figures may be obstructed, further distorting individual perceptions of reality, hindering interaction with information, and diminishing the capacity for informed decision-making. Moreover, disinformation has been observed to exploit emotional responses, instilling polarisation and distrust, particularly through AI-generated content and social media manipulation:

> "I about those groups in Facebook. Which, like I remember this example, that they. I saw that like someone I don't know, a bad guy is generating Pictures of our militants Ukrainian militants captured. I'm just from the captivity or something like that. Give me give me money. Oh, like everything connected to that and those kind of AI generated images, They had a lot and lot of shares, likes, et cetera. And I think that for me it is visible that it it is AI but for someone like my mom it is not and they are really playing a lot With the feelings Etcetera. So we also have this like video created. To Oh, how to say it, to divide the country. And. Like, I mean that they choose the topic where they're like, something like someone who work for Russia and they produce a lot of content to support those topics."
>
> — Interviewee 2

Within the realm of the model, such tactics support the theoretical premise that adversarial entities engaging in disinformation exert effort s to deceive, thus increasing the verification cost x for the consumer. As AI-generated content becomes more sophisticated, the cognitive burden required to detect manipulation

rises, further exacerbating the asymmetry between producers of disinformation and information consumers.

Elaborating on informational effects, participants also identified the impact of misinformation on social trust and confidence. For example, an interviewee shared their frustration with the misleading statements of local political figures which may have led to emotional distress and misperceptions of reality:

> "Our Head of the region Luhansk Luhansk region in Ukraine he was sharing on the beginning of the war. He shared information that. 90% of Luhansk region is destroyed. Well, it wasn't true. He said that in like, if you will go to the fact so. The information about 90% of buildings destroyed was about one city but not about whole region. And by sharing this information He also played on emotions of people and That was Not so. I was angry with that because he was official, representative and head of my region, and during that time my parents was in occupation, so I knew that at least my home is Not damaged, and the city also. We had electricity bill during that time, but he said to the media and to everyone that also like their region is without electricity as well. So it wasn't true."
>
> — Interviewee 2

Particularly, this case illustrates that misinformation originating from official sources may have serious repercussions. When information is misrepresented or exaggerated, it may engender heightened anxiety, loss of trust, and distorted understanding of reality. Substantiated by the evidence regarding risk aversion and cognition, theoretical implications suggest that individuals repeatedly encountering misinformation from trusted institutions may recalibrate their verification strategies, either by increasing effort x to cross-check sources or by disengaging entirely owing to distrust.

Similarly, in conjunction to the immediate emotional distress it generates, disinformation was perceived by participants as a tool for systematically eroding confidence, trust, and a sense of security. One interviewee specifically described

how disinformation exploits negative emotions, deliberately constructing misleading narratives that provoke distress and uncertainty:

> "ITypically they propose you just negative information or just. Information which Leads you to. Well, it appeals to some negative feelings of. Yeah, like Sadness or some? Oh, some negative. Yeah. For example, there is a, as they say, about losses on the front line or some some soldier who came back from the frontline, but. He he he did not find some some support from public authority and so on. So on that when you typically say is a they they lead you. Sorry, just. Typically they lead you to some special website. Which which you you never you never seen before. And so, so so far all that is like in fact it creates some. Some feeling or. Yeah, with security or non security there's some. Loss in confidence in trust."
>
> — Interviewee 3

As evidenced by this account, disinformation is not merely a source of misleading content but an active mechanism for shaping psychological responses, fostering distrust, and amplifying societal divisions. The interviewee further reflected on its polarising effects, emphasising how adversarial actors strategically deploy disinformation to fragment communities and diminish collective resilience:

> "Conscious efforts or strategy tactics of our enemy. To generate this confusion in society and which finally needs to well, is this isolation separation of people we believe is that of costly, it has economic impact. We are strong when we are together, we are strong and we can win the war. Frontline and only war. Or we can survive. If you are together so all these tactics, they are oriented to decrease your your level of yeah, belonging to some some group of yeah with some really I don't know need long term purposes or goals and so on you start to isolate you to generate confusion generate some some negative feeling and so on"
>
> — Interviewee 3

These reflections align with the prediction that exposure to disinformation increases the cognitive and emotional burden associated with information processing. In excess of the direct effort x required for verification, individuals must

also contend with the broader strategic implications of disinformation, which may alter perceptions of security and in addition to amplifying cognitive costs, may precipitate shifts in risk preferences and modify the subjective utility associated with being informed. The adversarial nature of such campaigns not only distorts information landscapes but also necessitates adaptive responses from individuals and institutions alike. Moreover, the disinformation campaigns also appear to have tangible effects on decision-making. One participant reflected on how exposure to disinformation may provoke feelings of insecurity and uncertainty, causing indecision:

> "Yeah, you can affect the decision making. Just because people which feel insecure, isolated, sceptical, they cannot take right. This is not. They cannot demonstrate leadership. They cannot take responsibility for decisions and sometimes they prefer to just to avoid decision making They are so uncertain in the future. They are so so unconfident in themselves. They say, OK, maybe it's not the time now that they see it's not time now to take this with a possibility."
>
> — Interviewee 3

Consistent with the model, this observation reinforces the assumption that disinformation not only disrupts critical information assessment but also reduces confidence in decision-making abilities. As exposure to deceptive content increases, individuals may become more susceptible to external influences and less capable of making effective decisions in uncertain situations.

To top it off, the emotional and cognitive impact of misinformation and disinformation has been linked to diminished productivity for some individuals, demonstrating the far-reaching consequences of disinformation beyond information processing, encroaching upon cognitive resources and impairing the ability to engage meaningfully with daily tasks:

> "With the 1st and 2nd year of war. Full scale war in Ukraine. And. My Efficiency on work at work was low. Because for as you told me that. I could. I couldn't if I saw some news bad news. I couldn't concentrate on my work."
>
> — Interviewee 2

The constant barrage of deceptive information, particularly when it elicits strong emotional responses, disrupts cognitive stability, leading to attentional depletion and reduced capacity for sustained focus. In this context, the model predicts that repeated exposure to manipulative content not only intensifies verification efforts but also imposes long-term cognitive costs, further diminishing the ability of individuals to process information efficiently and engage in critical reasoning.

Taking these observations into account, the model posits that the efforts exerted by malicious actors to deceive individuals, whether through disinformation or other manipulative strategies, drive an increase in verification effort x as reflected by the tendency of x to rise with increasingly negative values of s in Figure 4.8. This intensification reflects the cognitive burden imposed by the need to critically evaluate content and recognise potential deception. To reiterate, the model predicts that regardless of whether information is classified as misinformation or disinformation, the manipulative characteristics of such content catalyse individuals to invest additional cognitive resources in verifying its accuracy as misrepresentations are expected. This amplification of effort x, in turn, influences emotional responses, trust in information ecosystems, and overall decision-making processes.

Integrating these findings, the interviews provide empirical validation of the model assumption that information consumption decisions are shaped not only

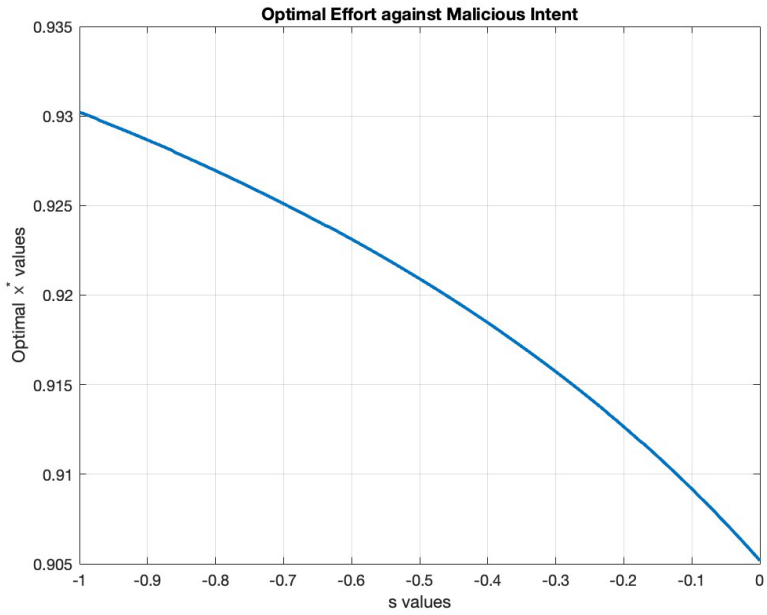

Figure 4.8: Optimal Effort against Malicious Intent

by cognitive constraints but also by adversarial manipulation. Unlike misinformation, which primarily imposes a passive verification cost, disinformation actively amplifies the burden on the consumer, necessitating an increase in verification effort x, while simultaneously increasing cognitive costs, which may erode the benefits derived from the information. This dynamic underscores the strategic nature of information engagement, demonstrating that verification behaviours are contingent upon both the anticipated presence of deceptive actors and the perceived risks associated with exposure to misinformation and disinformation. As a result, individuals must continuously recalibrate their approach to information processing, weighing the costs of verification against the potential consequences of erroneous belief formation.

4.9 Results of the Analysis

The empirical findings from the interviews provide substantial validation for the theoretical model of online information consumption, particularly in relation to the parameters of risk aversion (CRRA utility function), cognitive effort (x), verification cost ($C(x)$), and the strategic efforts of malicious actors (s). Participants consistently articulated an enhanced cognitive burden when engaging with disinformation, corroborating the assumption that deceptive content necessitates an increase in cognitive effort x. Disinformation, as opposed to misinformation, is strategically crafted to mislead, thereby amplifying the need for more intensive assessment. This increase in cognitive effort is reflected in the escalating costs of information verification $C(x)$, where the utility of the information, in terms of its truthfulness and relevance, becomes more difficult to assess due to the emotional and psychological manipulations embedded within the content. The data from the interviews thus lends robust support for the theoretical proposition that disinformation uniquely influences both cognitive and emotional resources, thereby raising the verification costs compared to misinformation.

In terms of risk aversion, the model is integrated within the expected utility framework incorporating a constant relative risk aversion (CRRA) utility function, which accounts for how individuals adjust their behaviour as the perceived risk of engaging with disinformation rises. Participants demonstrated heightened uncertainty and emotional distress when exposed to disinformation, illustrating how cognitive and emotional costs influence their utility function. As cognitive effort x increases, individuals with a higher degree of risk aversion will experience diminishing returns in utility, which incentivises a shift toward information avoid-

ance or disengagement. The CRRA utility function captures this dynamic, as it predicts that individuals will adjust their consumption behaviour depending on the relative costs of verification $C(x)$ and the perceived risk. When cognitive and emotional costs rise, the utility of engaging with information decreases, further reinforcing the tendency to avoid or reduce the effort required to verify information. Accordingly, the model captures the nuanced relationship between risk aversion and the increasing verification costs, aligning with empirical observations of disengagement from high-risk information.

The distinction between misinformation and disinformation, as uncovered by the interviews, further emphasises the postulation that disinformation, due to its targeted manipulation, incurs a higher verification cost $C(x)$. The strategic nature of disinformation, designed to exploit emotional vulnerabilities and societal anxieties, increases the cognitive burden x beyond simple factual verification, incorporating emotional and psychological evaluations that influence decision-making. As malicious actor effort s increases, especially through the use of advanced technologies like AI-generated content, the complexity of distinguishing deceptive information from reliable sources further escalates. This shift towards more sophisticated manipulation techniques corroborates the model's incorporation of malicious actor effort s, which predicts that increasing disinformation complexity raises verification costs and induces deeper cognitive and emotional strain on the information consumer. The expected utility optimisation encapsulates how, as the verification cost $C(x)$ rises with increased effort from malicious actors, individuals with higher risk aversion will perceive an even greater disincentive to engage with potentially deceptive content, leading to lower utility from information consumption.

The data also highlight broader adaptive responses driven by rising verification costs. As verification effort x increases, individuals recalibrate their trust in information sources and engage in defensive strategies, such as avoidance or skepticism, particularly when faced with disinformation. These adaptive strategies are captured in the CRRA utility function, where individuals reduce their exposure to high-risk information when the associated costs $C(x)$ exceed the perceived benefit. This aligns with the assumption that as the verification cost increases, individuals may either increase their effort to verify x or reduce their engagement with information altogether, depending on their risk aversion and perceived utility. The CRRA utility function thus replicates that individuals with higher risk aversion will be more likely to disengage, as the perceived utility of verifying information diminishes. The findings confirm the prediction that individuals adaptively adjust their behaviour based on verification costs, reaffirming the link between verification effort, risk aversion, and disengagement.

Furthermore, the model predicts that as the effort exerted by malicious actors s increases, especially through AI-driven disinformation campaigns, the cognitive and emotional burden x required for verification rises, amplifying the cost function $C(x)$. This non-linear increase in verification costs is consistent with the experiences shared by interviewees, who reported escalating difficulty in distinguishing trustworthy from deceptive content. The interviews underscore the premise of the model that disinformation, particularly when combined with the strategic efforts of malicious actors, not only increases the cognitive load but also manipulates emotional responses, distorting perceptions and decisions. As cognitive effort and emotional distress compound the costs of verification, the expected utility component of the model reflects how individuals adjust their behaviour, with

risk-averse individuals opting for avoidance strategies as the verification effort becomes more demanding. Consistent with the model, the evidence substantiates that that the higher malicious actor effort increases cognitive load and reinforces disengagement strategies among risk-averse individuals.

Thus, the empirical data corroborate the theoretical model by illustrating how cognitive effort x, emotional responses, and malicious actor effort s interact in moulding information consumption decisions. The findings validate the assumption that disinformation, due to its strategic manipulation of both cognitive and emotional resources, incurs higher verification costs $C(x)$ than misinformation, which in turn influences decision-making and societal dynamics. The model effectively captures these dynamics, as it predicts that increasing costs lead individuals, particularly those with higher risk aversion, to reduce engagement with deceptive content, recalibrate their trust in information sources, and ultimately reshape their information consumption behaviour. Taken into account, the results attest to the capacity of the model to predict information consumption patterns, demonstrating the interplay between individual risk preferences, cognitive costs, and emotional manipulation.

Putting it succinctly, the empirical evidence serves to calibrate the theoretical model by informing how the strategic design of disinformation, in conjunction with escalating verification costs, manifests in shifts in both individual decision-making and broader societal behaviour. The findings validate the core theoretical assumptions, particularly the interplay between cognitive effort, emotional responses, and malicious actor efforts, which together shape patterns of information consumption. By highlighting how these factors influence trust recalibration, engagement with deceptive content, and adaptive strategies, the evidence affirms

the overall capacity of the model to encapsulate the complex dynamics at play. Specifically, the insights underscore the ability of the model to capture how rising verification costs, driven by the deliberate manipulation inherent in disinformation, alter the consumption, trust, and processing of information in the digital era. In this way, the qualitative data from the semi-structured interviews reinforce the robustness of the model, providing a comprehensive understanding of how verification costs and emotional responses shape information consumption in the face of disinformation.

4.10 Improving the Position of Information

Following the analysis of qualitative data, which examined the cognitive effort, risk aversion, and perceived distinctions between misinformation and disinformation, and in turn mapped onto the theoretical model of online information consumption, the interviews proceeded to elucidate several nuances regarding the cognitive and emotional challenges individuals face when processing information in the context of online environments. The findings, as discussed with the participants, reveal that these challenges are exacerbated by the manipulation of cognitive biases and the varying epistemic standards among individuals, thereby warranting not only a reconsideration of information consumption patterns but also significant improvements in information production and dissemination. Within the theoretical framework, these insights suggest that the level of cognitive effort x exerted, and the associated cognitive costs $C(x)$ play a critical role in defining the degree to which information is engaged with or dismissed. The effectiveness of information is thus not solely determined by its accuracy but

also by the format, emotional resonance, and the cognitive ease with which it is consumed.

One of the key challenges that emerged from the interviews in enhancing information dissemination stems from the divergence in the epistemic standards of individuals, as content that is considered credible and acceptable was observed to vary significantly across audiences. That is, individuals have different capacities to engage with and interpret content. During the interviews, it was noted that individuals frequently assess information not by its factual accuracy or logical coherence but rather by its format, style, and emotional appeal. This concurs with the theoretical model positing that cognitive costs $C(x)$ influence the perceived value of information, leading individuals to prioritise content that minimises cognitive effort x. For example, an interviewee described an instance where a well-substantiated argument was dismissed in favour of a YouTube video featuring exaggerated visual and auditory stimuli:

> "I was facing situation when person talks on the level which is beyond my, below. My kind of starting point. And when I asked to give me so give me example who. Who makes you believe that it's actually right argument you are giving now and he send me YouTube video where I saw absolutely insane guy with some flashing lights. You know, attention, attention. Important information. So it was like a joke to me. And he says, you know what, what you are talking to me, it's just not really want. I don't understand what you talk about this guy. He's talking language I can understand. And that's why I trust him, because. I understand what he talks. Maybe you think it's too simple or too primitive, but he gets my understanding and it's more important because what you explain, I don't think it's really want what you say or it's make some impact or this absolutely different sphere, but that guy. Says things that are easier to understand in life is easy, and it shouldn't be that complex as you trying to present."
>
> — Interviewee 1

I was facing situation when person talks on the level which is beyond my, below. My kind of starting point. And when I asked to give me

> so give me example who. Who makes you believe that it's actually right argument you are giving now and he send me YouTube video where I saw absolutely insane guy with some flashing lights. You know, attention, attention. Important information. So it was like a joke to me. And he says, you know what, what you are talking to me, it's just not really want. I don't understand what you talk about this guy. He's talking language I can understand. And that's why I trust him, because. I understand what he talks. Maybe you think it's too simple or too primitive, but he gets my understanding and it's more important because what you explain, I don't think it's really want what you say or it's make some impact or this absolutely different sphere, but that guy. Says things that are easier to understand in life is easy, and it shouldn't be that complex as you trying to present.
>
> — Interviewee 1

Essentially, cognitive preferences affect engagement with information, and this dynamic can be interpreted within the model by recognising that individuals seek to minimise cognitive costs $C(x)$, often opting for content that is cognitively effortless rather than analytically rigorous. Consequently, personal biases, prior beliefs, and habitual modes of information consumption may significantly influence the reception of information. When traditional methods of presenting factual content fail to align with these preferences, information may be disregarded irrespective of its accuracy. Furthermore, certain individuals may reject information that conflicts with their pre-existing beliefs, particularly when such information is perceived as a direct attack on their worldview. This may be further complicated by the defensive reactions elicited when information contradicts deeply held beliefs, leading to outright dismissal of otherwise credible sources:

> "I personally never go against the source of information because it's absolutely useless and and then it switch the topic of discussion. So we start. We suddenly discussing the source instead of the matter. So I never do that if he believes in that, then I can help it with his. It's how he receives his information and finds it comfortable, so I can't deal with that. What I would rather do I would consider concerns personal concerns of this person of that individual and what is important? What he cares about what exactly effects His life or his

> decision and what specific, OK, Flashing lights, so on, OK. But what is the message? How they got him, how they got him. So I would rather challenge the message instead of source."
>
> — Interviewee 1

Given the insights drawn from the discussions and their alignment with the theoretical model, information designed to counteract falsehoods must be carefully framed to avoid immediate rejection, particularly when addressing audiences that are predisposed to skepticism or hostility towards certain narratives. As regards the model, individuals with lower cognitive effort x are less likely to engage with complex counterarguments, requiring information to be presented in a manner that minimises cognitive costs $C(x)$ while maximising persuasive impact. Rather than focusing on discrediting sources directly, an interviewee deliberated that the alternative approaches such as challenging the logical consistency of misleading narratives or introducing competing messages that resonate more closely with audience values may prove more effective. This observation is consistent with the implication of the model that information is most effective when it balances cognitive effort x and information costs $C(x)$ against perceived informational benefits.

In addition, the mode of information dissemination may also influence its effectiveness significantly. Different audiences tend to engage with information through distinct platforms, namely, TikTok, Instagram, Facebook, or traditional news outlets, with each fostering unique norms of information processing:

> "The all the channels of information you try to get your audience and everywhere on TikTok in one audience on Instagram, another one Facebook sort one and and so on and so on. And some other communities. You just chat in. I don't know. Economist, New York Times. Whatever. So it's different, different audience. And different context you provide through all those channels."

— Interviewee 1

To combat fragmented information consumption, there is a clear need for tailoring communication strategies to the specific expectations and consumption habits of target audiences. The model reinforces this by illustrating how cognitive costs $C(x)$ vary across platforms, with certain formats such as short-form video content, reducing effort x while increasing engagement. For the purpose of ensuring message accessibility, adapting content to conform to platform-specific communication techniques while maintaining factual robustness is therefore required.

Beyond the methods of information delivery, the interviews highlighted the importance of cultivating emotional intelligence in communication strategies to further enhance the likelihood of message reception. Moreover, emotional framing affects cognitive effort x, as content that resonates emotionally may reduce the perceived costs of processing complex information. Recognising and responding to emotional triggers in audiences can reduce cognitive resistance and foster constructive engagement with an interviewee stressing the importance of structured dissemination approaches:

> "I think we should be much more vigilant and careful about, you know. How to manage information how we disseminate information how to well manage."
>
> — Interviewee 3

Overall, the perspectives shared by the interviewees appear to be reflective of broader calls for proactive endeavours in developing structured approaches for improving public communication and propose several interrelated strategies

for improving information provision. A key consideration is ensuring comprehensibility without oversimplification by adapting information to align with the epistemic baseline of the audience. In addition, engaging with the underlying narratives of misinformation strategically, rather than dismissing it outright, may allow to address cognitive biases and encourage reconsideration. Furthermore, the mode of dissemination should reflect platform-specific norms to improve accessibility while preserving factual integrity. Likewise, narrative framing, particularly through relatable stories and analogies, may facilitate cognitive ease and strengthens emotional resonance, thereby improving message retention. Finally, mitigating defensive reactions remains essential, as framing information in ways that do not directly challenge deeply held beliefs reduces resistance and fosters greater receptivity.

Consequently, the findings accentuate that effective information provision extends beyond being factually accurate. Grounded in the theoretical framework of cognitive effort and its associated costs, information must be structured in a manner that corresponds to audience cognitive capacities and engagement preferences. The analysis of the qualitative through the prism of the model offers a nuanced understanding of these cognitive and emotional dynamics as well as potential approaches to address the challenges of misinformation and disinformation. This advances the comprehension of individual cognitive boundaries, enabling information providers to develop strategies that augment both accessibility and impact, ultimately contributing to more effective communication in an increasingly fragmented information environment.

4.11 Discussion

Building on the complexities of information dissemination in the digital age, this chapter critically examined the dissemination of false information, distinguishing between misinformation and disinformation. Situating this analysis within a formal model designed to capture the strategic interactions underlying information consumption when verification incurs cognitive and opportunity costs, this chapter explored how the nature and consequences of falsehoods vary depending on their origins and intent. To contextualise this distinction, while misinformation is generally characterised by its sporadic and unintentional nature, disinformation is systematically deployed in environments where its impact is significantly magnified by prevailing socio-political conditions. Specifically, contexts marked by conflict, political unrest, and electoral competition offer fertile ground for the proliferation of disinformation, as such settings are often characterised by heightened uncertainty, deep societal divisions, and competing, often contradictory, narratives. The deliberate nature of disinformation, coupled with its ability to exploit informational asymmetries and psychological vulnerabilities, renders it an extraordinarily potent tool for shaping public opinion, steering political discourse, and advancing strategic objectives. These inherent characteristics not only distinguish disinformation from other forms of false information, such as inadvertently propagated misinformation, which, although may lead to serious consequences, lacks the strategic precision and intentionality of disinformation, but also emphasise its profound implications for decision-making, making it a critical phenomenon that demands rigorous examination.

Corroborating the more episodic nature of misinformation, the ongoing war

in Ukraine serves as an exemplary case study, emphasising the pervasive and strategically central role of disinformation in contexts susceptible to manipulation. The findings derived from the analysis underscore that disinformation is not merely a peripheral or ancillary feature of contemporary warfare but also a central strategic instrument deployed with precision and intent. The Ukrainian case study vividly illustrates how disinformation influences not only immediate tactical decisions, such as the coordination of military operations, including drone strikes, ambushes, and other critical engagements, but also long-term strategic objectives, such as destabilising public trust, shaping international narratives, and manipulating policy responses. In such an environment, the systematic and relentless deployment of disinformation further exacerbates the already formidable challenges associated with information verification, thereby distorting decision-making processes and eroding public confidence in credible sources of information. In this regard, the Ukrainian case study offers invaluable insights into the complex dynamics of information consumption in contexts where disinformation is pervasive, strategically orchestrated, and continuously deployed.

In such a high-stakes political context, the consequences of acting on false information may extend to matters of personal safety, health, and even survival, thereby intensifying the implicit willingness to pay, whether in time, cognitive effort, or foregone opportunities, to obtain credible information whose expected utility exceeds that of unverified content. Under the CRRA utility specification employed in the model, higher risk aversion magnifies the perceived disutility of such losses, implying that individuals require a proportionally greater informational benefit to restore expected utility. This pattern is consistent with the observed behaviour of users in conflict settings, where verification effort func-

tions as a form of risk mitigation against potentially catastrophic outcomes. By contrast, in lower-stakes environments such as everyday consumer decisions or entertainment content, the same structure of trade-offs persists, but the magnitude of perceived loss is smaller, leading to a lower optimal verification effort.

Effectively, conflict-affected informational environments such as that of the war in Ukraine provide a robust foundation for calibrating and validating the model by assessing the optimal levels of effort, as predicted by the model, against the actual efforts expended in verifying information, thereby advancing the understanding of how individuals engage with and navigate the complexities of information in environments prone to widespread informational manipulation.

To begin the calibration of the model, a mixed-methods approach, as outlined in the literature Yin [2011], was selected, acknowledging the limitations of available data and the necessity of a more granular investigation of information verification processes, extending beyond the explanatory scope of quantitative methods alone. By adopting this methodological approach, qualitative and quantitative techniques are integrated, facilitating a comprehensive analysis of the strategic interactions underlying information consumption. The qualitative component, derived from semi-structured interviews, provides an in-depth understanding of the cognitive and behavioural processes individuals employ when engaging with disinformation, while the quantitative model, informed by game theory, captures decision-making mechanisms within environments characterised by information asymmetry and uncertainty. By synthesising the two approaches, the model is firmly grounded in both theoretical principles and empirical observations, enabling the calibration of parameter values and evaluating whether the outputs of the model are within the defined boundaries. Additionally, it allows for a struc-

tured assessment of the trade-offs in information consumption established in the model, determining whether they are both realistic and contextually appropriate. Effectively, the mixed methods approach not only strengthened the theoretical foundations of the model but also generates valuable insights into the optimal levels of effort individuals allocate to verifying information, as demonstrated in the Ukrainian case study, bridging the gap between abstract game-theoretic constructs and the complexities of real-world disinformation dynamics.

Following from the methodological framework, the qualitative phase involved systematic coding and thematic analysis of interview data to extract key patterns in information verification behaviours. This process ensured that the model's assumptions reflected the decision-making strategies employed by individuals in real-world disinformation contexts. The integration of qualitative and quantitative methods facilitated a robust analysis by creating a feedback loop between empirical data and theoretical modelling. The qualitative insights were critical in fine-tuning the assumptions embedded in the game-theoretic model, ensuring that they were not only theoretically sound but also grounded in the practical realities of disinformation verification. To be specific, the identified themes informed the formulation of strategic interaction scenarios in the game-theoretic model, aligning theoretical constructs with empirically observed behaviours. In parallel, the quantitative component leveraged these insights to define parameter values and establish equilibrium conditions, facilitating an examination of how variations in verification effort influence decision outcomes. By iteratively refining the model through empirical validation, the approach ensured that theoretical predictions remained anchored in behavioural evidence, reinforcing the applicability of the model to complex online environments. This iterative process, combining the

flexibility of qualitative data with the precision of quantitative modelling, provides a comprehensive view of the decision-making landscape, exploring strategic interactions under varying levels of information uncertainty and illuminating the cognitive and behavioural factors shaping these decisions. Ultimately, the mixed-methods approach offers valuable insights into the complex interplay between individual choices and disinformation dynamics in a conflict context.

As the primary qualitative data instrument, the study employed semi-structured interviews, with the interview questions tailored to elucidate the intricacies of cognitive, strategic, and behavioural approaches to information verification. Guided by the qualitative research framework, the questions were meticulously designed to probe the multifaceted nature of information engagement across both professional and personal contexts. Not only was the design intended to gather data for testing the assumptions of the model concerning the efforts, costs and constraints of cognition, but also sought to elicit the insights and lived experiences of the participants in a manner avoiding the imposition of preconceptions or biases. In this endeavour, rather than steering interviewees, the interviews aimed to establish an open, non-directive environment, allowing participants to articulate their challenges and realities, thereby fostering a deeper, more comprehensive understanding of the environment in which information verification occurs. This exploratory approach facilitated the collection of rich, contextually grounded data, which enabled a more precise tuning of the model parameters and an enhanced alignment with real-world decision-making processes. Emphasis was placed on uncovering the cognitive burdens and institutional pressures such as organisational protocols, time constraints, and socio-cultural influences, that affect verification efforts, alongside the varying degrees of strain participants

experience in these processes. In this fashion, the interview design ensured the study remained attuned to the dynamic and evolving complexities of the informational landscape as regards the emerging issues and impacts of misinformation and disinformation. This, in turn, reinforced the relevance the model bears to contemporary disinformation dynamics, strengthening both its empirical foundation and theoretical robustness.

In the unfolding analysis, the results from the interviews were systematically mapped into the quantitative model of information consumption, revealing key boundaries regarding the cognitive costs and efforts in the face of disinformation campaigns, thereby strengthening its applicability for assessing real-world decision-making in the domain of information assessment. Through the interviews, key themes emerged that were critical to the model, including risk aversion, cognition and the cost of information, as well as distinctions between misinformation and disinformation. Additionally, broader inquiries into the potential improvements in information quality and the uncertainty inherent to information consumption were addressed. The model proved effective in validating the fundamental trade-offs, particularly the balance between the perceived benefits of accessing accurate information and the cognitive burdens imposed by the verification process against the backdrop of costs of misleading information. Furthermore, the insights gained from the interviews revealed additional aspects essential for improving the quality and delivery of information, such as communication strategies, the framing of ideas to counter disinformation, and the emotional aspects of information delivery. These findings not only anchors the model in real-world complexities but also illuminates its relevance in informing strategies aimed at improving information verification efforts. Given the connec-

tion between theoretical constructs and practical application, the model provides valuable guidance to incentivise more effective decision-making in the increasingly complex and uncertain informational landscapes, where alleviating cognitive burdens and mitigating verification costs appear to be essential in improving the ability to navigate misinformation and disinformation effectively.

Parameterising the risk preferences, the findings from the ongoing war in Ukraine highlight the dynamic nature of risk aversion in online information consumption. Specifically, as the conflict unravelled, individuals initially exhibited strong engagement with information, driven by the perceived urgency and necessity of staying informed. Over the course of war, however, many participants shifted towards avoidance or filtering strategies, adjusting their preferences to mitigate the cognitive and emotional burdens associated with constant exposure to disinformation. These findings demonstrate that risk preferences are not static, but evolve in response to contextual factors, emotional reactions, and perceived stakes. The predictions of the model concur with these observations, as individuals appear to calibrate their information-seeking behaviours according to the anticipated risks and rewards, engaging in greater information verification efforts when they perceive higher potential costs from inaccurate or misleading content. The adaptive nature of the observed patterns underlines the importance of risk aversion in governing the verification efforts individuals are willing to exert, especially in high-stakes, disinformation-prone environments such as the Ukrainian conflict.

As the analysis transitions to the role of cognition and the influence of cognitive costs in the verification of information, it examines how individuals weigh the mental effort required against the potential benefits of acquiring reliable in-

formation. Captured by the transaction cost function, cognitive effort increases alongside costs associated with verification, prompting individuals to adopt less intensive strategies or disengage entirely when the perceived value of the information fails to justify the effort expended. Drawn from the interviews, the empirical evidence substantiates this assumption, demonstrating that cognitive costs may become prohibitively high, leading to disengagement or selective reliance on specific sources to minimise verification efforts. Additionally, the findings reinforce the non-linearity of cognitive costs, as verification appears to be progressively hindered when these costs increase. Effectively, the analysis uncovers that verification costs are aggravated by factors such as cognitive capacity, available resources, and external pressures, including professional obligations and time constraints. In response, individuals apply a variety of tactics to manage and mitigate cognitive load, including selective engagement, reliance on trusted sources, and filtering emotionally charged information. These strategies reflect a dynamic trade-off between the cognitive burden of verification and the imperative to remain adequately informed. Furthermore, the findings corroborate the postulation that verification efforts are dynamic, evolving not only as a function of individual decisions but also external constraints. Having consolidated these factors, the framework offers a comprehensive and robust explanation of the cognitive processes underlying information verification in environments characterised by uncertainty and risk.

Being central to situating risk preferences, cognition, and their associated costs, the investigation delves into the distinction between misinformation and disinformation, which proves essential in understanding the decision-making processes and strategies employed by individuals navigating complex informational

landscapes. Within the model, misinformation, broadly defined, refers to the unintentional spread of incorrect or misleading information, often arising from errors, misinterpretations, or miscommunications. In contrast, disinformation is characterised by a deliberate and strategic effort to deceive, where falsehoods are crafted with the explicit intention of manipulating perceptions and influencing decisions. The malicious intent inherent in disinformation introduces an adversarial dimension to the model, requiring consumers to heighten their verification efforts in order to actively counter the distortions introduced by malign actors. The formulation of the model, therefore, positions exposure to disinformation as an adversarial signal, compelling individuals to adjust their cognitive resources based on the anticipated risks and the malicious intent behind the content.

As the empirical data from interviews affirms, the theoretical framework resonates with the conceptual distinction between misinformation and disinformation, affirming that participants recognise the unintentional errors of flawed reporting or misunderstanding inherent in misinformation and differentiate it from the purposeful, deceitful and coordinated nature of disinformation. However, the results also display that, while the model predicts increased verification effort in response to disinformation, many participants prioritise the overall trustworthiness of the source and the coherence of the information rather than actively categorising the content as misinformation or disinformation. This insight led to a further refinement of the model, integrating the notions that, although individuals may be cognisant of the potential for deception, their verification efforts are also shaped by broader criteria of reliability, source credibility, and prior knowledge. Irrespective of recognising disinformation, deviations from informational expectations, viewed through the lens of the model, trigger inferences of inten-

tional or erroneous delivery of misleading content, thereby inducing heightened verification efforts. In this line of reasoning, the theoretical assumptions about the escalation of verification efforts in response to deceptive content are validated by the findings from the interviews which underline the adaptive nature of verification behaviours. Essentially, the complexity of information verification transcends the binary distinction of misinformation versus disinformation, reflecting a more dynamic and context-dependent decision-making process that incorporates not only the perceived intent behind the information but also the credibility and consistency of the content. This interplay between cognitive effort, trust, and strategic manipulation indicates the significance of these factors in modelling information consumption in environments characterised by uncertainty and the risk of manipulation.

Further probing into the cognitive and emotional impacts of misinformation and disinformation, the findings reveal that exposure to such content has far-reaching effects, influencing decision-making and exacerbating societal division. Evidenced by the recounts of the participants during times of crisis, the emotional responses triggered by misleading information, ranging from anxiety and insecurity to fear, profoundly shape personal perceptions and collective societal attitudes. In such instances, the manipulation of emotional states may drive individuals to invest greater cognitive resources into verifying content, as they seek to guard against potential deception. As the conflict in Ukraine continues, many interviewees described the psychological toll of persistent exposure to fabricated content, particularly when it aligns with existing fears or preconceptions. These reactions, as the model predicts, intensify the emotional burden of information consumption, ultimately leading to confusion and negative thinking. More im-

portantly, such emotional responses compromise decision-making capabilities and foster a breakdown in social trust, especially when false narratives disproportionately impact vulnerable groups, such as the elderly, who may struggle to assess the veracity of information. On the one hand, the model, in this case, underscores the pivotal role of verification effort in mitigating these emotional and cognitive tolls, as individuals are forced to engage more deeply with the content to ensure its credibility, balancing the risks of deception with the psychological costs of exposure. On the other, the results also indicate that these escalating cognitive and emotional costs may be excessive, leading some individuals to disengage from verification efforts altogether, or invest minimal effort in verifying information, as the model also anticipates. This disengagement arises from the overwhelming burden posed by constant exposure to disinformation, pushing individuals to avoid confronting the uncertainty and distress that accompany rigorous verification processes, as they seek to reduce the cognitive strain and emotional distress associated with navigating falsehoods

In line with the cognitive dynamics of the model, the increasing sophistication of disinformation tactics, such as the emergence of AI-generated content, further complicates the ability to distinguish between reliable and deceptive information, thereby amplifying the cognitive burden required for effective verification. This corroborates the prediction that manipulative strategies, whether rooted in disinformation or misinformation, prompt an increase in verification effort as individuals expect deception and thus adjust their cognitive resources accordingly. The interviews highlight that these heightened cognitive efforts not only exacerbate the emotional distress experienced but also reinforce a cycle of mistrust, fear, and heightened vigilance. Furthermore, misinformation originating from official

sources often exacerbates the issue by undermining public confidence and societal cohesion, whereas, more generally, falsehoods disseminated by trusted sources accelerate the breakdown in trust within information ecosystems. In light of the responses from the interviews, the cumulative effect of deceptive tactics appears to impair decision-making on a broader societal scale, given that the emotional and cognitive costs detract from the capacity to process information effectively. This pattern empirically validates the postulation of the model that the adversarial nature of disinformation amplifies both cognitive and emotional burdens, ultimately influencing individual decision-making and broader societal outcomes. The observed behaviour of individuals continuously recalibrating their approach to information processing aligns with the cognitive framework wherein the costs of verification are weighed against the risk of forming erroneous beliefs. However, when the cognitive load becomes exceedingly overwhelming, and in agreement with the model, some individuals may reduce or cease verification efforts, resulting in disengagement and less informed decision-making.

Overall, the empirical findings substantiate the theoretical model of online information consumption, demonstrating the intricate interplay between cognitive effort, verification costs, risk aversion, and the strategic efforts of malicious actors. The data affirm that disinformation, unlike misinformation, is strategically designed to mislead, thereby necessitating increased cognitive effort x and elevating verification costs $C(x)$, particularly as emotional and psychological manipulations further obscure credibility assessment. The model, situated within the expected utility framework and incorporating a CRRA utility function, captures how rising cognitive and emotional costs influence behavioural adaptations, with individuals exhibiting heightened uncertainty and distress when confronted with disinforma-

tion. As verification costs escalate, those with greater risk aversion experience diminishing utility, leading to information avoidance and disengagement, a pattern mirrored in the empirical data. Independent whether explicit or implicit, the distinction between misinformation and disinformation becomes particularly salient as targeted manipulation amplifies the verification burden, reinforcing the model in the assertion that adversarial efforts s by malicious actors or deviating trusted sources exacerbate the difficulty of distinguishing deceptive content from reliable sources. This concurs with the prediction that as the complexity of deception increases, verification costs rise non-linearly, inducing cognitive strain and further disincentivising engagement. The findings also illuminate adaptive responses, as individuals recalibrate trust, adopt defensive scepticism, or disengage entirely when the costs of verification exceed perceived benefits, in accordance with the CRRA utility function.

Beyond its empirical alignment, as the first formal attempt to model information credibility assessment, the model offers a novel contribution by integrating cognitive and emotional costs into an expected utility framework, thereby providing a more comprehensive representation of decision-making under uncertainty in information environments. By adopting Bayesian updating for belief revision and explicitly accounting for the interplay between cognitive effort, risk aversion, and adversarial strategies, the model extends beyond conventional approaches that primarily focus on rational updating mechanisms. This broader perspective enhances its applicability across diverse digital ecosystems, where manipulative strategies continuously evolve. Moreover, being able to anticipate behavioural adaptations in response to rising verification costs underscores the practical value of the model, informing interventions aimed at mitigating the personal and soci-

etal impact of disinformation. Ultimately, the evidence supports the contribution of the model to reflecting and anticipating information consumption behaviours by encapsulating how cognitive and emotional burdens interact with strategic disinformation efforts to reshape decision-making and societal trust. The qualitative data not only reinforce the theoretical assumptions but also emphasises the robustness the model possesses in capturing the evolving challenges of digital information environments, where rising assessment costs driven by manipulative strategies modify trust, alter engagement patterns, and redefine information consumption dynamics.

Explored through the lens of cognitive effort, risk aversion, and the perceived distinctions between misinformation and disinformation, the analysis of qualitative data, integrated into quantitive model through a mixed-methods approach, offers profound insights into the challenges faced by individuals in processing information within online environments. As evidenced by the interviews, these cognitive and emotional obstacles are exacerbated by the manipulation of cognitive biases and the diverse epistemic standards across individuals. Such findings necessitate not only a reconsideration of prevailing information consumption patterns but also significant improvements in the strategies employed for information production and dissemination. The insights drawn from the interviews, when mapped onto the theoretical model of online information consumption, reveal that the cognitive effort x exerted and the associated cognitive costs $C(x)$ are pivotal in determining the degree to which information is engaged with or dismissed. Thus, the effectiveness of information is not solely contingent on its factual accuracy but equally determined by its presentation, emotional resonance, and the cognitive ease with which it is processed.

A notable challenge that emerged from the discussions centres on the divergence in epistemic standards, as individuals demonstrated varied capacities to engage with and interpret content. During the interviews, it became prominent that individuals often assess information not by its factual accuracy or logical consistency but by the format, style, and emotional appeal it conveys. This dynamic aligns with the theoretical assertion of the model that cognitive costs $C(x)$ determine the perceived value of information, leading individuals to prioritise content that minimises cognitive effort x. In this regard, the preference for cognitively effortless content over analytically rigorous material reflects a broader trend where personal biases, prior beliefs, and habitual modes of information consumption significantly influence engagement. When traditional modes of presenting factual content fail to resonate with these preferences, information is often disregarded regardless of its accuracy. Furthermore, as seen in the interview excerpts, information that conflicts with pre-existing beliefs is frequently dismissed outright, particularly when perceived as an attack on personal worldview, underlining the emotional and cognitive complexities that shape information processing.

From a practical standpoint, the findings of the Ukrainian case study underscore the critical importance of framing information in ways that minimise cognitive costs while maximising persuasive impact, particularly when addressing audiences predisposed to scepticism or hostility towards certain narratives. This approach aligns with the theoretical model, which posits that individuals with cognitive or ideological constraints are less likely to engage with complex counterarguments, necessitating a presentation of information that both simplifies cognitive processing and maximises its persuasive potential. The model further highlights the importance of context, revealing how variations in platform-specific

consumption habits influence the effectiveness of message delivery. As demonstrated by the interviewees, different platforms foster distinct norms for information processing, underscoring the need for tailored communication strategies that account for the cognitive and emotional preferences of target audiences. By adapting content to the epistemic baseline of the audience and addressing underlying cognitive biases, such strategies can more effectively combat the challenges of misinformation and disinformation.

Ultimately, the empirical data and theoretical framework converge to emphasise that effective information provision goes beyond mere factual accuracy and must also match the mental and emotional capacities of its audience. The use of a mixed-methods approach, integrating qualitative insights with the quantitative model, significantly strengthens the findings, uncovering the subtleties of how psychological and affective factors influence information consumption decisions. The effectiveness of the model to integrate these dynamics provides a sophisticated framework for understanding how individuals navigate the complexities of online information environments. By explicitly considering cognitive effort, risk aversion, and adversarial strategies, the model enhances the understanding of how information consumption behaviours manifest. Moreover, the competence the model displays in anticipating behavioural adaptations as consequence of rising verification costs reinforces its practical value, offering actionable insights for improving the effectiveness of communication strategies. Thus, the model not only contributes to advancing theoretical knowledge but also offers significant value in addressing the real-world challenges posed by both unwittingly and purposefully inaccurate information, paving the way for more informed and effective communication practices in an increasingly fragmented digital landscape.

Chapter 5

Conclusions

5.1 The Architecture of Subtle Influence: Orchestrating Online Behaviour

In the contemporary digital age, the unprecedented expansion of internet access and digital technologies has transformed the ways in which individuals seek, exchange, and engage with information. Online platforms have substantially broadened the scope of civic participation, enabled transnational knowledge flows, supported real-time access to educational resources, and fostered the emergence of decentralised public discourse. This informational abundance has democratised communication, allowing individuals not only to consume but also to create and disseminate content across diverse media formats. From enhancing public awareness in crisis contexts to facilitating the mobilisation of collective action, the benefits afforded by digital connectivity are profound and wide-ranging. Moreover, the commercialisation of online services has provided users with personalised ex-

periences, tailored recommendations, and convenient access to a plethora of goods and services, further enriching the digital ecosystem. However, these gains are accompanied by a parallel set of structural vulnerabilities. The same infrastructures that enable knowledge-sharing also serve as conduits for pervasive data extraction, intrusive surveillance practices, and the commercialisation of behavioural profiling. The erosion of informational privacy and the opacity of algorithmic curation mechanisms have intensified concerns regarding the manipulation of attention, the shaping of preferences, and the amplification of deceptive content.

Amid these dynamics, the proliferation of misinformation and disinformation has emerged as a salient concern, particularly within environments marked by epistemic uncertainty and high informational volatility. The decentralisation of content production, while empowering, has disrupted conventional epistemic gatekeeping mechanisms, thereby accelerating the circulation of unverified claims and strategically distorted narratives. Traditional filters such as editorial oversight and institutional verification have become increasingly displaced by algorithmically driven forms of content prioritisation, which optimise for engagement rather than accuracy. As a result, individuals are required to navigate an increasingly fragmented and polarised information ecosystem, wherein the boundaries between credible and manipulative content are often obfuscated. These challenges are further compounded by cognitive and temporal constraints that limit the capacity users have for sustained evaluative scrutiny. The frictionless nature of digital information exchange, though efficient, exacts cognitive costs that disproportionately affect users with limited resources or expertise, creating asymmetries in their ability to discern and respond to informational threats.

Empirical observations indicate that, when confronted with informational ex-

cess and ambiguity, individuals tend to adopt cognitively economical strategies for processing content. Heuristics such as reliance on trusted sources, conformity to social norms, or consistency with prior beliefs offer expedient, though imperfect, means of evaluating credibility. These heuristics, while facilitating rapid decision-making, also render users susceptible to confirmation bias, selective exposure, and affective polarisation. Moreover, the prevalence of algorithmically curated feeds reinforces feedback loops that prioritise salience and emotional resonance over evidentiary robustness, thereby deepening epistemic fragmentation. Within this context, individuals frequently resort to pragmatic judgement strategies in lieu of rigorous verification, revealing the tension between informational accessibility and evaluative adequacy. These behavioural regularities highlight the importance of understanding information consumption as a dynamic interplay between cognitive constraints, strategic adaptation, and environmental design, rather than as a static process.

While real-world decision processes often rely on heuristics that simplify cognitive effort, the present model represents their overall effect through a smooth and continuous cost function $C(x)$. In practice, heuristic behaviour could be expressed through a piecewise or stepwise formulation in which verification costs remain relatively constant over certain intervals of effort and then rise abruptly once a cognitive threshold is reached. A step-function representation of $C(x)$ would therefore reflect the presence of rule-based or threshold behaviour but would also make the optimisation problem non-differentiable and analytically intractable. The continuous specification adopted in this thesis functions as a stylised approximation of such locally flattened regions and retains analytical tractability while reflecting the gradual increase in perceived cognitive cost that accompanies

greater verification effort.

To provide a formal representation of these dynamics, this thesis introduced a game-theoretic model that conceptualises the strategic interaction between information consumers and producers under conditions of uncertainty. Central to the model is the optimisation problem faced by individuals when determining the level of effort to invest in assessing the credibility of available information. The framework accounts for the trade-off between the utility derived from accurate knowledge and the cognitive costs associated with its acquisition. In doing so, it offers a structured lens through which to examine observed behavioural tendencies, the diffusion of misinformation, and the susceptibility of users to epistemic manipulation. By modelling the decision architecture underlying content engagement, the analysis contributes to a more nuanced understanding of how strategic behaviours evolve in disinformation-prone environments.

This modelling choice is supported by the qualitative findings where respondents described stable but bounded patterns of verification effort rather than abrupt behavioural changes. Their references to trusting familiar sources or ceasing verification once fatigue was reached are consistent with a cost curve that increases smoothly within realistic ranges of effort. The continuous specification of $C(x)$ therefore provides a practical representation of aggregate heuristic behaviour while preserving the possibility of deriving analytical solutions.

Confronting the escalating and multifaceted challenges inherent within the digital ecosystem, encompassing the pervasive dissemination of illicit and deleterious content alongside salient concerns regarding the safeguarding of user well-being, a spectrum of legislative instruments has been enacted globally to regulate digital content and fortify user protection. In the United Kingdom, the Online

Safety Act 2023 represents a pivotal legislative endeavour to counter a constellation of online harms, including the proliferation of illegal content and material demonstrably injurious to both juvenile and adult demographics. The Act establishes a tiered Duty of Care upon online platforms, categorised according to user volume and functional capacities, with the most expansive services subject to the most exigent obligations, mandating the undertaking of comprehensive risk assessments and the implementation of robust preventative mechanisms designed to impede the dissemination of harmful content. Empowered to enforce these statutory provisions, the Office of Communications (Ofcom), the United Kingdom's communications regulator, wields the authority to impose substantial financial penalties and even instigate the blocking of access to non-compliant digital services.

However, the primary legislative thrust of the Online Safety Act 2023 towards proscribing illegal content has engendered scholarly critique concerning its putative limitations in effectively resolving the ubiquitous and often insidious issues of misinformation and disinformation, particularly during periods of heightened sociopolitical sensitivity such as electoral cycles or public health crises. While the Act incorporates a discrete "false communications offence" targeting the knowing transmission of demonstrably false information intended to induce non-trivial psychological or physical harm, its circumscribed scope and the evidentiary burden of proving malicious intent render it a potentially blunt instrument against the more diffuse and strategically crafted campaigns of misleading information. Furthermore, despite mandating enhanced transparency regarding content moderation protocols, the Act does not explicitly stipulate proactive measures for the specific mitigation of disinformation. The ongoing academic and public discourse

surrounding its capacity to robustly address disinformation, alongside enduring concerns regarding the preservation of fundamental freedoms of expression, remains a critical locus of scholarly inquiry. The current regulatory approach, while structurally robust, does not directly capture the behavioural trade-offs explicated in the game-theoretic model, wherein the cognitive and temporal costs of verification often offset the benefits associated with consuming accurate information. This tension is exacerbated when users form expectations that content may be intentionally misleading or strategically divergent from prior knowledge, further increasing the psychological burden and deterring engagement altogether.

Integrating insights from behavioural and game theoretic model of information consumption offers a potentially fruitful avenue for enhancing regulatory efficacy in this domain. It becomes analytically apparent that user conduct frequently reflects an economisation of cognitive resources, wherein individuals rationally weigh the perceived utility of engaging with content against the cognitive and temporal costs associated with its verification. This theoretical perspective suggests that legislative interventions, and indeed voluntary codes of practice aimed at addressing the salient lacuna in the current Act, might achieve greater efficacy through the incorporation of strategic mechanisms designed to reduce the cognitive load on users, thereby potentially disrupting the propagation of misinformation.

Within this theoretical framework, public policy can be understood as acting upon either the cost or the benefit component of the user decision problem. Measures that improve the accessibility and clarity of verified information or that provide supportive digital infrastructure, effectively reduce the slope of $C(x)$ by lowering cognitive or temporal costs. Conversely, interventions that enhance the

reliability and visibility of trustworthy content increase the perceived benefit from verification and thereby influence the expected utility component of the decision. By linking these parameters directly to behavioural outcomes, the model clarifies how regulatory and technological measures can be used to influence individual verification effort and mitigate the spread of disinformation.

Empirical findings from the model indicate that under high verification costs, especially in emotionally or politically charged settings, users may resort to partial engagement, selective exposure, or even complete withdrawal from information environments, a phenomenon that can undermine informed participation and civic trust. Therefore, regulatory conditions that simultaneously reduce cognitive burden and incentivise more deliberate engagement with content, such as simplified verification cues or the integration of platform-endorsed fact-checking tools, could meaningfully shift user strategies toward more critical consumption. As the Online Safety Act evolves through its phased implementation, incorporating such behavioural insights into its framework could prove crucial.

Implementing behavioural insights in practical policy design is inherently complex because the relevant cognitive cost parameters vary across users, cultural contexts, and digital platforms. Empirical identification of such parameters requires detailed behavioural data that are not easily observable in real time. Policymakers may therefore need to rely on experimental evidence, pilot programmes, or adaptive feedback mechanisms to estimate how interventions alter the effective shape of $C(x)$ and change verification effort. Continuous evaluation and recalibration would be essential to align theoretical expectations with observed behavioural responses.

Complementary to the legislative architecture embodied by the Online Safety

Act 2023, the AGENCY project constituted a significant research undertaking, strategically funded by the Strategic Priority Fund on Protecting Citizens Online of UKRI. This triennial, multidisciplinary initiative rigorously examined the premise that the enhancement of online agency, operationalised as the amplification of user autonomy and control within digital environments, represents a critical mechanism for the endogenous mitigation of digital safety risks for citizenry. Through a comprehensive program of empirical investigation and theoretical modelling, the project aimed to generate actionable insights and develop implementable frameworks designed to empower users and foster a more resilient information ecosystem. The methodological scope encompassed the formulation of a policy playbook advocating for 'agency by design' principles for technology provision, the instantiation of user-centric resources such as the FemTech Shield, and the scholarly analysis of collective digital agency as a strategic response to online misinformation.

The empirical and analytical outputs of this research offer a substantial contribution to the scholarly and policy discourse surrounding the effective regulation of online harms, including the persistent challenges of misinformation and disinformation that the Online Safety Act seeks to remedy and resolve. The systematic investigation of user behaviour under conditions of asymmetric information and varying cognitive costs of verification, coupled with the development of innovative tools such as automated fact-verification methodologies and AI-driven content detection systems, provides empirically grounded insights directly relevant to refining regulatory strategies. Furthermore, the sustained focus AGENCY cast on the empowerment of digitally vulnerable populations and the analysis of regulatory arbitrage opportunities in emerging technological domains, such as the

Internet of Things, underscores the interconnectedness of diverse vectors of online harm and the consequent imperative for a holistic approach to digital safety governance. Consequently, the evidence-based recommendations and user-centric frameworks developed through this research represent a valuable input for policymakers and regulators seeking to enhance the efficacy and societal welfare impact of the Online Safety Act in fostering a more resilient and trustworthy digital public sphere.

Another significant policy framework in the international landscape is the Digital Services Act (DSA) of the European Union, enacted in 2022, which imposes specific obligations on Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) to address the systemic risks associated with disinformation. The DSA mandates regular risk assessments and the implementation of mitigation strategies, such as crisis response mechanisms, content removal protocols, and enhanced transparency measures, to limit the amplification of false or misleading content. These obligations are supplemented by a voluntary Code of Practice, which encourages platforms, advertisers, and fact-checkers to adopt measures including the demonetisation of disinformation, increased transparency in political advertising, and the empowerment of users and independent verifiers. Although the DSA enhances regulatory oversight and encourages accountability in content moderation, its reliance on a self-regulatory framework and the absence of a precise legal definition of disinformation, focused primarily on the intent and potential harm of deceptive content, introduce interpretative flexibility that may compromise enforcement consistency.

From a game-theoretic perspective, the DSA primarily addresses the institutional and structural constraints that shape platform behaviour, aiming to realign

incentives through regulatory pressure, transparency obligations, and potential reputational costs. However, it does not explicitly account for the decentralised, micro-level strategic behaviours of users who interact with content under cognitive limitations and asymmetries of information. Users may be conceptualised as boundedly rational agents who evaluate the perceived utility of engagement against the cognitive and opportunity costs of verification. In the absence of mechanisms that lower these costs or reshape user payoff structures, misleading content may continue to proliferate, particularly when it resonates with pre-existing beliefs or yields emotional or cognitive gratification. The model demonstrates that in environments prone to disinformation, especially where source reliability is uncertain or inconsistent, users may strategically disengage, become more sceptical, or avoid content altogether. Emotional responses such as distrust, sadness, and fatigue often exacerbated when individuals must repeatedly verify information in both professional and personal settings may reinforce avoidance behaviours.

These insights underscore the need for regulatory frameworks that are not only punitive or reactive but also anticipatory, embedding conditions that facilitate user comprehension and verification while promoting media literacy across generational and technological divides. Consequently, while the DSA introduces necessary constraints on platforms, it may be insufficiently granular in capturing the endogenous dynamics of user decision-making and the informational externalities that sustain disinformation flows. Incorporating behavioural-game-theoretic insights into the regulatory framework could enable more accurate modelling of user responses and facilitate the design of targeted interventions, thereby enhancing the capacity DSA has to consider the evolving tactics of disinformation

dissemination in complex platform-mediated environments.

Shifting the focus to the protection of vulnerable demographic, the United States has passed the Kids Online Safety Act (KOSA) which aims to protect children online by requiring internet service platforms to default to the highest possible privacy settings for minors, provide tools for parents to safeguard their children, and facilitate the reporting of harmful content. While these measures are commendable in protecting vulnerable populations, KOSA does not directly tackle the complex issue of disinformation and its impact on user behaviour. Applying the game-theoretic model developed herein, it may be analysed how users, especially minors, interact with potentially harmful content and make decisions regarding the effort they invest in verifying information. Given that verification is both cognitively taxing and time-consuming, younger users, who may lack advanced media literacy or are less familiar with source credibility heuristics, may be particularly vulnerable to persuasive but misleading content. This suggests the need for interventions that not only restrict harmful exposure but also equip users with critical engagement strategies aligned with their cognitive development.This analysis could inform legislative strategies that balance protection with the promotion of critical engagement skills among users.

In Australia, a different regulatory emphasis is evident in the Online Safety Amendment (Social Media Minimum Age) Act 2024 seeking to restrict the use of social media by minors under the age of 16. The legislation imposes monetary punishments on social media companies that fail to take reasonable steps to prevent minors from creating accounts on their services. While this measure aims to protect younger users from potential harms, it does not directly tackle the issue of disinformation. Examining this through the lens of game theory, the

influence of such age restrictions on user behaviour can be analysed, particularly regarding the effort invested in verifying information and the potential for minors to seek alternative, less regulated online spaces. Without addressing the cognitive dimensions of information interaction, such restrictions may inadvertently drive users toward platforms with less oversight, where misinformation is more prevalent and verification even more challenging.

A more interventionist approach to online content regulation can be observed in South Korea, where the government has implemented stringent internet censorship laws to combat cyberbullying and the spread of false information. The Korea Communications Standards Commission (KCSC) actively monitors and takes action against content related to cyberbullying, defamation, and the spread of misinformation and has the authority to suspend or delete web postings deemed harmful, illegal, or violating their standards. While these measures aim to maintain online safety, they raise concerns about freedom of expression and the potential for overreach and politically motivated censorship. From a game-theoretic perspective, users may adapt to such regulations by altering their online behaviours, potentially seeking ways to circumvent censorship or adjusting their content-sharing practices. Strategic source selection, scepticism toward platform moderation, and differential trust in domestic versus foreign sources may all shift in response to perceived overreach, introducing new behavioural equilibria that may not align with the intended outcomes of censorship-driven regulation.

Beyond the aforementioned legislative efforts, a multitude of other nations have instituted or are contemplating legislation aimed at fortifying online safety and regulating digital content. In Ireland, the Broadcasting Authority of Ireland has been superseded by Coimisiún na Meán, established under the Online Safety

and Media Regulation Act 2022. This commission is charged with overseeing online safety and media regulation within the country. Similarly, Regulatory Authority for Audiovisual and Digital Communication in France, known as Arcom, plays a central role in supervising online content and ensuring compliance with national standards. In South Africa, the Film and Publication Board serves as the regulatory authority, addressing online harms and promoting a secure digital environment. Enacted as Online Safety Act, No. 9 of 2024, Sri Lankan legislative framework has established the Online Safety Commission, a body vested with substantial authority to evaluate and expunge prohibited online content, a mandate authorities frame as crucial for handling cybercrimes and ensuring national stability. However, this development has elicited salient perspectives highlighting the potential for undue restrictions and infringement upon fundamental freedom of expression.

While the statutory regimes and regulatory bodies address misinformation to varying degrees, a notable lacuna often lies in their specific and comprehensive mechanisms for tackling the nuanced challenges of disinformation, particularly its sophisticated manipulation through AI and its rapid cross-border dissemination. Furthermore, their enforcement capabilities and harmonised international cooperation may be insufficient to effectively counter the global and often deliberately obfuscated nature of contemporary misinformation campaigns. Finally, a consistent and robust emphasis on media literacy and critical thinking education as a preventative measure, rather than solely reactive regulation, is frequently underdeveloped within these frameworks.

Integrating these legislative frameworks with the game-theoretic model of content verification introduced herein provides critical insights into the behavioural

dynamics underpinning the consumption and dissemination of disinformation as well as the efficacy of online content regulations. By accounting for the cognitive costs associated with information verification, the model suggests that legislative efforts may be more effective if they incorporate mechanisms that not only deter the spread of harmful content but incentivise critical engagement with content. As the model demonstrates, users adjust their verification efforts in response to perceived costs, benefits, and expected source reliability. Where verification becomes prohibitively costly or the perceived benefits of accuracy are low, engagement may diminish entirely, reinforcing information silos or apathy. These effects are particularly pronounced when users are repeatedly exposed to misleading content across multiple contexts encompassing professional, social, and personal settings, thereby eroding trust, increasing emotional fatigue, and diminishing motivation to scrutinise further information.

Drawing upon the behavioural dynamics highlighted by the model,, regulatory paradigms may transcend the conventional focus on penalising online platforms for the propagation of inaccurate information towards incentivising the development of tools that reduce the cognitive effort required for users to assess information credibility. This might include visual salience cues, AI-assisted content clustering based on source reliability, or incentive-compatible nudges that reward interaction with verified information. Aligning the incentives of information producers and consumers with broader societal goals, such as reducing disinformation, could lead to more effective and adaptive regulatory frameworks inducing more deliberative, cognitively efficient, epistemically sound, and socially responsible decision-making online.

Despite the significant strides offered by current legal instruments such as

the Online Safety Act, the Digital Services Act and others in mitigating online harms, their effectiveness can be enhanced by integrating insights from behavioural science and game theory. By addressing the cognitive and strategic factors influencing user engagement with online content, these policies can be refined to better combat misinformation and disinformation, fostering a more informed and resilient digital public sphere. In this fashion, not only would the responsiveness of regulatory mechanisms to evolving threat landscapes be improved but also support the development of intervention strategies that are both scalable and context-sensitive. Accordingly, policy frameworks may more effectively match the behavioural dynamics of optimal information consumption, thereby promoting trust, accountability, and democratic resilience in digital environments, contributing to a more informed and resilient online public sphere in the long term.

As these considerations are reinforced by the results from the information consumption model, which indicate that as the strategic efforts of malicious actors intensify, thereby increasing the cognitive burden of verification, individuals, particularly those with higher risk aversion, exhibit a tendency toward information avoidance or disengagement. This highlights the importance of regulatory interventions that extend beyond punitive measures targeting platforms and malicious content producers. Instead, emphasis should also be placed on empowering users by reducing the cognitive costs associated with discerning credible information, ultimately fostering sustained engagement with high-quality content in increasingly complex digital ecosystems.

5.2 The Calculus of Credibility: A Concluding Synthesis

The global impact of misinformation and disinformation has become increasingly evident in recent years, with far-reaching consequences that extend across public health, democratic integrity, and social stability. A prominent example is the spread of false accounts surrounding COVID-19, which significantly contributed to vaccine hesitancy, exacerbating the toll of the pandemic and resulting in preventable deaths. Similarly, election interference campaigns, which utilise disinformation tactics, have undermined democratic processes in several countries, eroding public trust in institutions. For instance, disinformation campaigns were observed in the 2016 US Presidential Election, the 2017 French Presidential Election, and the 2019 European Parliament Election. Compounding this issue, the proliferation of conspiracy theories, particularly those surrounding the 2020 U.S. presidential election, has fuelled political polarisation and even incited violence, as evidenced by the January 6th Capitol attack. The 2022 Russian invasion of Ukraine further demonstrates the pervasive power of disinformation, with both state and non-state actors disseminating false narratives about the origins, progress, and global implications of the conflict to manipulate public opinion, justify military actions, and destabilise international support. The emergence of AI-generated content adds a new layer of complexity to this issue, enabling the mass production and propagation of highly convincing but misleading information at an unprecedented scale. These examples call for a more in-depth understanding of the dynamics governing online information consumption, as well as the development of robust, evidence-based countermeasures. Despite ongoing efforts from

governments, digital platforms, and civil society organisations to combat these threats through fact-checking initiatives, content moderation policies, and media literacy campaigns, a comprehensive theoretical framework remains essential to inform these strategies and address the underlying mechanisms that facilitate the spread of misinformation and disinformation.

To summarise, this thesis has presented a dynamic approach to modelling and analysing online information consumption, distinguishing between misinformation and disinformation, and examining the impact of cognitive and emotional costs on user behaviour. While much of the existing research remains primarily centred on the conceptualisation, deconstruction and characterisation of information evaluation processes and user credibility assessments, the current body of literature lacks a systematically formalised theoretical framework necessary to model the complexities of online information consumption. In answering the central research question, which explores how individuals navigate the complexities of online information consumption and the implications of cognitive and emotional costs on their decision-making processes, the thesis finds that the model effectively captures the strategic interactions that shape information engagement in digital environments. Specifically, the calibration of the model, grounded in the analysis of data from environments saturated with misinformation, robustly substantiates the influence of cognitive effort, perceived risk, and adversarial manipulation on individual information processing. These factors emerged as critical in shaping user responses to online content and are essential for understanding the behavioural dynamics underpinning the spread and reception of misinformation and disinformation.

In the development of a behavioural theory of online harms and information

consumption, Chapter 1 embarks on a comprehensive review of decision-theoretic frameworks to establish a foundation for understanding the behavioural dimension of online decision-making. Drawing on a broad spectrum of economic theories of choice, and engaging critically with the empirical and theoretical challenges that have emerged in response to them, the review ultimately reaffirmed the analytical centrality of the expected utility framework. Despite well-documented paradoxes and behavioural inconsistencies that have inspired a range of alternative formulations, expected utility theory was retained owing to its structural coherence, interpretive flexibility, and continued relevance in representing general behavioural regularities, particularly in aggregate contexts where extreme or anomalous patterns lie beyond the scope of primary concern. This examination is further complemented by a review of contributions from information economics, particularly those concerned with the valuation of information, signal credibility, and information asymmetries. These perspectives were instrumental in shaping an understanding of how users evaluate, prioritise, and respond to information online, especially under conditions of uncertainty and informational overload. Together, these theoretical strands underpin the overarching framework developed in this thesis to capture and formalise patterns of strategic behaviour and information engagement in digital environments characterised by uncertainty and potential harm.

To this end, Chapter 3 introduces the development of a formal model of online information consumption, marking a departure from traditional approaches that primarily focus on descriptive characterisations of user evaluation processes. Laying the general framework for theorising information engagement under uncertainty, the foundational structure provides the basis upon which the construction

of the subsequent model extensions is enabled. By integrating insights from behavioural economics, the model incorporates bounded rationality, cognitive limitations, and belief formation processes, with particular emphasis on Bayesian learning as the mechanism through which individuals update their beliefs in response to new information. The model elucidates how individuals balance the endogenous desire for knowledge acquisition with the associated cognitive costs and potential social standing implications, capturing the tension between epistemic goals and the burdens of information processing and acknowledges the influence of factors such as self-deception and motivated belief formation, processes for which Bayesian updating provides a theoretical underpinning. Thereby, this thesis establishes a foundational framework that facilitates the analysis of the trade-offs individuals navigate when engaging with information in environments characterised by cognitive limitations and socially situated incentives, providing a solid groundwork upon which subsequent model adjustments and extensions may be made to accommodate evolving research needs.

Chapter 4 advances the theoretical framework by focusing on more nuanced misinformation scenarios, specifically delving into the dynamics of disinformation. Distinguishing itself from generalised concept of misinformation usually construed as but not limited to the unintentional spread of falsehoods, disinformation is a calculated tool wielded by malevolent actors, characterised by the deliberate, strategic manipulation of information with malicious intent, often targeting public opinion or political agendas. The model is extended to integrate the role of malevolent actors within the model, expanding its reach to environments where disinformation is purposefully propagated. The incorporation of disruptive actions highlights how such entities strategically destabilise the Bayesian updating

process, impacting how users assess and update their beliefs, particularly in assessing the veracity of information. By considering the costs that such deliberate efforts of distortion by malicious actors impose on individuals, the model is refined to account for the additional cognitive and emotional burdens users face when distinguishing between genuine and intentionally misleading content. This more elaborate approach deepens the analysis of decision-making under uncertainty as individuals navigate online information, considering both the psychological costs of processing and the strategic actions of those spreading disinformation. Through this refinement, the model not only illuminates the strategic dynamics at play in disinformation campaigns, but also offers a more comprehensive framework for examining and understanding user engagement in contexts where deception is deliberately and frequently systematically orchestrated.

Finally, building upon the preceding analysis, Chapter 5 synthesises empirical findings from semi-structured interviews with Ukrainian participants, integrating these insights with a game-theoretic model of information consumption. The chapter investigates how verification efforts are shaped by a complex interplay of cognitive evaluations, emotional reactions, and contextual factors, highlighting how individuals navigate the verification process when faced with persistent disinformation in a high-stakes environment. The mixed-methods approach, combining qualitative data from participant interviews with quantitative elements of the model, provides a nuanced understanding of the decision-making process under information uncertainty. The results reinforce the theoretical framework by demonstrating that verification efforts are influenced not only by perceived intent but also by factors such as source credibility, prior knowledge, and the broader socio-political environment. Notably, emotional responses, particularly in conflict

contexts, escalate verification costs, often leading to disengagement or minimal effort. This escalation suggests that beyond a certain threshold, the cognitive and emotional burdens associated with verification become overwhelming, prompting individuals to reduce their efforts or disengage entirely. Furthermore, the analysis captures how the growing sophistication of disinformation tactics, such as the use of AI-generated content, adds further complexity to information verification, a finding that has significant implications for future research and intervention strategies. Ultimately, the thesis highlights the adaptive nature of information verification behaviours and validates the capacity the model has to predict the increased cognitive and emotional burdens individuals experience before these efforts become prohibitively costly when exposed to potentially misleading content. The model outcomes suggest that the dynamics of online information consumption are influenced not only by strategic intentions but also by emotional and cognitive responses to the broader socio-political context, emphasising the intricate relationship between individual verification behaviours and the evolving tactics of disinformation campaigns.

While the present thesis develops a novel model of online information consumption under uncertainty, the analytical framework retains a user-centric orientation, concentrating on the decision-making process of individuals confronted with potentially deceptive content. The broader information environment, however, entails a complex strategic interaction between multiple actors, whose behaviours are treated as exogenous to the model. Most notably, the actions of malicious agents engaging in the deliberate dissemination of misleading narratives are incorporated as external shocks rather than as outcomes of an endogenous strategic calculus. Although the model implicitly assumes that such actors de-

rive utility from the spread of disinformation, it does not formally account for their strategic optimisation problem, including how they might adjust the intensity or subtlety of deceptive efforts in response to user vigilance or platform-level interventions, nor the potential trade-offs between deception intensity and exposure risk modelled explicitly. This abstraction, while analytically necessary to preserve the tractability of the user verification decision, constrains the scope of insight into how adversarial strategies may evolve in response to changes in user engagement, verification patterns, or institutional responses.

A promising extension of this research would involve representing heuristic behaviour more explicitly through a stepwise or piecewise-flat cost function. In such a setting, $C(x)$ would remain constant over limited intervals of effort, reflecting periods during which additional verification incurs no perceptible cognitive burden. Once a cognitive threshold is reached, effort costs would rise abruptly, creating a pattern that resembles a step ladder. This structure would capture how individuals use shortcuts or rules of thumb in decision-making under uncertainty. However, such a formulation would make the optimisation problem non-smooth and analytically intractable, requiring simulation or numerical techniques instead. The continuous function used in the present model therefore provides a tractable approximation of these behavioural patterns while maintaining interpretability and analytical precision.

Future research could enrich the analytical reach of this framework by endogenising the actions of malicious agents within a fully strategic framework, in which their choices are conditioned on both user effort and the anticipated costs of exposure or removal. Incorporating additional agents such as platforms and content intermediaries, whose algorithmic amplification, moderation policies, and

incentive structures critically shape the flow and visibility of information online, would enable a more holistic treatment of the information ecosystem. Modelling these actors as strategic agents with their own objective functions, such as maximising engagement, managing reputational risk, or minimising regulatory exposure, would allow for a richer understanding of the feedback loops and second-order effects that emerge in multi-agent settings, especially when incentives are misaligned, wherein users, adversaries, and platforms interact under conditions of imperfect information and competing incentives. This expanded framework would open up new analytical avenues related to game-theoretic modelling of signalling, screening, and adverse selection under disinformation, offering greater traction on the structural fragilities of the digital information economy and informing the design of mechanisms and regulations to counter online manipulation.

Although this thesis explicitly incorporates behavioural dimensions, such as bounded rationality and cognition costs into the modelling of user verification, future work could extend this foundation by considering other factors such as motivated reasoning and confirmation bias, embedding these dynamics in more complex institutional and platform-specific contexts. In particular, analysing how platform responses, such as automated content moderation, algorithmic throttling, or verification prompts, interact with user cognition would allow for a more comprehensive understanding of system-wide outcomes. In parallel, a stronger policy and governance lens may be applied by examining the effect that different institutional configurations, ranging from algorithmic transparency requirements and penalties to the introduction of verification subsidies or nudges, bear on strategic behaviour within the system. Investigating the potential for regulatory interventions that realign incentives, reduce verification costs, or internalise

reputational externalities could contribute meaningfully to ongoing debates in digital platform regulation, especially in relation to asymmetric power, user autonomy, and information integrity. For instance, exploring interventions such as algorithmic transparency mandates, subsidised verification tools, or sanctions against persistent amplification of deceptive content may advance theoretical inquiry into the ways in which incentive structures configure strategic behaviour across users, platforms, and influence strategic behaviour across users, platforms, and a broader constellation of actors, including both adversarial and cooperative entities operating within digital information systems.

Pursuant to these considerations, the empirical validation, whether conducted through audit-based platform studies, behavioural experiments, or observational data on disinformation flows, may be significantly enriched by the application of AI-driven techniques. Approaches such as adversarial testing, wherein models are systematically exposed to deliberately manipulated or deceptive data inputs, offer a means of evaluating model robustness under adversarial conditions. Additionally, fine-tuning large language models (LLMs) on disinformation-specific datasets may further improve their capacity to detect subtle and evolving forms of misinformation. The incorporation of real-time data analytics, leveraging the structural dynamics of social networks and content diffusion patterns, may also strengthen the relevance of the model to real-world contexts. Deployed either independently or in combination, these advanced methods may serve to enhance the external validity of the model, anchoring subsequent refinements in empirical complexity and informing the development of adaptive, context-sensitive and evidence-based counter-disinformation strategies.

The broader social consequences of disinformation should be interpreted with

caution. Its effects depend on structural conditions such as the degree of political polarisation, levels of institutional trust, and the design of digital recommendation systems. Future work should consider how these contextual features shape the effectiveness of individual verification strategies and their aggregate impact on information quality. Extending the model to a network or multi-agent framework could provide a more comprehensive understanding of how individual cognitive costs and heuristic behaviours combine to influence collective informational stability.

References

Abbott, A. (2014). *The system of professions: An essay on the division of expert labor*. University of Chicago press. 179

Abelson, R. P. (1985). Decision making and decision theory. *The Habdbook of Social Psychology 1*, 231–309. 36

Acquisti, A. (2004). Privacy in electronic commerce and the economics of immediate gratification. In *Proceedings of the 5th ACM conference on Electronic commerce*, pp. 21–29. 5

Acquisti, A., A. Friedman, and R. Telang (2006). Is there a cost to privacy breaches? an event study. 75

Acquisti, A. and J. Grossklags (2005). Privacy and rationality in individual decision making. *IEEE Security & Privacy 3*(1), 26–33. 6

Acquisti, A., L. K. John, and G. Loewenstein (2013). What is privacy worth? *The Journal of Legal Studies 42*(2), 249–274. 8

Acquisti, A., C. Taylor, and L. Wagman (2016). The economics of privacy. *Journal of Economic Literature 54*(2), 442–92. 4, 5, 50, 51

Ahmad, R., A. Komlodi, J. Wang, and K. Hercegfi (2010). The impact of user experience levels on web credibility judgments. *Proceedings of the American Society for Information Science and Technology 47*(1), 1–4. 95

Ainslie, G. (1992). *Picoeconomics: The strategic interaction of successive motivational states within the person*. Cambridge University Press. 131

Ainslie, G. (2001). *Breakdown of will*. Cambridge University Press. 131

Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in economics*, pp. 235–251. Elsevier. 51, 70

Alexopoulos, C., S. Virkar, M. A. Loutsaris, A.-S. Novak, and E. Loukis (2020). Analysing legal information requirements for public policy making. In *Electronic Participation: 12th IFIP WG 8.5 International Conference, ePart 2020, Linköping, Sweden, August 31–September 2, 2020, Proceedings 12*, pp. 95–108. Springer. 167

Allais, M. (1953). Violations of the betweenness axiom and nonlinearity in probability. *Econometrica 21*, 503–546. 30, 31, 38

Allcott, H. and M. Gentzkow (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives 31*(2), 211–236. 81

Allen, B. (1990). Information as an economic commodity. *The American Economic Review 80*(2), 268–273. 78

Alloy, L. B. and L. Y. Abramson (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology: General 108*(4), 441. 124, 125

Alvarado-Alvarez, C., I. Armadans, M. J. Parada, and M. T. Anguera (2021). Unraveling the role of shared vision and trust in constructive conflict management of family firms. an empirical study from a mixed methods approach. *Frontiers in Psychology 12*, 629730. 163

Alwashmi, M. F., B. Fitzpatrick, J. Farrell, J.-M. Gamble, E. Davis, H. Van Nguyen, G. Farrell, and J. Hawboldt (2020). Perceptions of patients regarding mobile health interventions for the management of chronic obstructive pulmonary disease: mixed methods study. *JMIR mHealth and uHealth 8*(7), e17409. 163

Arazy, O. and R. Kopak (2011). On the measurability of information quality. *Journal of the American Society for Information Science and Technology 62*(1), 89–99. 95

Arrow, K. J. (1962). The economic implications of learning by doing. *The Review of Economic Studies 29*(3), 155–173. 51

Arrow, K. J. (1963). Liquidity preference. *Lecture VI in Lecture Notes for Economics 285*, 33–53. 23

Asch, S. E. (1951). Effects of group pressure on the modification and distortion of judgments. 92

Atkinson, A. B. et al. (1970). On the measurement of inequality. *Journal of Economic Theory 2*(3), 244–263. 42

Bail, C. A., B. Guay, E. Maloney, A. Combs, D. S. Hillygus, F. Merhout, D. Freelon, and A. Volfovsky (2020). Assessing the russian internet research

agency's impact on the political attitudes and behaviors of american twitter users in late 2017. *Proceedings of the National Academy of Sciences 117*(1), 243–250. 87

Bandura, A. (1997). Selfefiicczcy: The exercise of control. new york: Wh in. 123

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology 52*(1), 1–26. 90

Bargh, J. A., M. Chen, and L. Burrows (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology 71*(2), 230. 90

Bargh, J. A., P. M. Gollwitzer, A. Lee-Chai, K. Barndollar, and R. Trötschel (2001). The automated will: nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology 81*(6), 1014. 90

Barnes, S. B. (2006). A privacy paradox: Social networking in the united states. 5

Barth, S. and M. D. De Jong (2017). The privacy paradox–investigating discrepancies between expressed privacy concerns and actual online behavior–a systematic literature review. *Telematics and Informatics 34*(7), 1038–1058. 5, 33

Becker, S. W. and F. O. Brownson (1964). What price ambiguity? or the role of ambiguity in decision-making. *Journal of Political Economy 72*(1), 62–73. 39

Bénabou, R. (2013). Groupthink: Collective delusions in organizations and markets. *Review of Economic Studies 80*(2), 429–462. 132

Bénabou, R. and J. Tirole (2002). Self-confidence and personal motivation. *Psychology, Rationality and Economic Behaviour: Challenging Standard Assumptions 117*(3), 19–57. 122, 123, 124, 125, 126, 128, 129, 131, 132, 135, 136

Bénabou, R. and J. Tirole (2003). Intrinsic and extrinsic motivation. *The Review of Economic Studies 70*(3), 489–520. 124

Berglas, S. and E. E. Jones (1978). Drug choice as a self-handicapping strategy in response to noncontingent success. *Journal of Personality and Social Psychology 36*(4), 405. 127

Berlinski, N., M. Doyle, A. M. Guess, G. Levy, B. Lyons, J. M. Montgomery, B. Nyhan, and J. Reifler (2023). The effects of unsubstantiated claims of voter fraud on confidence in elections. *Journal of Experimental Political Science 10*(1), 34–49. 87

Bernoulli, D. (1954). Exposition of a new theory on the measurement of risk. *Econometrica 22*(1), 23–36. 7

Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical Statistics 24*(2), 265–272. 41

Blessing, L. T. and A. Chakrabarti (2009). *DRM: A design reseach methodology*. Springer. 168

Bloom, N. and J. Van Reenen (2010). New approaches to surveying organizations. *American Economic Review 100*(2), 105–109. 192

Boardman, A. E., D. H. Greenberg, A. R. Vining, and D. L. Weimer (2017). *Cost-benefit analysis: concepts and practice*. Cambridge University Press. 1

Boczkowski, P. J., E. Mitchelstein, and M. Matassi (2018). "news comes across when i'm in a moment of leisure": Understanding the practices of incidental news consumption on social media. *New Media & Society 20*(10), 3523–3539. 88

Bolker, E. D. (1966). Functions resembling quotients of measures. *Transactions of the American Mathematical Society 124*(2), 292–312. 28

Bolker, E. D. (1967). A simultaneous axiomatization of utility and subjective probability. *Philosophy of Science 34*(4), 333–340. 28

Bor, A. and M. B. Petersen (2022). The psychology of online political hostility: A comprehensive, cross-national test of the mismatch hypothesis. *American Political Science Review 116*(1), 1–18. 81

Bortolotti, L. (2009). Delusions and other irrational beliefs. 134

Brenner, M. E. (2012). Interviewing in educational research. In *Handbook of complementary methods in education research*, pp. 357–370. Routledge. 189

Briggle, A., K. Waelbers, and P. Brey (2008). *Current issues in computing and philosophy*, Volume 175. Ios Press. 82

Briley, D. A. and R. S. Wyer Jr (2002). The effect of group membership salience on the avoidance of negative outcomes: Implications for social and consumer decisions. *Journal of Consumer Research 29*(3), 400–415. 91, 134

Broome, J. (1991). Utility. *Economics & Philosophy 7*(1), 1–12. 10

Brown, T. L., R. T. Croson, and C. Eckel (2011). Intra-and inter-personal strategic ignorance: A test of carrillo and mariotti. Technical report, working paper. 133

Brunnermeier, M. K. and S. Nagel (2008). Do wealth fluctuations generate time-varying risk aversion? micro-evidence on individuals' asset allocation. *American Economic Review 98*(3), 713–736. 28

Buchanan, T. and V. Benson (2019). Spreading disinformation on facebook: Do trust in message source, risk propensity, or personality affect the organic reach of "fake news"? *Social Media+ Society 5*(4), 2056305119888654. 87

Buchbinder, M., A. Browne, T. Jenkins, N. Berlinger, and L. Buchbinder (2023). Hospital physicians' perspectives on occupational stress during covid-19: A qualitative analysis from two us cities. *Journal of General Internal Medicine 38*(1), 176–184. 166

Butler, P. V. (2000). Reverse othello syndrome subsequent to traumatic brain injury. *Psychiatry 63*(1), 85–92. 134

Carlin, P. S. (1992). Violations of the reduction and independence axioms in allais-type and common-ratio effect experiments. *Journal of Economic Behavior & Organization 19*(2), 213–235. 30

Carrillo, J. D. and T. Mariotti (2000). Strategic ignorance as a self-disciplining device. *The Review of Economic Studies 67*(3), 529–544. 133

Cavusoglu, H., B. Mishra, and S. Raghunathan (2004). The effect of internet security breach announcements on market value: Capital market reactions for

breached firms and internet security developers. *International Journal of Electronic Commerce 9*(1), 70–104. 75

Chalhoub, G., M. J. Kraemer, N. Nthala, and I. Flechais (2021). "it did not give me an option to decline": A longitudinal analysis of the user experience of security and privacy in smart home products. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–16. 167

Chellappa, R. K. and R. G. Sin (2005). Personalization versus privacy: An empirical examination of the online consumer's dilemma. *Information Technology and Management 6*, 181–202. 8

Chen, S. and S. Chaiken (1999). The heuristic-systematic model in its broader context. 84

Chew, S. H., W. Huang, and X. Zhao (2020). Motivated false memory. *Journal of Political Economy 128*(10), 3913–3939. 132, 134, 135, 136

Choi, W. and B. Stvilia (2015). Web credibility assessment: Conceptualization, operationalization, variability, and models. *Journal of the Association for Information Science and Technology 66*(12), 2399–2414. 95

Clifford Astbury, C., A. Demeshko, E. Gallo-Cajiao, R. McLeod, M. Wiktorowicz, C. Aenishaenslin, K. Cullerton, K. M. Lee, A. Ruckert, A. Viens, et al. (2024). Governance of the wildlife trade and the prevention of emerging zoonoses: a mixed methods network analysis of transnational organisations, silos, and power dynamics. *Globalization and Health 20*(1), 49. 165

Cohn, R. A., W. G. Lewellen, R. C. Lease, and G. G. Schlarbaum (1975). Individual investor risk aversion and investment portfolio composition. *The Journal of Finance 30*(2), 605–620. 28

Connaway, L. S., T. J. Dickey, and M. L. Radford (2011). "if it is too inconvenient i'm not going after it:" convenience as a critical factor in information-seeking behaviors. *Library & Information Science Research 33*(3), 179–190. 94

Couldry, N. and U. A. Mejias (2019). The costs of connection: How data is colonizing human life and appropriating it for capitalism. In *The costs of connection*. Stanford University Press. 93

Creswell, J. W. and J. D. Creswell (2017). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage publications. 169, 170

Cropf, R. A. (2008). Benkler, y.(2006). the wealth of networks: How social production transforms markets and freedom. new haven and london: Yale university press. 528 pp. *Social Science Computer Review 26*(2), 259–261. 93

Culnan, M. J. and P. K. Armstrong (1999). Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organization Science 10*(1), 104–115. 6

Dai, J., H. Chen, and Z. Zhang (2024). How does a historical linguistic landscape influence tourists' behavioral intention? a mixed-method study. *Tourism Management Perspectives 53*, 101294. 165

De Finetti, B. (1972). *Probability, induction and statistics: The art of guessing*. New York: John Wiley and Sons. 29

De Gramatica, M., F. Massacci, W. Shim, A. Tedeschi, and J. Williams (2015). It interdependence and the economic fairness of cybersecurity regulations for civil aviation. *IEEE Security & Privacy 13*(5), 52–61. 161, 162, 173, 174, 175, 177

de Gramatica, M., F. Massacci, W. Shim, U. Turhan, and J. Williams (2017). Agency problems and airport security: Quantitative and qualitative evidence on the impact of security training. *Risk Analysis 37*(2), 372–395. 161, 162, 170, 173, 186

Del Vicario, M., A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi (2016). The spreading of misinformation online. *Proceedings of the National Academy of Sciences 113*(3), 554–559. 81

Dinev, T. and P. Hart (2006). An extended privacy calculus model for e-commerce transactions. *Information Systems Research 17*(1), 61–80. 6

Doctorow, C. (2024). *The internet con: How to seize the means of computation*. Verso Books. 93

Dohmen, T., A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association 9*(3), 522–550. 12, 16

Dunning, D. (2001). *On the motives underlying social cognition*. Wiley Online Library. 132

Eckel, C. C. and P. J. Grossman (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization 68*(1), 1–17. 7

Ecker, U. K., B. K. Sze, and M. Andreotta (2021). Corrections of political misinformation: no evidence for an effect of partisan worldview in a us convenience sample. *Philosophical Transactions of the Royal Society B 376*(1822), 20200145. 81

Eeckhoudt, L. and P. Godfroid (2000). Risk aversion and the value of information. *The Journal of Economic Education 31*(4), 382–388. 59

Eeckhoudt, L., C. Gollier, and H. Schlesinger (2011). *Economic and financial decisions under risk*. Princeton University Press. 19, 23, 44, 48, 59

Eil, D. and J. M. Rao (2011). The good news-bad news effect: asymmetric processing of objective information about yourself. *American Economic Journal: Microeconomics 3*(2), 114–138. 133

Elliott, K., A. Patacconi, J. Swierzbinski, and J. Williams (2019). Knowledge protection in firms: A conceptual framework and evidence from hp labs. *European Management Review 16*(1), 179–193. 161, 162, 177

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics 75*(4), 643–669. 34, 38

Elster, J. and J. E. Roemer (1993). *Interpersonal comparisons of well-being*. Cambridge University Press. 11

Enders, A. M., J. Uscinski, C. Klofstad, and J. Stoler (2022). On the relationship between conspiracy theory beliefs, misinformation, and vaccine hesitancy. *Plos One 17*(10), e0276082. 81

Evans, J. S. B., J. L. Barston, and P. Pollard (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition 11*(3), 295–306. 86

Eysenbach, G. and C. Köhler (2002). How do consumers search for and appraise health information on the world wide web? qualitative study using focus groups, usability tests, and in-depth interviews. *Bmj 324*(7337), 573–577. 94

Fareri, D. S. and M. R. Delgado (2014). Social rewards and social networks in the human brain. *The Neuroscientist 20*(4), 387–402. 89

Fazio, R. H. and M. P. Zanna (1981). Direct experience and attitude-behavior consistency. In *Advances in Experimental Social Psychology*, Volume 14, pp. 161–202. Elsevier. 128

Fennema, H. and P. Wakker (1997). Original and cumulative prospect theory: A discussion of empirical differences. *Journal of Behavioral Decision Making 10*(1), 53–64. 36

Feri, F., C. Giannetti, and N. Jentzsch (2016). Disclosure of personal information under risk of privacy shocks. *Journal of Economic Behavior & Organization 123*, 138–148. 4

Festinger, L. (1954). A theory of social comparison processes. *Human Relations 7*(2), 117–140. 125

Festinger, L. (1962). Cognitive dissonance. *Scientific American 207*(4), 93–106. 90

Finetti, B. d. ([1937]1992). Foresight: Its logical laws, its subjective sources. In *Breakthroughs in statistics*, pp. 134–174. Springer. 29

Fingarette, H. (1985). Alcoholism and self-deception. In M. Martin (Ed.), *Self-Deception and Self-Understanding*. University Press of Kansas. 127

Firat, T. and A. Bildiren (2024). Developmental characteristics of children with learning disabilities aged 0–6 based on parental observations. *Current Psychology 43*(4), 2909–2921. 167

Fischer, P., E. Jonas, D. Frey, and S. Schulz-Hardt (2005). Selective exposure to information: The impact of information limits. *European Journal of Social Psychology 35*(4), 469–492. 86

Fishburn, P. C. (1970). Utility theory for decision making. Technical report, Research analysis corp McLean VA. 11

Fishburn, P. C. (1977). Mean-risk analysis with risk associated with below-target returns. *The American Economic Review 67*(2), 116–126. 42

Flanagin, A. J. and M. J. Metzger (2000). Perceptions of internet information credibility. *Journalism & Mass Communication Quarterly 77*(3), 515–540. 94

Flanagin, A. J. and M. J. Metzger (2011). From encyclopaedia britannica to wikipedia: Generational differences in the perceived credibility of online encyclopedia information. *Information, Communication & Society 14*(3), 355–374. 94

Fletcher, R., D. Radcliffe, D. Levy, R. Nielsen, and N. Newman (2015). *Reuters Institute digital news report 2015: supplementary report*. Reuters Institute for the Study of Journalism. 88

Flintham, M., C. Karner, K. Bachour, H. Creswick, N. Gupta, and S. Moran (2018). Falling for fake news: investigating the consumption of news via social media. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–10. 84

Fogg, B. J. (2003). Prominence-interpretation theory: Explaining how people assess credibility online. In *CHI'03 extended abstracts on human factors in computing systems*, pp. 722–723. 95

Fogg, B. J., C. Soohoo, D. R. Danielson, L. Marable, J. Stanford, and E. R. Tauber (2003). How do users evaluate the credibility of web sites? a study with over 2,500 participants. In *Proceedings of the 2003 conference on Designing for user experiences*, pp. 1–15. 97

Fotopoulou, A., M. A. Conway, M. Solms, S. Tyrer, and M. Kopelman (2008). Self-serving confabulation in prose recall. *Neuropsychologia 46*(5), 1429–1441. 134

Friend, I. and M. E. Blume (1975). The demand for risky assets. *The American Economic Review 65*(5), 900–922. 28

Fritch, J. W. and R. L. Cromwell (2001). Evaluating internet resources: Identity, affiliation, and cognitive authority in a networked world. *Journal of the American Society for Information Science and Technology 52*(6), 499–507. 80

Fritch, J. W. and R. L. Cromwell (2002). Delving deeper into evaluation: Exploring cognitive authority on the internet. *Reference Services Review 30*(3), 242–254. 80

Garrett, R. K. (2009). Echo chambers online?: Politically motivated selective exposure among internet news users. *Journal of Computer-Mediated Communication 14*(2), 265–285. 86

Garrett, R. K. (2017). The "echo chamber" distraction: Disinformation campaigns are the problem, not audience fragmentation. *Journal of Applied Research in Memory and Cognition 6*(4), 370–376. 87

Garrett, R. K., J. A. Long, and M. S. Jeong (2019). From partisan media to misperception: Affective polarization as mediator. *Journal of Communication 69*(5), 490–512. 87

Gerber, N., P. Gerber, and M. Volkamer (2018). Explaining the privacy paradox: A systematic review of literature investigating privacy attitude and behavior. *Computers & Security 77*, 226–261. 6

Gialdini, C., A. R. Michel, M. Romero, S. Ramos, G. Carroli, B. Carroli, R. G. P. de León, M. V. Ortiz, and A. Lavelanet (2024). Multicountry research on comprehensive abortion policy implementation in latin america: a mixed-methods study protocol. *BMJ Open 14*(1), e073617. 167

Gigerenzer, G. and P. M. Todd (1999). *Simple Heuristics that make us Smart*. New York: Oxford University Press. 98, 99

Gilbert, D. T. and J. Cooper (1985). Social psychological strategies of self-deception. In M. W. Martin (Ed.), *Self-Deception and Self-Understanding: New Essays in Philosophy and Psychology*, pp. 75–77. Lawrence: University Press of Kansas. 125

Gilbert, D. T., S. T. Fiske, and G. Lindzey (1998). *The handbook of social psychology*, Volume 1. Oxford University Press. 125, 132

Gilboa, I. and D. Schmeidler (2004). Maxmin expected utility with non-unique prior. In *Uncertainty in economic theory*, pp. 141–151. Routledge. 40

Gilovich, T. (2008). *How we know what isn't so*. Simon and Schuster. 127

Grama-Vigouroux, S., S. Saidi, A. Berthinier-Poncet, W. Vanhaverbeke, and A. Madanamoothoo (2020). From closed to open: A comparative stakeholder approach for developing open innovation activities in smes. *Journal of Business Research 119*, 230–244. 167

Greenwald, A. G. (1980). The totalitarian ego: Fabrication and revision of personal history. *American Psychologist 35*(7), 603. 128

Grinberg, N., K. Joseph, L. Friedland, B. Swire-Thompson, and D. Lazer (2019). Fake news on twitter during the 2016 us presidential election. *Science 363*(6425), 374–378. 87

Guess, A. M. and B. A. Lyons (2020). Misinformation, disinformation, and online propaganda. In N. Persily and J. A. Tucker (Eds.), *Social Media and Democracy: The State of the Field and Prospects for Reform*, pp. 10–33. Cambridge University Press. 87

Guess, A. M., B. Nyhan, and J. Reifler (2020). Exposure to untrustworthy websites in the 2016 us election. *Nature Human Behaviour 4*(5), 472–480. 87

Guiso, L. and M. Paiella (2008). Risk aversion, wealth, and background risk. *Journal of the European Economic Association 6*(6), 1109–1150. 28

Hamal, K. and J. R. Anderson (1982). A note on decreasing absolute risk aversion among farmers in nepal. *Australian Journal of Agricultural Economics 26*(3), 220–225. 27

Handa, J. (1977). Risk, probabilities, and a new theory of cardinal utility. *Journal of Political Economy 85*(1), 97–122. 37

Hannak, A., P. Sapiezynski, A. Molavi Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, and C. Wilson (2013). Measuring personalization of web search. In *Proceedings of the 22nd international conference on World Wide Web*, pp. 527–538. 85

Hanoch, G. and H. Levy (1975). The efficiency analysis of choices involving risk. In *Stochastic Optimization Models in Finance*, pp. 89–100. Elsevier. 41

Hassoun, A., I. Beacock, S. Consolvo, B. Goldberg, P. G. Kelley, and D. M. Russell (2023). Practicing information sensibility: How gen z engages with online information. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–17. 80, 91, 134

Haugstvedt, H. and S. E. Tuastad (2023). "it gets a bit messy": Norwegian social workers' perspectives on collaboration with police and security service

on cases of radicalisation and violent extremism. *Terrorism and Political Violence 35*(3), 677–693. 161, 162, 179, 180

Hayek, F. (1945). The use of knowledge in society. *The American Economic Review 35*(4), 519–530. 50

Heckman, J. J. and Y. Rubinstein (2001). The importance of noncognitive skills: Lessons from the ged testing program. *American Economic Review 91*(2), 145–149. 132

Heider, F. (2013). *The psychology of interpersonal relations*. Psychology Press. 125

Henke, J. B., S. K. Jones, and T. A. O'Neill (2022). Skills and abilities to thrive in remote work: What have we learned. *Frontiers in Psychology 13*, 893895. 167

Hicks, J. R. and R. G. Allen (1934). A reconsideration of the theory of value. part i. *Economica 1*(1), 52–76. 9, 11

Higgins, K. and S. BuShell (2018). The effects on the student-teacher relationship in a one-to-one technology classroom. *Education and Information Technologies 23*, 1069–1089. 169

Hilligoss, B. and S. Y. Rieh (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management 44*(4), 1467–1484. 82

Hirshleifer, J. (1978). The private and social value of information and the reward to inventive activity. In *Uncertainty in economics*, pp. 541–556. Elsevier. 66

Holmström, B. (1979). Moral hazard and observability. *The Bell Journal of Economics 10*(1), 74–91. 51, 76

Holt, C. A. and S. K. Laury (2002). Risk aversion and incentive effects. *American Economic Review 92*(5), 1644–1655. 16

Howe, M. L. and M. H. Derbish (2010). On the susceptibility of adaptive memory to false memory illusions. *Cognition 115*(2), 252–267. 134

Howe, M. L., S. R. Garner, M. Charlesworth, and L. Knott (2011). A brighter side to memory illusions: False memories prime children's and adults' insight-based problem solving. *Journal of Experimental Child Psychology 108*(2), 383–393. 134

Hughes, A. (2019). A small group of prolific users account for a majority of political tweets sent by u.s. adults. 87

Iyengar, S. and K. S. Hahn (2009). Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication 59*(1), 19–39. 86

James, W. (1890). The principles of psychology (cleveland, oh. 123, 124

Jeffrey, R. C. (1965). *The Logic of Decision*. New York: McGraw-Hill. 28

Johnson, T. J. and B. K. Kaye (2000). Using is believing: The influence of reliance on the credibility of online political information among politically interested internet users. *Journalism and Mass Communication Quarterly 77*(4), 865–879. 80

Johnson, T. J., R. Wallace, and T. Lee (2022). How social media serve as a super-spreader of misinformation, disinformation, and conspiracy theories regarding

health crises. In *The Emerald handbook of computer-mediated communication and social media*, pp. 67–84. Emerald Publishing Limited. 87

Joinson, A. N., U.-D. Reips, T. Buchanan, and C. B. P. Schofield (2010). Privacy, trust, and self-disclosure online. *Human–Computer Interaction 25*(1), 1–24. 5

Jones, E. E., F. Rhodewalt, S. Berglas, and J. A. Skelton (1981). Effects of strategic self-presentation on subsequent self-esteem. *Journal of Personality and Social Psychology 41*(3), 407. 128

Kahneman, D. (2013). Thinking, fast and slow. In D. Kahneman (Ed.), *Choices, Values, and Frames*, pp. 1–12. Cambridge University Press. 97

Kahneman, D., P. Slovic, and A. Tversky (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge university press. 97

Kahneman, D. and A. Tversky (1979a). Prospect theory: An analysis of decision under risk. *Econometrica 47*(2), 263–291. 30, 31, 33, 34, 35, 36

Kahneman, D. and A. Tversky (1979b). Prospect theory: An analysis of decision under risk. *Econometrica 47*(2), 263–292. 97

Kihlstrom, R. E., D. Romer, and S. Williams (1981). Risk aversion with random initial wealth. *Econometrica: Journal Of The Econometric Society 49*(4), 911–920. 24

Kim, A. and A. R. Dennis (2019). Says who? the effects of presentation format and source rating on fake news in social media. *Mis Quarterly 43*(3), 1025–1039. 94

Kim, D. (2012). Interacting is believing? examining bottom-up credibility of blogs among politically interested internet users. *Journal of Computer-Mediated Communication 17*(4), 422–435. 95

Kim, K.-S., S.-C. J. Sin, and T.-I. Tsai (2014). Individual differences in social media use for information seeking. *The Journal of Academic Librarianship 40*(2), 171–178. 80

Kimball, M. S. (1989). Precautionary saving in the small and in the large. 65

Klayman, J. and Y.-W. Ha (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review 94*(2), 211. 86

Knight, F. H. (1921). *Risk, Uncertainty and Profit*, Volume 31. Houghton Mifflin. 29

Korn, C. W., T. Sharot, H. Walter, H. R. Heekeren, and R. J. Dolan (2014). Depression is related to an absence of optimistically biased belief updating about future life events. *Psychological Medicine 44*(3), 579–592. 124

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin 108*(3), 480. 132

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics 112*(2), 443–478. 131

Laibson, D. (2001). A cue-theory of consumption. *The Quarterly Journal of Economics 116*(1), 81–119. 131

Lange, O. (1934a). The determinateness of the utility function. *The Review of Economic Studies 1*(3), 218–225. 11

Lange, O. (1934b). Notes on the determinateness of the utility function – iii. *The Review of Economic Studies 2*(1), 75–77. 11

Lasser, J., S. T. Aroyehun, A. Simchon, F. Carrella, D. Garcia, and S. Lewandowsky (2022). Social media sharing of low-quality news sources by political elites. *PNAS Nexus 1*(4), pgac186. 87

LeRoy, S. F. and L. D. Singell Jr (1987). Knight on risk and uncertainty. *Journal of Political Economy 95*(2), 394–406. 29

Levy, J. S. (1992). An introduction to prospect theory. *Political Psychology 13*(2), 171–186. 33, 34

Lin, C.-C., A. Y. Huang, and O. H. Lu (2023). Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. *Smart Learning Environments 10*(1), 41. 163

Lin, I. Y. and A. S. Mattila (2021). The value of service robots from the hotel guest's perspective: a mixed-method approach. *International Journal of Hospitality Management 94*, 102876. 165

Littrell, S., E. F. Risko, and J. A. Fugelsang (2021). 'you can't bullshit a bullshitter'(or can you?): Bullshitting frequency predicts receptivity to various types of misleading information. *British Journal of Social Psychology 60*(4), 1484–1505. 87

Loewenstein, G. and D. Prelec (1992). Anomalies in intertemporal choice: Evidence and an interpretation. *The Quarterly Journal of Economics 107*(2), 573–597. 131

Lucassen, T., R. Muilwijk, M. L. Noordzij, and J. M. Schraagen (2013). Topic familiarity and information skills in online credibility evaluation. *Journal of the American Society for Information Science and Technology 64*(2), 254–264. 95

Lucassen, T. and J. M. Schraagen (2011). Factual accuracy and trust in information: The role of expertise. *Journal of the American Society for Information Science and Technology 62*(7), 1232–1242. 95

MacCrimmon, K. R. and S. Larsson (1979). Utility theory: Axioms versus 'paradoxes'. In *Expected utility hypotheses and the Allais paradox*, pp. 333–409. Springer. 39

MacCrimmon, K. R. and D. A. Wehrung (1990). Characteristics of risk taking executives. *Management Science 36*(4), 422–435. 12

MacKenzie, A. and I. Bhatt (2020). Lies, bullshit and fake news: Some epistemological concerns. *Postdigital Science and Education 2*, 9–13. 87

Mao, J. C. (1970). Survey of capital budgeting: Theory and practice. *Journal of Finance 25*(2), 349–360. 44

Marwick, A. and R. Lewis (2017). Media manipulation and disinformation online. Technical Report 359, Data & Society Research Institute, New York. 87

Maxwell, J. (2008). Designing a qualitative study. 189, 190

McDonald, A. and L. F. Cranor (2010). Beliefs and behaviors: Internet users' understanding of behavioral advertising. In *TPRC 2010*. 5

McKay, R., R. Langdon, and M. Coltheart (2005). Paranoia, persecutory delusions and attributional biases. *Psychiatry Research 136*(2-3), 233–245. 134

McNamara, G., C. Robertson, T. Hartmann, and R. Rossiter (2022). Effectiveness and benefits of exercise on older people living with mental illness' physical and psychological outcomes in regional australia: A mixed-methods study. *Journal of Aging and Physical Activity 31*(3), 417–429. 165

Merriam, S. B. and E. J. Tisdell (2015). *Qualitative Research: A Guide to Design and Implementation*. John Wiley & Sons. 169

Merton, R. K., M. Fiske, and P. L. Kendall (2003). *The focused interview: A manual of problems and procedures*. TPB. 189

Metzger, M. J. (2007). Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology 58*(13), 2078–2091. 95, 97

Metzger, M. J. and A. J. Flanagin (2013). Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics 59*, 210–220. 86, 94

Metzger, M. J., A. J. Flanagin, K. Eyal, D. R. Lemus, and R. McCann (2003). Bringing the concept of credibility into the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. *Communication Yearbook 27*(1), 293–335. 80

Metzger, M. J., A. J. Flanagin, and R. B. Medders (2010). Social and heuristic

approaches to credibility evaluation online. *Journal of Communication 60*(3), 413–439. 82

Metzger, M. J., A. J. Flanagin, P. Mena, S. Jiang, and C. Wilson (2021). From dark to light: The many shades of sharing misinformation online. *Media and Communication 9*(1), 134–143. 87

Mischel, W., E. B. Ebbesen, and A. M. Zeiss (1976). Determinants of selective memory about the self. *Journal of Consulting and Clinical Psychology 44*(1), 92. 133

Mishan, E. J. and E. Quah (2020). *Cost-benefit analysis*. Routledge. 1

Moravec, P. L., R. K. Minas, and A. R. Dennis (2019). Fake news on social media. *MIS Quarterly 43*(4), 1343–A13. 84

Moscati, I. (2013). How cardinal utility entered economic analysis: 1909–1944. *The European Journal of the History of Economic Thought 20*(6), 906–939. 11

Moser-Plautz, B. (2024). Barriers to digital government and the covid-19 crisis–a comparative study of federal government entities in the united states and austria. *International Review of Administrative Sciences 90*(2), 402–418. 167

Mosleh, M. and D. G. Rand (2022). Measuring exposure to misinformation from political elites on twitter. *Nature Communications 13*(1), 7144. 87

Mumtaz, S. and S. Nadeem (2022). Understanding the integration of psychological and socio-cultural factors in adjustment of expatriates: An aum process model. *Sage Open 12*(1), 21582440221079638. 167

Nachman, D. C. (1982). Preservation of "more risk averse" under expectations. *Journal of Economic Theory 28*(2), 361–368. 24

Naranjo-Zolotov, M., O. Turel, T. Oliveira, and J. E. Lascano (2021). Drivers of online social media addiction in the context of public unrest: A sense of virtual community perspective. *Computers in Human Behavior 121*, 106784. 90

Nisbett, R. E. and T. D. Wilson (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review 84*(3), 231. 124

Norberg, P. A., D. R. Horne, and D. A. Horne (2007). The privacy paradox: Personal information disclosure intentions versus behaviors. *Journal of Consumer Affairs 41*(1), 100–126. 8

Nyhan, B. and J. Reifler (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior 32*(2), 303–330. 93

O'Donoghue, T. and M. Rabin (1999). Doing it now or later. *American Economic Review 89*(1), 103–124. 131

O'keefe, D. J. (2015). *Persuasion: Theory and research*. Sage Publications. 94

Özdemir, G. et al. (2020). The effect of social intelligence levels of school principals on their leadership behaviours: A mixed method research. *International Journal of Eurasian Education and Culture 5*(8), 270–300. 164

Paravisini, D., V. Rappoport, and E. Ravina (2017). Risk aversion and wealth: Evidence from person-to-person lending portfolios. *Management Science 63*(2), 279–297. 28

Pareto, V. (1972). *Manual of Political Economy*. London: Macmillan. Edited by Ann S. Schwier and Alfred N. Page. 9, 11

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. penguin UK. 85

Pashler, H. (1998). The psychology of attention. 132

Pennycook, G. and D. G. Rand (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition 188*, 39–50. 93

Pennycook, G. and D. G. Rand (2020). Who falls for fake news? the roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality 88*(2), 185–200. 87

Petersen, M. B., M. Osmundsen, and K. Arceneaux (2023). The "need for chaos" and motivations to share hostile political rumors. *American Political Science Review 117*(4), 1486–1505. 87

Phelps, E. S. and R. A. Pollak (1968). On second-best national saving and game-equilibrium growth. *The Review of Economic Studies 35*(2), 185–199. 131

Pirolli, P. (2005). Rational analyses of information foraging on the web. *Cognitive Science 29*(3), 343–373. 83

Pötzsch, S. (2008). Privacy awareness: A means to solve the privacy paradox? In *IFIP Summer School on the Future of Identity in the Information Society*, pp. 226–236. Springer. 5

Pratt, J. W. (1964). Risk aversion in the small and in the large. *Econometrica 32*(1/2), 122–136. 23, 24

Puri, M. and D. T. Robinson (2007). Optimism and economic choice. *Journal of Financial Economics 86*(1), 71–99. 124

Quiggin, J. (1981). Risk perception and the analysis of risk attitudes. *Australian Journal of Agricultural Economics 25*(2), 160–169. 37

Quiggin, J. (1982). A theory of anticipated utility. *Journal of Economic Behavior & Organization 3*(4), 323–343. 37

Quiggin, J. (2012). *Generalized expected utility theory: The rank-dependent model*. Springer Science & Business Media. 37

Quiggin, J. (2014). Non-expected utility models under objective uncertainty. In *Handbook of the Economics of Risk and Uncertainty*, Volume 1, pp. 701–728. Elsevier. 36

Qureshi, J. A., S. M. F. Padela, S. Qureshi, and S. Baqai (2021). Exploring service brand associations: A consumers' perspective in rising service economy. *Studies of Applied Economics 39*(2), 1–17. 167

Ramsey, F. P. (1931). *The foundations of mathematical and other logical essays*. Routledge and K. Paul. 29

Rialti, R., A. Marrucci, L. Zollo, and C. Ciappei (2022). Digital technologies, sustainable open innovation and shared value creation: evidence from an italian agritech business. *British Food Journal 124*(6), 1838–1856. 167

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology 53*(2), 145–161. 80

Riley, J. G. (2001). Silver signals: Twenty-five years of screening and signaling. *Journal of Economic Literature 39*(2), 432–478. 70

Robbins, L. R. B. (1932). *The nature and significance of economic science*, Volume 2. Macmillan London. 11

Robertson-Lang, L., S. Major, and H. Hemming (2011). An exploration of search patterns and credibility issues among older adults seeking online health information. *Canadian Journal on Aging/La Revue Canadienne Du Vieillissement 30*(4), 631–645. 94

Romer, D. and K. H. Jamieson (2021). Patterns of media use, strength of belief in covid-19 conspiracy theories, and the prevention of covid-19 from march to july 2020 in the united states: survey study. *Journal of Medical Internet Research 23*(4), e25215. 81

Rosen, A. B., J. S. Tsai, and S. M. Downs (2003). Variations in risk attitude across race, gender, and education. *Medical Decision Making 23*(6), 511–517. 15

Ross, S. A. (1973). The economic theory of agency: The principal's problem. *The American Economic Review 63*(2), 134–139. 73

Ross, S. A. (1981). Some stronger measures of risk aversion in the small and

the large with applications. *Econometrica: Journal Of The Econometric Society 49*(3), 621–638. 24

Rothschild, M. and J. Stiglitz (1978). Equilibrium in competitive insurance markets: An essay on the economics of imperfect information. In *Uncertainty in economics*, pp. 257–280. Elsevier. 51, 70, 76

Salancik, G. R. et al. (1977). Commitment and the control of organizational behavior and belief. *New Directions in Organizational Behavior 1*, 54. 124

Samuelson, P. A. (1937). A note on measurement of utility. *The Review of Economic Studies 4*(2), 155–161. 11

Sandmo, A. (1971). On the theory of the competitive firm under price uncertainty. *The American Economic Review 61*(1), 65–73. 27

Savage, L. J. (1954). *The Foundations of Statistics*. New York: John Wiley and Sons. 28, 29, 30

Sbaffi, L. and J. Rowley (2017). Trust and credibility in web-based health information: a review and agenda for future research. *Journal of Medical Internet Research 19*(6), e218. 94

Schacter, D. L. (1996). *Searching for Memory: the brain, the mind, and the past*. New York: Basic Books. 128

Schilling, L. and S. Seuring (2023). Mobile financial service-enabled microbusinesses driving sustainable value creation in emerging markets. *Technological Forecasting and Social Change 192*, 122596. 167

Schmeidler, D. (1989). Subjective probability and expected utility without additivity. *Econometrica: Journal Of The Econometric Society 57*(3), 571–587. 37

Schneider, S. L. and L. L. Lopes (1986). Reflection in preferences under risk: Who and when may suggest why. *Journal of Experimental Psychology: Human Perception And Performance 12*(4), 535. 7

Schoemaker, P. J. (1982). The expected utility model: Its variants, purposes, evidence and limitations. *Journal of Economic Literature 20*(2), 529–563. 9, 12

Schoemaker, P. J. (1990). Are risk-attitudes related across domains and response modes? *Management Science 36*(12), 1451–1463. 12

Scholz-Crane, A. (1998). Evaluating the future: A preliminary study of the process of how undergraduate students evaluate web sources. *Reference Services Review 26*(3/4), 53–60. 94

Segal, U. and A. Spivak (1990). First order versus second order risk aversion. *Journal of Economic Theory 51*(1), 111–125. 23

Seligman, M. E. (2006). *Learned optimism: How to change your mind and your life*. Vintage. 123

Shearer, E. and A. Mitchell (2021). News use across social media platforms in 2020. 84

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics 69*(1), 99–118. 98

Simon, H. A. (1978). Rationality as process and as product of thought. *The American Economic Review 68*(2), 1–16. 32

Slovic, P. and A. Tversky (1974). Who accepts savage's axiom? *Behavioral Science 19*(6), 368–373. 30, 39

Smith, H. J., T. Dinev, and H. Xu (2011). Information privacy research: an interdisciplinary review. *MIS Quarterly 35*(4), 989–1015. 5

Spence, M. (1973). Job market signaling. In J. J. McCall (Ed.), *Uncertainty in Economics*, pp. 355–374. Academic Press. 51, 70, 75

Starbird, K. (2019). Disinformation's spread: bots, trolls and all of us. *Nature 571*(7766), 449–450. 81

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature 38*(2), 332–382. 37

Stets, J. E. and P. J. Burke (2000). Identity theory and social identity theory. *Social Psychology Quarterly 63*(3), 224–237. 90

Stigler, G. J. (1961). The economics of information. *Journal of Political Economy 69*(3), 213–225. 51

Stiglitz, J. E. (1975). The theory of" screening," education, and the distribution of income. *The American Economic Review 65*(3), 283–300. 51, 76

Stiglitz, J. E. (1983). Risk, incentives and insurance: The pure theory of moral hazard. *The Geneva Papers on Risk and Insurance-Issues and Practice 8*(1), 4–33. 51, 73, 76

Stiglitz, J. E. (2000). The contributions of the economics of information to twentieth century economics. *The Quarterly Journal of Economics 115*(4), 1441–1478. 78

Stiglitz, J. E. (2017). The revolution of information economics: the past and the future. Technical report, National Bureau of Economic Research. 97

Stone, E. F. and D. L. Stone (1990). Privacy in organizations: Theoretical issues, research findings, and protection mechanisms. *Research in Personnel and Human Resources Management 8*(3), 349–411. 4

Strotz, R. H. (1973). *Myopia and inconsistency in dynamic utility maximization*. Springer. 131

Sundar, S. S. (2008). *The MAIN model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative Cambridge, MA. 82

Sundar, S. S., H. Kang, M. Wu, E. Go, and B. Zhang (2013). Unlocking the privacy paradox: do cognitive heuristics hold the key? In *CHI'13 extended abstracts on human factors in computing systems*, pp. 811–816. 6

Tajfel, H. (1981). *Human groups and social categories: Studies in social psychology*. Cambridge, UK: Cambridge University Press. 90

Tase, A., M. Z. Ni, P. W. Buckle, and G. B. Hanna (2022). Current status of medical device malfunction reporting: using end user experience to identify current problems. *BMJ Open Quality 11*(2), e001849. 166

Taylor, S. E. and J. D. Brown (1988). Illusion and well-being: a social psychological perspective on mental health. *Psychological Bulletin 103*(2), 193. 126

Taylor, S. J., R. Bogdan, and M. L. DeVault (2015). *Introduction to qualitative research methods: A guidebook and resource*. John Wiley & Sons. 168

Tirole, J. (2009). Cognition and incomplete contracts. *The American Economic Review 99*(1), 265–294. 100, 104, 113, 114, 115, 116, 118, 119, 121, 122

Toma, C. L. and J. T. Hancock (2013). Self-affirmation underlies facebook use. *Personality and Social Psychology Bulletin 39*(3), 321–331. 90

Tormala, Z. L., P. Briñol, and R. E. Petty (2006). When credibility attacks: The reverse impact of source credibility on persuasion. *Journal of Experimental Social Psychology 42*, 684–691. 94

Tormala, Z. L., P. Briñol, and R. E. Petty (2007). Multiple roles for source credibility under high elaboration: It's all in the timing. *Social Cognition 25*, 536–552. 94

Tucker, J. A., A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan (2018). Social media, political polarization, and political disinformation: A review of the scientific literature. 87

Tufekci, Z. (2008). Grooming, gossip, facebook and myspace: What can we learn about these sites from those who won't assimilate? *Information, Communication & Society 11*(4), 544–564. 8

Turel, O. and B. Osatuyi (2021). Biased credibility and sharing of fake news on social media: Considering peer context and self-objectivity state. *Journal of Management Information Systems 38*(4), 931–958. 91

Tversky, A. and C. R. Fox (1995). Weighing risk and uncertainty. *Psychological Review 102*(2), 269. 15

Tversky, A. and D. Kahneman (1974). Judgment under uncertainty: heuristics and biases. *Science 185*(4157), 1124–1131. 99

Tversky, A. and D. Kahneman (1981). The framing of decisions and the psychology of choice. *Science 211*(4481), 453–458. 97

Tversky, A. and D. Kahneman (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty 5*(4), 297–323. 37

Uscinski, J., A. M. Enders, C. Klofstad, and J. Stoler (2022). Cause and effect: On the antecedents and consequences of conspiracy theory beliefs. *Current Opinion in Psychology 47*, 101364. 81

Vaismoradi, M., J. Jones, H. Turunen, and S. Snelgrove (2016). Theme development in qualitative content analysis and thematic analysis. *Journal of Nursing Education and Practice 6*(5), 100–110. 165

van Poelgeest, R., A. Schrijvers, A. Boonstra, and K. Roes (2021). Medical specialists' perspectives on the influence of electronic medical record use on the quality of hospital care: Semistructured interview study. *JMIR Human Factors 8*(4), e27671. 166

Vermaak, M. and H. M. de Klerk (2017). Fitting room or selling room? millennial female consumers' dressing room experiences. *International Journal of Consumer Studies 41*(1), 11–18. 167

Von Neumann, J. and O. Morgenstern ([1944]2007). Theory of games and economic behavior. In *Theory of games and economic behavior*. Princeton university press. 7, 10, 11, 28

Vosoughi, S., D. Roy, and S. Aral (2018). The spread of true and false news online. *Science 359*(6380), 1146–1151. 81, 93

Vraga, E. K. and L. Bode (2020). Defining misinformation and understanding its bounded nature: Using expertise and evidence for describing misinformation. *Political Communication 37*(1), 136–144. 81

Weber, E. U., A.-R. Blais, and N. E. Betz (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making 15*(4), 263–290. 12

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology 39*(5), 806. 126

Wilder, S. (2005). Information literacy makes all the wrong assumptions. *Literacy and Numeracy Studies 14*(1), 69–72. 94

Xiao, X., P. Borah, and Y. Su (2021). The dangers of blind trust: Examining the interplay among social media news use, misinformation identification, and news trust on conspiracy beliefs. *Public Understanding of Science 30*(8), 977–992. 81

Yin, R. K. (2011). *Applications of case study research*. Sage publications. 161, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 175, 177, 180, 188, 190, 191, 248

Young, A. L. and A. Quan-Haase (2013). Privacy protection strategies on facebook: The internet privacy paradox revisited. *Information, Communication & Society 16*(4), 479–500. 6

Yurtseven, N. and U. Akpur (2018). Perfectionism, anxiety and procrastination as predictors of efl academic achievement: a mixed methods study. *Novitas-ROYAL (Research on Youth and Language) 12*(2), 96–115. 163

Yurtseven, N. and S. Dulay (2022). Career adaptability and academic motivation as predictors of student teachers' attitudes towards the profession: A mixed methods study. *Journal of Pedagogical Research 6*(3), 53–71. 165

Zuboff, S. (2023). The age of surveillance capitalism. In *Social theory re-wired*, pp. 203–213. Routledge. 93

Zulman, D. M., M. Kirch, K. Zheng, and L. C. An (2011). Trust in the internet as a health resource among older adults: analysis of data from a nationally representative survey. *Journal of Medical Internet Research 13*(1), e1552. 94, 95