LUISS

Department of Political Science

Bachelor of Arts in Politics: Philosophy and Economics

Course of Behavioural Economics and Psychology

Architectures of Influence: AI-Powered Nudging, Investor Behaviour, and the Future of Financial Markets.

Prof. Giacomo Sillari

Supervisor

Elina Damianidou

Candidate

Academic year 2024/2025

TABLE OF CONTENTS

ABSTRACT

This thesis examines how artificial intelligence (AI) affects behavioural influence in the financial sector through dynamic and hyper-personalised nudge configurations. It builds upon behavioural economics and nudge theory principles, studying the evolution from static and generalised interventions to instantaneous, data-driven choice architectures.

The first part of the analysis examines AI-powered investment platforms such as Robinhood, Betterment, and Wealthfront to explore how such technologies impact financial decision-making at the individual level. It examines how principles such as agency, autonomy, and informed consent are affected by the use of AI, stressing aspects of algorithmic opacity, data asymmetry, and gamification of financial interfaces.

The thesis then expands upon these nudging principles and considers how they can have systemic impacts when amplified across user bases and aggregated through platform design. Case studies explored to assess whether and how AI-enabled behavioural design can produce synchronised behaviours, volatility, or emergent behaviours signalling financial instability.

By bridging the individual and systemic levels, this thesis adds to new research in the intersecting literatures of algorithmic governance and behavioural finance. It scrutinises the implications of ethical and regulatory concerns regarding AI nudging. It ultimately advocates for the reconceptualisation of transparency, accountability and user safety in the development of financial technologies.

Introduction

In the evolving landscape of financial decision-making, artificial intelligence is emerging as a catalyst that alters human behaviour rather than assists. The role of financial intermediaries – traditionally played by financial advisors and institutions – is increasingly being supplanted by algorithmic interfaces (robo-advisors, mobile banking applications, or AI trading platforms). These are not impartial positions; they are designed to inform, persuade, and sometimes even influence decisions. The behavioural nudges, or interventions, materially embedded into the design of these algorithmic interfaces are evidence of how to nudge users into behaviour that is thought to be in their best interest. While nudges have been seen before, particularly in targeted ads and offerings, they have existed as fixed or static outcomes for a particular segment of users over time. However, the emergence of AI-enabled personalisation indicates we can now witness nudging, not as fixed and universal interventions but also dynamic and with user individuality at the centre stage. This opens a new chapter in behavioural economics and financial technology relating to significant and topical concerns related to decision-making, oversight of choice, and systemic stability in the digital future. Nudging was introduced by Richard Thaler and Cass Sunstein in their 2008 book *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Their premise is that the framed choices presented, or choice architecture, can also influence behaviour in a way that does not infringe on freedom of choice. Nudges are not dictums; they rely on cognitive biases and heuristics, which can shape the way individuals frame decisions that substantiate better decision-making. Nudging has been used in aspects of the financial work environment, such as pension enrolment, paying off credit cards, saving, and making investment decisions. Various governments and organisations see the appeal in nudging because of its subversive nature and cost-effectiveness, to the extent of creating behavioural insights teams to nudge at scale. However, traditional nudging has a cost: it is generally static, one-size-fits-all, and slow to respond to user feedback.

The innovation of artificial intelligence and machine learning has changed the landscape of behavioural nudging. With access to vast amounts of user data - transaction histories, browsing behaviour, psychometric dispositional patterns, and risk aversion - artificial intelligence systems can nudge with a high degree of precision. This has led to what some academics have termed ‘hypernudging’ or algorithmic nudging, meaning digital platforms using real-time analytics to tailor decision environments in ways that are often difficult to ascertain for the user. A key example of this is the generic placement of fruit at eye level in a university cafeteria versus an urgent text sent by a service just moments before they are likely to check their bank or savings account, recommending an investment product at the moment of changing market sentiment, or subtly adjusting the colour, frame, or timing of financial nudges/alerts based on prior engagement assessments.

We see the transformation of finance nowhere more than in retail finance. Investment services like Wealthfront, Betterment, and Robinhood have incorporated behavioural cues into their digital interfaces with the oft-stated aim of helping users avoid common behaviour biases related to trading, such as overtrading, loss aversion, present bias, and under-diversification. A robo-advisor might suggest a user shift their investments into low-fee index funds, or an interactive investment app may suggest a user “round up” all purchases to invest spare change. These services often trumpet their product to democratise finance and “help” their clients develop better financial behaviours. Whilst, when well-designed, ethically deployed, and rigorously determined through experimentation, AI-nudging has the potential to help users learn about finance and investing, reduce harmful behaviours related to cognitive biases, and serve the overall goal of long-term investment improvement. However, AI-nudging does present strong risks, both individually and systemically. The fact that a user may begin to overly rely on algorithmic advice perhaps shifts the locus of control and judgement to a non-transparent and opaque algorithm that they do not fully understand the reasoning behind.

Traditional financial advice is often transparent in its intent, data, and resulting ramifications of its recommendations. Some investment enabling service algorithms lack transparency, accountability, and mechanisms to provide options for user feedback or check that the user is on the right track. Algorithmic manipulation can occur, where the nudger has goals that may not be aligned with the users. For instance, an investment service may nudge users towards high-margin products or trading behaviours that may provide high engagement and increased profits for the company. Still, it could take away from the user's long-term interest. This simultaneous boundary between persuasion and manipulation raises ethical issues in nudging, especially if the user does not realise they are being nudged or does not fully understand their financial recommendation.

At a macroeconomic level, it is also possible that widespread use of similar algorithms across platforms may systematically induce financial risk. AI-nudging behaviour can amplify herding behaviour, where large groups of investors are simultaneously nudged to similar strategies, asset classes, or reactions to market signals. Systematic herding during volatile times could exacerbate systemic risk to peak-market busts, like the 2010 Flash Crash or, more recently, with algorithmic trading behaviour exhibited within cryptocurrency markets. The interconnectedness of AI systems moves from portfolio behaviour to shifts in entire financial ecosystems and the diligence of how policymakers and regulators dismantle or strengthen those financial ecosystems on individual and macroeconomic levels. Although much of the literature to this point has explored the efficacy of nudging as an overall improvement to individual behaviour (e.g., increased savings), research has comparatively neglected how algorithmic personalisation, ethical limits of behavioural nudging beyond informed consent, and the broader computation implications for market stability and regulatory matters may best mitigate these risks and improve behaviours overall. This thesis seeks to outline and bridge this gap by exploring the nuances around the AI-nudging, shift of associated behaviours, how AI-nudging differs from traditional nudges, and whether this contributes to a systemised risk.

The thesis is organised around four central research questions to steer this exploration:

1. How do AI-powered nudges affect investor decisions and individual behaviour?

2. In what ways do AI-powered nudges differ from traditional, static nudges?

3. What ethical and regulatory dilemmas arise from AI-powered nudging in finance?

4. Can AI-powered nudging create systemic risks like market fragility, herd behaviour, or loss aversion?

This research uses a qualitative, interdisciplinary approach to explore these research questions and the various components contributing to AI-powered nudging in the financial industry. It incorporates academic literature from behavioural economics, finance, human-computer interaction, AI ethics, financial sector and regulators literature, and lastly, case studies of individual technologies. The methodology is mainly literature-based, hence the qualitative nature, emphasising critical synthesis and comparison. Of special interest is the interaction between user experience design, machine-learning algorithms, and behavioural knowledge, referred to, for this thesis, as the "architecture of influence" in digital finance. The thesis structure will reflect this trajectory of analysis. Chapter 1 discusses the theoretical foundations of AI-powered nudges and their background, beginning with the origins of nudge theory in behavioural economics and expanding into newer literature on algorithmic personalisation and hyper-nudging. Chapter 1 will also introduce ethical and regulatory dilemmas to provide context for the later analysis. Chapter 2 will focus on the micro-level of analysis and how AI-funded nudging systems affect individual investor behaviour across a few digital platforms. Drawing upon case studies and the bulk of empirical literature, Chapter 2 will explore the benefits, limitations, and unintended consequences of digital dialogues towards reordered

strategies. Chapter 3 will address the macro-level, studying the effects of general AI-powered nudging tools on universal settings and challenges to financial regulation. Then, the conclusion synthesises the results, reflects on the ethics and social consequences of AI-powered nudges in general, and suggests future research. The thesis aims to advance interdisciplinary dialogue at the intersection of behavioural economics, AI, and financial regulation. It aims to contribute to a deeper understanding of this phenomenon, as well as both the opportunities and dangers of AI-powered nudging in finance.

CHAPTER 1: THEORETICAL FRAMEWORK

1.1 Nudge Theory and Behavioural Economics

In recent years, contemporary economic thought has progressively recognised that individuals sometimes do not act per the traditional rational choice models. The transition from classical views of rational choice to new financial models largely began with the work of psychologists Daniel Kahneman and Amos Tversky, while creating the field of behavioural economics. This intellectual discipline utilises psychology to provide explanations for the systematic errors associated with humans deviating from rational behaviour. Behavioural economic models do not view the individual as a rational utility maximiser, but as one with cognitive limits, subject to heuristics, feelings, and context.

Nudge theory was introduced by Thaler and Sunstein (2008), who describe nudges as "any aspect of the choice architecture that alters people's behaviour in a predictable way without forbidding any options or significantly changing their economic incentives" (p. 6). Nudge theory does not set out to limit or prevent an individual's options or opportunities for decision-making, but to help improve decisions by nudging people to better outcomes while allowing for individual freedom of choice. Thaler and Sunstein's (2008) nudge theory developed the idea of libertarian paternalism; that is, allowing someone to make better decisions, benefitting themselves (or society), but in a non-controlling way.

The financial context is particularly conducive to nudging, because decisions in this space are often complex, feedback is often deferred, and decisions are often made in an emotional state. Nudge theory has thus emerged in diverse areas, including retirement savings, investment diversification, debt repayment, and consumer product choice. As Cai (2019) notes, financial decisions are often subject to the influence of psychological biases that detract from long-term flourishing. The nudges in the financial domain aim to reduce friction, encourage reflection, and re-position priorities - largely through subtle information design, and changes to how options are presented.

These dilemmas have become more acute with the ongoing development of algorithmic and data-driven nudging. Unlike earlier nudges that were typically predefined, implemented uniformly, and visible to the initial chooser, machine learning nudging has transformed previous notions about behavioural influencing. Behavioural influence is no longer static or generalizable; it is dynamic, hyper-personalized, and increasingly opaque.

Hypernudging, as defined by Yeung (2019), refers to a methodology that utilizes artificial intelligence and big data to generate dynamic choice architectures that continuously adapt in real time to individual behaviour.

These new nudges can learn behaviours and updating themselves all based on a user’s prior decision-making, affective state, and contextual information. The hypernudge is not simply an intermittent nudge, as it becomes an ongoing behavioural environment that has the possibility of shaping not only individual decision-making, but, possibly, the construction of preferences themselves. It is important to recognise that traditional nudging frameworks did not consider real-time or adaptive mechanisms when initially developed. While established nudging attempted to correct for recurring tendencies - like impulsivity or acceptability of the status quo - today’s nudging platforms may leverage those same patterns to predict, respond,

and reinforce future behaviours. Such inherent adaptivity weakens traditional aspects and assumptions of nudge theory including deliberate boundedness, ethical sanctity, and user-centred agendas.

The transition of nudging from policy to algorithmic influence is further complicated by the opacity of the AI systems driving our behaviour. Rather than choice being embodied in a menu, order form, or other competitive variable, choice in a digital context is embedded within algorithms, personalisation engines, and data-for-free loops, where the user may not only not perceive the nudge, but cannot even conceive the original or logical emergence of the nudge. Per Nyman (2023), what is increasingly being recognised is that these platforms are functioning as forms of behavioural governance, not merely nudging users toward intended goals, but carefully curating user experience through a balance of behavioural psychology and machine-based prediction.

This status change represents a broader ontological shift in influence, where the nudge is no longer a singular act of desire, but a behavioural infrastructure bolstered by real-time data collection, ongoing algorithmic profiling, and platform-based optimisation. Within this evolving ecology, the durability of the nudge and the use of seemingly human-like design signals creates a landscape where the boundaries between benevolence and exploitation and support and subtle control may become increasingly indiscernible.

The following section will explore how AI is opening nudging to a transition from a mechanistic choice framework to a fluid system for behavioural design. To understand this change is to make sense of how power, persuasion, and prediction increasingly blend inside the architecture of contemporary finance.

1.2 AI-Driven Nudging: The Evolution of Personalised Interventions

AI nudging refers to a range of touchpoints leveraging machine learning algorithms, behavioural insights and predictive analytics to create experience-led nudges, dynamic prompts, recommendations, or changes in an interface, which have intended consequences on user behaviour. The main difference being that where common nudges were delivered once at a time in a wide area (generalised) way (e.g., modelled with some questionable 'willingness to pay' or measures of influence), AI nudges create real-time nudges for the individual users, meaning that they will involve systems that adjust at every touchpoint. Every user touchpoint would be retrieved as a data point that would adjust models to behaviourally influence strategies in the furthest limit. As per Wagner (2021), we are moving from 'choice architecture', to something more able to be active in the model of 'behavioural orchestration', where we are not just are not only assisting decisions, but also shaping, directing, and even anticipating first-order behaviours.

This change is especially visible in financial decisions. Digital investment platforms like Robinhood, Betterment, and Wealthfront have embedded behavioural science in their design within the app or website, essentially eliminating the human advisor and replacing human interaction. Imagine a website that enables the desired behaviours through design. In these cases, a user can be 'nudged' to fund their account with leftover cash after a purchase, to 'stay in' when the market is down, or to 'rebalance' following very small changes in behaviour or in

the market. Nudges tend not to be discrete, but rather finely calibrated interventions triggered by algorithms developed to track user patterns, emotions, and financial behaviours. Robinhood provides a stark illustration. With its design for screen-based engagement and gamified features, such as animated celebrations and confetti, Robinhood nudges users into more frequent engagement, often encouraging the mindset for short-term trading. These nudges are not random free will restrictors; they are algorithmically selected nudges that are encouraging retention and increasing transaction volume. Betterment and Wealthfront were more mellow, nudging a user into long-term financial well-being through suggesting a goal, automated rebalancing, or prompting reported notifications with sense and nuance. However, even within this more passive approach, there is a behavioural element which strategically tries to reduce reactive choice in comparison to steady engagement.

The key aspect of AI-supported nudging is the level of micro-personalisation and temporal optimisation. It does not rest on standard demographic and common behaviour archetypes; AI nudging action practices create behavioural profiles - users are tagged for when they hesitate, the length of time it takes for them to decide, and how much they change emotions. AI nudging engagement then uses this knowledge to create the micro-personalised nudge, such that not only does it decide what to nudge, but when and how to nudge. Nudging then becomes cognitively synchronised, with every path being wrapped in the contextualised psychology of each user. As Nyman (2023) points out, the shift from prompts to nudges reflects a broader transition from decision support tools to mechanisms of behavioural governance. He describes AI-driven nudging as a subtle and pervasive form of influence embedded in the everyday infrastructures of digital technology and finance - quietly shaping behaviour through design.

Furthermore, traditional nudges are overtly conducted within the visible and recognised environments of choice, whereas the conduct of IT-based nudges is often algorithmically hidden. This observation engenders a deep asymmetry of knowledge and control: platforms will have considerable insight into user conduct; users generally have no idea of the systems which are informing their opportunities and shaping their decisions. This asymmetry represents a significant shift in ethical behaviour. Whilst traditional nudges, even when subtly contextualised, their logic is usually made available to the user. In most digital environments, nudges activation is ever more bearable under policy formation, proprietary models, feedback loops, and adaptive systems. A user may receive a push notification from their trading app telling them, “Stay the course!" when there is volatility. However, the logic of their original investment is buried in real-time predictions or user behaviour and tagline commercial priority. The result is not just personalised nudging but a remaking of agency, where choices are silently pre-determined by obscured systems. This has generated a buzz around hypernudging. More than one-off nudges, hypernudges move, shift, and recalibrate with every click, hesitation, and scroll. They evolve in lockstep with the user, creating an algorithmic feedback system of behavioural change. Hypernudging will not merely encourage better choices, it will manufacture better choices that serve a platform equally as much as user wellness.

On the other hand, well-designed, ethical, and transparent systems can correct for cognitive distortions, encourage financial literacy, and offer behavioural scaffolding for users who would not or could not access traditional advisory services. Nudges that promote dedicated

saving, automating portfolio rebalancing, or counteracting emotionally reactive behaviours could save users from themselves; meaning that the problem isn't influence, it is the quality and intentions of the design, acknowledgement of autonomous behaviours, safeguards against manipulation, and altruistic alignment with user welfare.

In short, the shift from static nudges to hyper-personalised nudges is an important moment in the practice of behavioural influence. These are fundamentally new forms of power, precision, and persuasion, and offer both potential and peril.

1.3 Ethical and Regulatory concerns

As AI becomes more embedded into behaviour design, the ethical and regulatory implications of AI-based nudging require greater commitment and attention than is currently being given. Although nudge theory was initially conceived as a gentle, transparent, and autonomy-preserving tool, it has evolved into the increasingly complex arena of algorithmic data-driven systems that raise a new set of challenges beyond the original ethical framework from which it emerged. According to Sunstein in *Why Nudge? (2014) and The Ethics of Nudging (2015),* the legitimacy of nudging stands upon criteria such as transparency, congruence with an intuitive notion of individual welfare, and the freedom of choice. These principles face challenges in AI-driven environments, where personalisation and opacity often obscure intent. Frameworks like FORGOOD (Fairness, Openness, Respect, Goals, Outcomes, and Dignity) provide a useful ethical lens for contemplating digital nudging interventions, especially where user welfare will likely play off against platform incentives within commercial ecosystems. While the benefits of AI-assisted nudging will undoubtedly make it more effective, they also make it more challenging to analyse, regulate, and hold one accountable. We have shifted from a notion of passive architecture of choice to one that is actively algorithmic. This glaring change raises significant agency, transparency, and responsibility issues, especially in digital financial contexts.

Ethical notions come from a fragile, functioning equilibrium between paternalism and freedom. Thaler and Sunstein (2008) articulated the idea of libertarian paternalism by arguing that it is possible and desirable to influence people's choices in ways that will benefit their welfare while also allowing them the autonomy to choose as they see fit. This idea rests upon several ethical conditions: that nudges are transparent, easily avoided, and consistent with the user's best interest as they define it. Even with these boundaries in mind, some critics have expressed concern that nudging can be manipulative, occur without consent, and can be asymmetrical in influence.

In AI-driven systems, the traditional notions of transparency and accountability are undetermined. AI-generated nudges are far from explicit default settings or opt-in checkboxes. Instead, they are often covert and operate in a feedback system that is increasingly complicated, yet invisibly dynamic on the backend, running predictive models powered by proprietary algorithms. Users typically do not notice or understand that some form of nudging has occurred, let alone how or why a suggestion or certain design surfaced, thus posing a radical asymmetry of information.

Taken together, this opacity presents a problem concerning informed consent. Consent is

steeped in the principle of informed consent, and in almost all examples of AI-powered systems, users provide their consent in the form of general terms of service or data sharing/collection agreements. Binns (2018) cautions that algorithmic systems will generally never provide any "explainability" to help reason for selecting interventions and for designers to articulate nudges in a specific design context. When users cannot be reasonably informed or when there are tacit assumptions behind the consent provisions, users' consent becomes a performative gesture, an artefact of legal reasoning rather than meaningful ethical considerations of design.

In addition, nudging was intended substantially for public goods (fewer health-related risk behaviours, higher retirement savings, less accumulated debt). Although some uses of AI-powered nudging can be construed to serve some proxy of user welfare, many are functioning in commercial ecosystems. This means the nudging strategies are now tuned for platforms' interests as opposed to serving the users' interests and may result in an ambiguous goals misalignment. For instance, a platform may merely nudge users to remain invested, while there exists strong turbulence in the markets not only for users not to panic sell, but rather, to keep the assets under management and the platform from withdrawal fees. Likewise, users may be nudged to use the platform more often, not to enhance financial literacy but to enhance app engagement metrics or transactional revenue. This kind of interest-driven behaviour is particularly troubling in financial markets where nudging can impact high-stakes decisions under uncertainty. The ethical stakes are higher in cases where users are not just relying on a platform for convenience, but view that platform as providing advice. The fiduciary distance between digital interfaces and users muddles the separation between service and suggestion. A human financial advisor must adhere to explicit ethical obligations, whereas the same cannot be assumed for algorithmic nudging systems seeking to optimise user welfare while concurrently seeking commercial gain without compromise. Given these ethical tensions, we can point to a need for existing regulatory space to emerge with effective regulatory structures. However, current regulations tend to be poorly equipped to deal with the behavioural and psychological elements of AI systems. Regulatory frameworks regarding finance tend to be focused on disclosure, risk management, and market integrity - areas where behavioural nudging has not been examined with the same rigour. Additionally, the majority of regulatory frameworks are highly reactive and rules-based. In contrast, AI systems are adaptive and probabilistic to continually alter modelled performance. This disjunction suggests that conventional regulatory methods are not only inadequate but are at times irrelevant.

New scholarship has recently identified some ways forward. One approach is the recommendation of placing algorithmic transparency requirements on platforms, hence requiring them to notify users when nudges are in use, explain its behavioural aim, and how the nudges were created. But transparency is only a small start. The leading edge will instead include behavioural impact assessments, a more co-operatively structured evaluation of signs indicating the system's nudges aligned with user welfare, the user's chosen autonomy, and were free of exploitation of their decision-making disposition toward digitally interacting. It may look like the existing service-oriented fairness or safety assessments used in AI ethics, but it will also explicitly surface a behavioural component.

Another approach is to instigate ethical design standards to govern behavioural interventions

like ethical standards in medicine or product safety assessments. Platforms would be obliged to explain how the nudges were evidence-based, minimally manipulative and, used as a user-centred design. It may also allow for friction-preserving design, where in the context of nudges users are required to pause, reflect, or at least consider their engagement with a nudge. Or as in the case of nudges, opt-out, thus allowing the user to disable nudging choices.

Regulatory safeguards should include the cumulative effects of nudges over time. Individual nudges may act benignly, but the combined effects of a user acting on hundreds of micro-interventions is far-reaching and structurally alters users' financial experience and arguably their identity. The cumulative effects are hard to identify if assessed in isolation, and scope may require longitudinal inquiry and mechanisms for ongoing regulatory monitoring. The cumulative effects also raise normative questions about what type of financial subjectivity is being developed.

In conclusion, despite the dangers, AI nudges offer the potential to genuinely support users, constructively promote their financial security, overcome cognitive biases, and assist in interacting with systems that previously left them outside of the limits of access. Thus, the goal should neither be to eliminate influence, nor contrive a way to eliminate it entirely, but instead to design ways to govern influence, making sure that in powerful digital infrastructures, behavioural interventions we use are compatible with human dignity, autonomy, and the user's long-term welfare.

CHAPTER 2: AI NUDGING ON THE INDIVIDUAL LEVEL

2.1 AI-Powered Nudging and Investor Behaviour

The emergence of AI in financial services is changing not only how people manage their money but also how their behaviours change. AI shape the design of digital banks, robo-advisors, and trading platforms. They design a 'user' interaction, predict, nudge, each interaction and adapt after every interaction. Most notable in this context is the new practice of using AI nudges, or engaging in data-informed interventions.

Importantly, these developments are more heightened in financial services and situations, because decision-making in financial contexts is challenging; futures are unclear, often clouded with some level of uncertainty, risk and consequently complexity, which promotes cognitive biases. Behavioural tendencies related to present bias, overconfidence, loss aversion, and status quo biases are powerful determinants of financial outcomes. AI-assisted nudging can learn from these tendencies and either exacerbate or mitigate them based on whatever compensates the platform. The line between personalisation and manipulation, and empowerment and dependency, becomes increasingly blurred as platforms supply these systems under the name of personalisation. Recent work addresses both the opportunity and risk of AI-enabled systems. Sadeghian and Otarkhani (2023) describe how banking applications using AI to support behavioural nudging effects change in how users spend, save, or borrow, with positive user engagement outcomes, but also potentially increased user dependency or decreased experiential learning over time. Oliveira and Leal (2020) argue that robo-advising platforms use behavioural knowledge to optimally build and manage user portfolios as a client-led strategy, while also being submerged in algorithmic processing. Behavioural nudging operates across a number of platforms and has a constitutive role in what choice means in the context of people working with one another and machines, and how that definition becomes more systematic or destructive.

This chapter will be divided into three related sections. The first section will address how AI nudging works as a placement for retail finance and personal investments, where the user is deciding how to save more money, investing for retirement or trying not to sway to emotional decision making through continuous nudges. The second section moves into retail trading and portfolio management with the stakes getting higher, but so do the effective contribution of nudges, potentially destabilising the user experience through the use of nudges that encourage reactive and repetitive behaviours. For example, in Robinhood, a popular app in the U.S. among new investors and traders, nudges often facilitate overconfidence and immediacy, allow for constant trading under the pretence that users possess skills of foresight. The third section looks at the behavioural transitions that AI systems are both profiting from and reinforcing, and the ethics of influencing users in systems that learn what they are most susceptible to or likely to, repeat in an altered form. The chapter hopes to think through the process of mapping this emerging terrain, as well as contribute to the developing behavioural apparatus of finance and (de)stabilising power, psychological distance and the familiarised design of agencies.

2.2 AI in Retail Finance and Personal Investment

The growth of intelligent financial platforms has significantly changed the landscape of personal finance. The traditional experience that was mainly static and dominated by the advisor has transitioned to a fluid, algorithmically driven process, in which digital systems manage budgets and automate a range of transactions while also anticipating needs, analysing behaviour, and directing choice in subtle ways. At the centre of this change is AI-driven nudging: using data, behavioural understanding, and predictive modelling to better and effectively influence financial behaviour in ways that are unnoticeable to the user; sometimes even unconscious.

Retail finance apps are leading the charge in this evolution. Apps like Cleo, Qapital, and Mint use AI to track user behaviour, set savings goals, flag unusual spending, and advise on steps to achieve financial goals. Importantly, they do not present information; they interpret and contextualise it, delivering prompts based on the user's history, preferences, and behavioural patterns. As Sadeghian and Otarkhani (2023) note, these platforms "understand the user's past to provide the appropriate nudges at the appropriate time and with the appropriate content that leads to the best possible financial showing." This is a kind of scaffolding of the digital user where staff behaviours are embedded in their daily behaviours in ways that render the decision-making process as natural and automatic responses to choreographed prompts.

This design is easy to see in automated savings solutions. Notifications nudged someone to "just save *$5* more this week" and some apps may use gamified visuals to reward good savings behaviour with badges or a progress bar. Nudges are based on known behaviour theories.

For example, what is known as present bias, as described by Laibson (1997), is a psychological bias where individuals value rewards that come sooner rather than larger (even exponentially) rewards that come later. AI platforms circumvent the present bias by changing future financial goals into current action based on future value but incorporating aspects that are felt as immediate value. This is also seen in loss aversion, as discussed by Kahneman and Tversky (1979) in prospect theory, with the opportunity of inaction or the loss of a potential savings situation being framed as a loss event, as a way of getting people to do something different from inaction to start proactively spending, and they subsequently positively impact their finances Early We will revisit this hindsight in the later conclusion. It is important to also highlight the effectiveness of these nudges. Increasing the effectiveness of nudges can be introduced through real-time personalisation. For instance, two users, with the same income, may receive completely different nudges - or prompts - based on a user profile, but also based on the user's previous actions undertaken, emotional moments triggering, and even what time of the day they are logged in to the app. The nudge becomes cognitively synchronised - in this case, personalised - both to the users and to their typical responses and actions. This is a major move away from previously described nudging and ideally describes an evolution of design that stays true to Thaler and Sunstein's (2008) original idea of libertarian paternalism and increases accessibility by using algorithms and machine learning.

These same principles are also being adopted in personal investment platforms, and especially in the space of robo-advisors. By employing services such as Betterment, Wealthfront, and N26 Invest, users receive algorithmically managed portfolios that are

designed to meet users' goals, risk profiles, and time horizons. However, the innovation is not limited to automated asset allocation but also how they weave nudges into the experience. For example, Oliveira and Leal (2020) state that current robo-advisors "nudge continued contributions, long-term commitment, and risk in terms of framing, such as framing carrying a sense of progress, temporal proximity to goals, and small nudges toward positive motivations." A graph that demonstrates retirement savings as "years covered" rather than raw currency, or a simple reminder messaging the user to stay invested during a market dip—this is a behavioural intervention disguised as interface design.

The platform is acting like a behavioural coach rather than solely an algorithmic portfolio manager at those perturbing times. The intent is not only to find the maximised returns on investment but to nudge the user away from emotionally distressful decisions (like panic selling or unreconstructed reallocation) that may also be motivated by loss aversion or availability bias—both of which can foster backward thinking that causes the user to exit a long-term plan. The platform's interventions help provide context and provide guidance that can alleviate both cognitive loads and emotional distress that accompany decision-making at these difficult times. Where we can identify this in digital interactions, we can pinpoint paternalism, not in the sense of controlling the user, but as algorithmic helpfulness.

The design considerations that make these interventions look like nudges are rooted in behavioural economics. Established principles of behavioural economics like the goal-gradient hypothesis (Hull, 1938) and anchoring effects (Tversky & Kahneman, 1979), are not just built in; they are adjusted diachronically. For example, if a user consumes content aimed at retirement promotion often enough, the platform may set the user up with more ambitious savings targets, compared to a user who experiences unexpected suboptimal cash-flows, for whom the same platform provides 'softer nudges', focused on liquidity. The ability of AI to identify and respond to these patterns accurately increases the efficacy of its nudges but also renders them less visible, and therefore, reinforces both their power and their ethical uncertainty.

This invisibility raises a particular challenge for the design of interfaces themselves. As Yeung (2017) and Wagner (2021) note, digital contexts can embed behavioural influence in visual hierarchies, layouts, and navigational features. Investment platforms often present portfolio choices in ways that nudge individuals into making certain choices—including those consistent with not only the user's risk tolerance, but the platforms revenue model. The prominence of buttons, colours, or language may encourage particular behaviours under the guise of neutrality. This type of framing encourages default bias (Samuelson & Zeckhauser, 1988) by nudging users towards already selected or default options, not necessarily because they are the best option, but because they require less cognitive effort.

It is important to note the influence of AI-powered nudges is not experienced equally by all users. Younger users, as digital natives accustomed to mobile interfaces and responsive design, may be particularly influenced by so-called real-time feedback and gamification nudges. On the other hand, older, or less financially literate, users may treat platform nudges (e.g., prompts) more heuristically as prescriptive advice rather than as choices. As noted by Shefrin (2000), the susceptibility to behavioural triggers is not only a function of individual cognition, but also social and demographic context. These inequalities have equity implications—behavioural nudges may not only guide users but stratify them based on their

financial competence and confidence, thus amplifying existing divides.

Yet, even with these concerns, AI nudging has strong normative claims - particularly since these systems are transparent, intuitive and user-centred. In a world where so many people do not have access to personalised financial advice, these platforms are an inexpensive, accessible, and scalable alternative. Nudging behaviours, aimed at overcoming indecision, avoiding impulsive decisions, and staying aligned with longer-term objectives, can be an important behavioural counterbalance to financial illiteracy and short-termism.

However, the promise of AI-powered nudging is nuanced. AI systems are adapting and increasingly anticipatory, and behavioural support can easily morph into behavioural dependence. If it seems the platform always "knows best," users may yield their decision-making instincts—not just to computer algorithms but to indiscernible architectures of influence. The line between support or guidance-breaking and automation-blurring, governs or is governed-becomes indiscernible.

This section has set forth ways in which AI-powered nudging is re-shaping retail finance and personal/investment, embedding behavioural influence and options to a much greater extent into the everyday aspects of digital finance. Whether it be objectives for automated savings, psychologically informed robo-advisors, today's investors are traversing systems that are not just reactive, but pre-emptive. These systems influence not just what choices are made, but conditions of possibilities for the choices. With that in mind, we will now turn to trading and portfolio management platforms, where helpful nudges become immediate, and stakes are rendered more volatile, while ethical lines become much more blurred.

2.3 AI Nudging in Trading and Portfolio Management

If retail finance platforms help establish users and habits around structure and long-term goals, trading platforms do the opposite by offering pathways to immediacy, uncertainty, and emotional charge. In trading applications, the stakes are higher, the feedback loops are tighter, and the action seems to hinge on emergent and combustible behaviour. That is where AI-enabled nudging in trading platforms takes on a different character that is more than supportive or advisory: it is often evocative, enticing, and occasionally disruptive. The more the user engages in cycles of buying, selling, and returning to action, the more immediate and interventionist the influence of AI-based nudging will generally be.

Trading platforms such as Robinhood, eToro, or Trading212 redefine what it means to be a retail investor, in terms of design and fidelity. The interfaces are sleek, gamified and finely tuned for engagement purposes. Users are typically nudged to move as fast as possible, even while they do not fully weigh or appreciate the implications of their decisions. They are encouraged to act quickly, aided by celebratory animations following every event that entails a successful trade or investment. The nudges are typically exogenous to the investing process: they comprise prompts and notifications, timed to moments of market volatility and movements or by visually appealing prompts to act. They also exist for nudging and now have little in common with a traditional decision aid, for example, emotional resonance, and salient tapping into cognitive biases and behavioural heuristics to trigger another cycle of action intended to deepen and perpetuate behaviour.

Of significant importance here is overconfidence. Odean (1998), and later Shefrin (2000), point to overconfidence in retail investors, which can be characterised at base level with the observation that retail investors are regularly too confident in their understanding and predictive ability, an illusion enhanced in users of platforms that reinforce each transaction with positive sentiment and satisfaction confirmed by transaction. Robinhood's confetti animation, since removed due to regulatory scrutiny, demonstrates this. Trades were presented not only as something you might do, but as something to be accomplished. The consequent psychological impact was not just uplifting; it was misleading and created a sense of competence, relying solely on momentum, which was often not based on tangible performance. By celebrating the action itself instead of the outcome, platforms engender a behavioural loop that makes trading habitual and compulsive.

A related trend is present bias, which is the tendency to prioritise immediate gratification over delayed outcomes. Laibson (1997) and Thaler and Benartzi (2004) both pointed out that present bias is especially pronounced when it comes to finances. Specifically, the benefits of restraint will be delayed, while the benefits of action will be immediate. Trading apps capitalise on consumers' present bias by creating high-frequency, high-feedback, and low-friction experiences. Emerging real-time prices, scrolling green and red, and executing trades are designed to minimise time, a function of temporal compression, where the future is eroded, and the present reigns. Nudges push users to react, rather than reflect.

Another bias is particularly powerful in times of market stress: loss aversion. In their prospect theory, Kahneman and Tversky (1979) described how people experience more pain in a loss than pleasure in an equivalent gain. In these moments, trading platforms use nudges to frame a turbulent market as an opportunity or to suggest to the trader to "stay the course" during volatility. While such nudges can reduce panic selling, they also keep the trader engaged and in-app. The ethical line may be quite fine. A prompt set up to slow down emotion-driven decisions may also dampen platform liquidity or fee revenue, even if disengagement may have been better for the user.

These pressures are often exacerbated by the social nudging functions of the platforms, especially in platforms like eToro that facilitate copy-trading and showcase "trending investor" data. These features incentivise herding behaviour and make users copy actions of individuals they believe are experts or are perceived as popular investors. While social learning is not necessarily unreasonable, its algorithmic delivery can amplify distorted signals. A trader whose recent success was mostly random can attract thousands of followers, not because they are good, but due to platform amplification. This produces a false appearance of concordance that further reinforces, as Banerjee (1992) has mentioned, informational cascades, shifts in behaviour that are dependent on what a perceived group is simultaneously doing rather than individual analysis.

AI is central to shaping these nudges. Along with platform affordances that have much less stagnant ranges in former digital financial platforms, trading apps are designed with algorithms that leverage machine learning to predict user reactivity, shape timing and tailor messaging according to the individual under scrutiny. An indecisive user may encounter a more forceful prompt; a user who might react more to social gregariousness may engage with even more community-based recommendations. These models don't just offer personalised prompts - they offer personalised behavioural environments that have been designed to

maximise engagement through a focus on predictive behavioural science.

This transition has important implications. When platforms discover not just what users do, but what they are most susceptible to, nudging becomes behaviourally targeted. In the long run, AI systems could reinforce existing biases rather than correct them, by capitalising on patterns that facilitate short-term engagement at the expense of long-term behaviour. For example, a user may be over-confident in their trading behaviour, and the AI nudges that confirm this behaviour will facilitate higher engagement while providing very little in support of an investment strategy. As Nyman (2023) describes, the platforms become not tools for empowerment, but what he calls “architectures of behavioural capture.

Unlike traditional advisors, who are bound by fiduciary duties to their clients, most trading platforms generate their revenues by order flow, spread differentials, or subscription fees, all business models that reward frequent activity. Hence, AI nudging creates a complexity of profit motive, conducting questions about whether an individual is receiving nudging that is in their interest, or whether the nudging is in the interest of the trading platform. A nudge to "discover trending assets" may feel like a valuable option, but in reality is a design feature prompting a trade. The semblance of aid hides a logic of storytelling and potential extraction, a change of behaviour from nudged guidance to nudging via monetisation.

The implications for AI-nudged users in trading platforms are therefore layered. On the one hand, the implications speak to the increasing sophistication of behavioural design and an ability to shape decision making in granular ways. On the other hand, the implications of AI nudging are an unsettling fragility of user autonomy, when placed within algorithmically optimised architectures where the core interest is to incite action, not promote reflection.

2.4 Behavioural Biases and their Algorithmic Exploitation

The incorporation of artificial intelligence into financial decision-making devices has prompted an unprecedented level of precision in understanding and influencing human behaviour. Yet, with this power comes a more complete entanglement with the same biases that behavioural economics attempted to expose and mitigate. AI systems are not neutral parts. They observe users' behaviours, are programmed to predict future behaviours, and increasingly, they alter the user experience to influence outcomes. They can functionally implement cognitive bias in both corrective and exploitative directions - a consideration worth serious inquiry. This section will consider how some degree of AI-informed nudging interacts with four of the foundational biases of behavioural economics: loss aversion, present bias, overconfidence, and status quo bias.

Loss Aversion: Increasing or Alleviating the Pain of Loss

First described by Kahneman and Tversky (1979) in their research on prospect theory, loss aversion refers to the human tendency to feel losses disproportionately to equivalent gains. In financial contexts, this means the bias comes through as an aversion to selling at a loss, an aversion to short-term declines, or an aversion to portfolio adjustments in the face of volatility, even when said portfolio readjustments may be a rational choice. AI systems may have a unique opportunity to act on this bias. On one hand, platforms such as Betterment and

Wealthfront suggest nudges that actively frame downturns in the market in ways that reassure and remind users to stay the course, to nudge them away from an emotionally reactive decision. These interventions could simply create an environment of protection, encouraging them to think longer term when the market is in flux. Nudges have the potential to mitigate loss by either choosing the frame to be temporary loss, or by putting it into the context of broader portfolio returns to lessen the psychological pain of losing money.

In contrast, there will be platforms, mainly exploiting users for engagement or trading activity, that will use the same bias to actively inhibit disengagement. There are likely systems that do not show individuals a loss, have some like performance, which may mitigate the perceived loss, or draw out some trending asset that just bounced back, effectively framing the decision to keep their capital with them as rational and potentially enticing. The emotional charge of loss aversion becomes an instrument of behaviour: a method of creating loyalty to a platform as much as it is to sound investment behaviour.

Present Bias: Current trade, future sacrifice

Present bias is central to Loewenstein's (1989) model of intertemporal choice. Present bias is the human tendency to place too much subjective weight on present reward relative to discounted future outcomes. It has been shown to be a major part of one's ability to save for the future or delay consumption, investment planning, and debt management, where immediate satisfaction or comfort supersedes long-term benefit.

In robo-advisory situations, users' present bias is generally unhelpfully resisted. Nudges that enable automatic contributions, feed-forward evidence of progress, and optimise present bias engage users in taking the abstract (retirement) and making it into tangible, shorter-term steps to narrow the distance. As Thaler and Benartzi (2004) showed in their Save More Tomorrow programme, commitment devices or default increases can leverage present bias toward a positive outcome by decreasing the pain of the sacrifice in the immediate context.

Conversely, a trading scenario is potentially more susceptible to employing present bias. Platforms like Robinhood or eToro provide an ongoing stream of real-time updates, instantaneous transactions and gamified feedback, all of which constantly reward engagement. Nudges give the users action and rewards a sense of immediacy - stock prices quickly advance across the screen, direct performance measurements, celebratory graphics - and all these factors collectively invite the user into a short-term decision-making cycle. The effect is an inhibition of reflection and the creation a set of reinforcement pathways for actions that are action-stimulation as opposed to strategy and gratification as opposed to growth.

By creating trading as a messy, but fun, gamified decision that focuses more on the experience of investing as opposed to the long-term consequences of the decision. The behavioural ramifications are significant: investors begin to look at each trade, not necessarily each movement, in a deliberate plan, but more like micro-events, part of a series of emotionally reinforcing action-cycles that are ultimately synchronised to their present bias and the compulsive need for real-time feedback loops.

Overconfidence: The Illusion of Competence in Digital Markets.

A well-understood bias in financial behaviour is overconfidence bias - the tendency to overestimate one's knowledge/skill, or ability to control outcomes. For example, Barber and Odean (2000), have found that overconfident traders tend to trade more, and in lies the potential to affect returns. Overconfidence is a significant contributor to the dynamics at play in digital trading platforms, which interface can itself create the illusion of control.

Generally, AI nudging obfuscates overconfidence. In the more conservatively behaving systems, like a traditional robo-advisor, the interface induces a lesser illusion of control because it restricts levels of customisation, and the focus is on risk-adjusted practices. Nudges present in these environments could include reminders regarding market uncertainty, enactments of loss scenarios, and nudges to remember risk profiles- interventions that induce an environment of reduced confidence and reflection.

In contrast, trading platforms generally create an industry of overconfidence using feedback design combined with space for personalisation. If a user performs well, they often receive notifications to congratulate them based on their performance and remove potential risk from re-entry based on luck or market. Furthermore, this only gets augmented when AI systems recognise these experiences and then begin to produce nudges that prompt more positive engagement, more leads for "trending stocks," and predictive trades, be they are successful at a given moment; these have the potential to increase the sense of having control.

Overconfidence in behaviour enclosed in algorithmic environments can reinforce biases through confirmatory bias, where the user is much more likely to seek and involve themselves in information that supports their preexisting thoughts, moving them in the same direction. AI systems trained to capitalise on engagement seem to learn that users with a predisposition for bullish movements will prioritise "bullish" content, leading users to greater risk-seeking profiles. In essence, they have programmed themselves as empathetic to bias and fall into behaviour extending a user's behavioural groove and acceptance rather than challenging it.

Status Quo Bias: The Safety of Defaults in Algorithmic Environments

Status quo bias, presented by Samuelson and Zeckhauser (1988), is the preference for current conditions over changing them even when a change is better. It is something we see across finance in the format of behavioural constructs involving reluctance to alter investment behaviour, re-evaluate risk profiles, or change retirement plans as personal circumstances evolve.

AI nudging algorithms will both exploit and correct this bias. On the positive side of the argument, in more cases than not, in robo-advisor environments, status quo bias is a part of the process. By creating automatic outcomes - like rebalancing, periodic contributions and pre assumed allocations - these platforms lead users to reasonable behaviours without any friction. Simply put, people will stick with the path of least resistance even if that path is not even the best for them!

Status Quo bias cannot be disconnected from its commercial implications. Some platforms have been built with defaults that are truly comprehension of their incentives - selection of a

product they have greater margins on, a selection at higher risk to increase returns (and fees), a default pre-identified premium service on default onboarding. Once the defaults are made, they stick. After all, the reason that a default is so powerful is because, similar to best behaviours exemplified by the status quo bias, once a user is stuck, they can often get stuck in bad defaults too.

This capacity for dual use implies an observation at this point where the parameters laid out by AI nudging possess a core underlying theme - that the same fundamental in behavioural terms can be used for good and to entrench asymmetries of information, power and commercial benefit. The construction of ethics is not in the presence of biases, but in the manner and intent of its operationalisation.

From Adaptive Personalisation to Behavioural Targeting

What emerges from the prior analysis is a unit of environment of behaviour that is highly personalised but is not necessarily empowering. AI systems learn the default biases that an individual is prone to, the nudges that work well, and lastly, the sequencing of nudges to enhance engagement over time. Over time, it is for these biases what could be described as a behavioural targeting model where content can be not only personalised but be uniquely tailored to influence!

This feedback loop - where the system modifies its intervention depending upon previous behavioural success - certainly begs a harder question: where is the line between helping users deal with biases and designing for their predictability? Nudging was originally intended as a soft cue, a nudge, not a behavioural mandate. But in AI environments, nudges are so finely tuned they can become invisible scripts that direct decisions in ways that the user is not easily aware of, let alone able to reject.

Even risk, even if behavioural biases are human constants, does not behave in the same way when engaged by AI. What matters is whether these biases are being utilised to minimise harm and support autonomy, or whether they are being monetised in ways that reject critical reflection and reinforce dependence. The call to action is not to remove nudge or to get rid of AI; it is to develop design ethics and regulatory frame that can differentiate between behavioural support in ways that offer agency vs. behavioural exploitation.

CHAPTER 3 – AI NUDGING AS MARKET PHENOMENON: EMPIRICAL ANALYSIS

3.1 AI Nudging, Systemic Risk, and Market Stability

As AI-powered nudging embeds itself into the architecture of modern finance, there is a shift away from the effects of individual behaviour and towards collective behaviour. What begins as a series of personalised, context-specific, micro-interventions - each designed to subtly push a user's choice - when scaled across millions of actors and users, can establish collective behaviour at the macro-scale that will in turn influence the markets themselves. The Financial Stability Board (2024) warns that the widespread use of similar AI models and shared data inputs across institutions can heighten correlations in trading behaviour, particularly under stress. This phenomenon, known as algorithmic convergence, can lead to synchronised reactions to market signals, amplifying volatility and threatening liquidity.

Having established a behavioural baseline in Chapter 1 and explored individual-level examination of investor engagement in Chapter 2, this discussion moves up a level of scale to interrogate the market-wide effects of intelligent behavioural influence. Chapter 3 expands on this idea by exploring how convergence moves upstream from institutional trading desks to retail platforms powered by AI. Although these platforms are often framed as democratising finance, they employ a framework of personalisation with digital nudges and constructed defaults that further align investor behaviour in possibly destabilising ways. As an increasing number of AI systems inform decision-making in institutional and retail environments, the line between reactive and reflexive market behaviour is gradually becoming obscured.

This is especially alarming in a context in which AI systems are not just predicting individual behaviours but optimising for engagement, retention, and predictability – a transformation Zuboff (2023) characterises as a shift from surveillance to behavioural modification.

As discussed in Chapter 2, platforms continue optimising nudging strategies using user data, thus delivering more tailored and time-sensitive nudges. However, when millions of users are nudged in similar manners to "stay the course" in a downturn, nudged to invest in an ongoing surge of an asset, or nudged to follow trending behaviours of other investors, the risk is that varied human behaviours become algorithmically homogenised. Behavioural economics documents that our financial systems not only indulge cognitive biases such as herd behaviour, availability heuristics, and overreaction. AI does not remove our cognitive bias; rather, it learns our bias, reproduces and potentially amplifies biases based on feedback loop inputs. As with individuals, the implications for institutions are profound. Investment banks, hedge funds, and high-frequency trading firms are increasingly using AI for strategy execution, risk management, and portfolio allocation. Even these systems are fundamentally reacting to behaviour-based inputs and are therefore subject to the same forms of algorithmic synchronisation. When both retail and institutional behaviours can be stimulated by behavioural signals, whether from market sentiment or volatility signals and performance dashboards, an AI ecosystem is created that induces a recursive dynamic that allows AI models to start 'predicting' and responding to other AI models. Sometimes referred to in the literature as reflexive modelling, the emergent spatiality of AI-nudged behaviours can lead to simultaneity of decision-making across the whole system, in ways that can lead to market instability.

This chapter is structured in three sections to study institution-level behaviours using AI nudging. Section 3.1 examines the impact of AI nudging at an institutional level, specifically the ways in which algorithmic trading systems will embed behavioural assumptions into all facets of a high-speed decision-making process. Section 3.2 assesses the emergent risks of behavioural convergence at a market level, especially when real-time AI nudging replaces slower, deliberative, or arguably rational financial behaviours. Finally, Section 3.3 applies a case study method using three distinct events to elaborate on how AI-induced nudging behaviours' cumulative presence and influence can add to instances of misallocation in markets.

3.2 Institutional AI Nudging and Algorithmic Trading

The behavioural impact of AI-driven nudging is not limited to the realm of individual investors. It penetrates institutional finance's operational logic, where algorithmic decision-making systems manage high-frequency trading, portfolio rebalancing, risk management, and compliance. In these environments, the human agent is no longer the primary target of behavioural interventions; the institutional interface - its dashboards, alerts, default analytics, and predictive overlays that provide the intelligence of nudging becomes the main target of intelligent nudging. A shift occurs from first-order nudging of individuals to second-order nudging of institutional increasingly rely on AI-based models nudged via behavioural signalling in the algorithmic environment.

Contemporary institutional trading systems are increasingly reliant on AI-based models that have been trained to detect, predict, and act instantaneously to market behaviour. These models don’t just observe behaviour; they affect it by adjusting risk exposures, rebalancing portfolios, or pausing execution on predictive signals based on algorithms that have been trained on market sentiment, volatility, and correlated history. In this way, institutional agents are nudged through data architecture: an alert over volatility tinted red, a risk dial trending downward in a dashboard or an auto-suggested likelihood, for example, represent subtle cues that govern action in a supposedly neutral interface. As Yeung (2019) illustrates in her exploration of ‘hypernudging’, behavioural guidance has transitioned from visible design decisions to data-based dynamic environments that continuously alter the architecture of decision-making contexts.

The established Western economic model assumes that institutions behave rationally to optimise based on complete information. Behavioural economics has contested these ideas for some time. For instance, Simon's (2010) bounded rationality and the heuristics and biases that were identified (Kahneman & Tversky, 1979) have application not only for individuals, but also for the heuristics in institutional mechanisms. For instance, when a portfolio manager uses either earnings or macroeconomic forecasts through an AI-powered interface that presents information in a stylised, simplified manner, cognitive biases such as anchoring may distort interpretation. The manager may overly rely on the initial data display, affecting how they interpret and act upon the information. Furthermore, despite loss aversion being associated with retail investors, it can manifest institutionally. For example, a risk dashboard that emphasises short-term drawdowns while downplaying long-term stability may nudge institutional actors toward overly conservative or reactive decisions. As algorithmic tools become more widespread, these behavioural biases are not eliminated – instead, they risk

becoming embedded within the design of financial systems. The logic of an AI system working to engineer diversification, velocity or speed implies that there is an expansion in the use of AI systems in decision-making. We observe yet another opportunity for automating herd behaviour that develops from features of correlated performance outcomes.

When institutions permit similar forecast models in training (on some overlapping datasets), a common logic underlying jointly held mechanisms to stimulates market convergence, which can happen without coordination. The behaviour reflected here shares characteristics; we frame this behaviour as algorithmic mimicry - a by-product of machine learning driven by competition, where what is frequently optimal in a strategy looks and feels similar to what others do - producing what Danielson and Uthemann (2023) develop as a paradox of a theoretically possibility for diversification to become a practice of convergence. We consider one area that this generates practically: simultaneously recalibrating portfolio changes to 'real-time'. An AI system generates portfolio changes relative to volatility, while large asset managers produce similar adjustments concurrently from similar defensive positions. On a larger scale, this can lead to liquidity issues where the race to sell creates a downward spiral, even if the sale of the assets is not occurring due to any new information. Some have referred to these panics, not as irrational but as rational stampedes. Many of the decisions remain logical when considered in isolation, but collectively create instability in the market system. This is in line with behavioural economist Shiller (2003), who argued that market behaviour is not only determined by fundamentals, but also by narrative contagion and feedback loop behaviours that AI systems can amplify. Importantly, AI models are increasingly engaged with sentiment analysis sourced from retail investor behaviour, social media, and news flow. This creates feedback loops in which institutional algorithms are nudged by the same behavioural signals that they partly create. For example, an increase in search traffic or social media mentions of an asset would likely create shifts in position within algorithmic systems, producing even more relevant mentions and attention. In this way, retail nudging quickly resembles institutional nudging, forming a reflexive relationship that tightens synchronised behaviour in the market.

This reflexivity was framed by George Soros (1994) as market participants' actions are based on their concept of the market, which itself affects the actions of those same participants in a circular way. In the realm of AI, this reflexivity is executed at an unprecedented algorithmic throughput. Institutional actors do much more than respond to the market; they also engage with AI-informed perceptions of the market, which can themselves be informed by previous institutional engagements. The result is a kind of closed feedback cycle of action and engagement that makes markets more aware of minor behavioural cues and less robustly tethered to fundamentals.

Moreover, the default logic baked into algorithmic tools has a different kind of status quo bias. Portfolio management systems, for instance, may come pre-loaded with strategies, benchmark comparisons, and rebalance frequency that make it difficult to diverge from. Samuelson and Zeckhauser (1988) defined a status quo bias as a preference for existing states, even if better states exist. Within institutional AI environments, that bias is structural: to deviate from an existing default requires justification, which adds friction to behaviourally normative rational choices, and reinforces alignment with systemic norms.

The ethical consequences of AI nudging in institutional contexts can be considered more

broadly with retail investors, but with much more serious implications. An individual's misguided decision may only set them back financially; however, an institution, algorithmically nudged into making an error, may create substantial systemic instability. Nonetheless, institutional responsibility remains harder to identify. When AI systems are treated as neutral tools, devoid of normative influences, the behavioural predispositions are no longer questioned. Because the behaviours of AI-infused systems remain opaque, transactions arise from systems with logics only partially understood - even by their designers - real accountability is impossible and regulatory intervention is problematic.

In addition, the absence of behavioural diversity from these systems increases the fragility of the broader financial ecosystem. Behavioural economics has generally recommended that diversity of preferences, heuristics, and time preferences is stabilising and dampens undue consensus-driven overreactions. In contrast, AI nudging seeks to operationalise average behavioural profiles and behave within the social norms of that average behaviour, thereby reducing variation in decision making. When AI nudges distort parameters uniformly across multiple institutions, entire systems become vulnerable to coordinated shocks.

This section has argued that institutional AI nudging is not merely an incidental design feature but is central to institutional market reactions. Behavioural assumptions, risk estimators, and predictive formats are algorithmically inscribed into the activity of institutional actors through procedurally nudged forms of behaviour, not through compulsory action, but through data architecture, interface design and default algorithmic pathways. The consequence of these nudges not only steers the internal decision-making process but collectively shapes market behaviours at scale, evidencing convergence, reflexive behaviours, and systemic fragility

3.3 Market Risks Introduced by AI Nudging

As behavioural nudging is increasingly embedded algorithmically across retail and institutional platforms, the price position and activity will start to reflect not a collection of independent financial judgements, but a converging set of behavioural responses driven by data, design, and digital influence. Each AI-endowed nudging prompt - whether a "stay invested" message, a trending asset suggestion, or an emotion-calibrated notification - may seem insignificant in the scheme of things, but collectively, nudges can fundamentally change the structure and tempo of the market. In this section, we explore how AI nudging presents risks not only to the autonomy of decision-makers but, of potentially greater importance, to the resilience and stability of the market itself.

At the heart of this evolution is a change in decision-making, not just who is making decisions but also how and under what constraints. As theorised by Fama (1970) in the Efficient Market Hypothesis, financial markets assumed that there were several heterogeneous, independent actors reacting to dispersed available information. However, behavioural economists, like Shiller (2003) and Banerjee (1992), have argued that markets are not impervious to herding, imitation, and reflexivity - these social coordination effects in which individuals make decisions based not just on fundamentals, but on what they anticipate others will do. When AI nudging is applied in scale, this will exacerbate previously understood behaviours because the nudges are built into familiar behavioural cues in every

user-facing interface.

In circumstances where AI models reveal similar optimally sounding intervention strategies across large amounts of users, and when platforms have their incentives revolving around maximising engagement and predictability of user behaviours, the nudging environment becomes homogeneous, delivering similar nudges at similar times. For example, during a market downturn, when a staggering number of users could all be nudged to take calm reappraisal, characterising volatility as opportunity, and even others who are nudged to buy low in the realised assets or stay put. While it is quite admirable for these nudges to be apt and perhaps even informed by behavioural prescriptions like loss aversion and status quo bias (Kahneman & Tversky, 1979; Samuelson & Zeckhauser, 1988), deploying these well-intentioned interventions widely sets of behaviours into a correlated condition that worsens the system's sensitivities.

These movements are heightened in a world of frictionless trading and investment interface design. Acting on a financial recommendation in years gone by, to whatever extent it increased the ability for someone to act, required some effort in the execution - you needed to phone a human, leave your home for a meeting, or create effort in protracted deliberative cognition. Today, all you need to act on a nudge is a single click trigger, an immediate transaction or gamified acceptance. The combination of the nudging suggestion and the pathway for execution means that the time between stimulus and behaviour to support the action is reduced, so that a transaction that may have once required consideration before execution has become an automatic transactional response. It is noteworthy, especially in AI-enabled finance, that while Thaler and Sunstein (2008) cautioned these nudges can be powerful even on the margin, now, the sensitivity of markets is not some abstract sense; it is infrastructure.

As previously stated, AI nudging can also cause feedback loops to amplify market movements. If enough users adopt the same signal - buy common stock XYZ or be encouraged not to sell XYZ - the behavioural cumulative action would signal the algorithms used to provide investment advice to users. This represents an instance of reflexivity, as hypothesised by Soros (1994), where perceptions of market participants help constitute the reality they are reacting to. In AI-driven markets, reflexivity is algorithmically mediated: instead of responding to signals and experiences, behavioural data is constantly mined, interpreted, and redeployed as new nudging constructs to baseline recognise and resolve. Thus, we are left with closed loops of self-reinforcing behaviour.

The risk is not volatility, but fragility, where the ability of the market to tolerate shocks is compromised through excessive behavioural coordination. Ecologically, a system will be more resilient when it has more heterogeneous agents responding differently to the same event. However, in a market operating in the predictable personalisation and behavioural architecture of AI human intermediations around similar OKRs, it is doing the opposite. Everyone is being nudged to "buy the dip", or "follow the trend" or "stay invested", depending on who they are and where they are situated in the pre-existing engagement profile mundaneness of proximal inference. We come yet another example of how technology built to 'correct' for irrationality, should it ever manifest at scale, can shed new alternatives of behavioural mapping.

Danielsson and Uthemann (2023) highlight that the widespread adoption of similar AI

models by financial institutions can lead to algorithmic convergence, where these models respond uniformly to market signals. This synchronicity can amplify market movements, potentially causing abrupt surges in order flows and rapid liquidity withdrawals. Such dynamics are not confined to high-frequency trading desks; they are increasingly evident in retail investment platforms. These platforms often employ pervasive, personalized nudges that influence investor behavior, further entrenching systemic patterns of response.

Behavioural instability is a systemic feature of today's digital economy when similar kinds of stimuli (along with push alerts, visual cues, and emotionally tailored framing) are immediately experienced by millions of users concurrently. One of the more subtle risks associated with AI-assisted nudging is illusory autonomy. The user thinks they are behaving autonomously based on personalised cues that encode their preferences, but preferences are themselves shaped, sorted, and reinforced by machine learning algorithms optimised for engagement. Whether intentional or not, algorithmic nudging is a behaviour in an opaque ecosystem, where users either do not know how their subsequent decisions are influenced, and regulators likewise have little transparency on how nudging decisions are made and calibrated. In other words, the illusion of choice masks algorithmic convergence.

Moreover, nudging pathways may diverge from systemic welfare. While platforms optimised for time-on-app or interaction frequency can nudge individuals into a greater engagement, they can also nudge collective resilience down. An "act now" nudge related to an opportunity may generate volume in the short term but detrimentally shift the marketplace's ability to self-correct. Such nudging incentives are rarely overtly revealed to the user, but reinforcing the design and rollout of nudging mechanisms may be commercially rational and behaviourally enticing. AI complicates this picture by allowing emphasised emotional data to inform and shape nudging strategies. As described in Chapter 1.3, favourable algorithms are evolving to read behavioural paths and emotional cues associated with behaviour, allowing a platform to shape nudging based on inferred expressions of emotional function. For example, if a user is showing suggestive signs of anxiety (e.g. after midnight sporadic app usage, continual checking of portfolio, or uncertainty in execution), they may receive nudges assuring their decision (notices), softer (visual) nudges, or a (cognitive) conservative nudge. It is worth noting that while this may ease the emotional burden, it constitutes emotional synchronisation or is where antecedent large cohorts of users are being nudged based on a common effect cue, effectively lowering behavioural dispersion during such times of stress. The consequences of collective emotional synchronisation are worsened at times of market change - when an investor's confidence is vulnerable, the information set is fluid, and volatility is high. At such times, nudge systems that were designed to increase calm in the environment for the investor play into self-induced panic. By universalising the 'calm message', it informs both the algorithms and the investors that fear abounds. Suppose a platform delays negative feedback from the investor when panicked. In that case, they may perpetuate an increasing temporal disconnect between market reality (demand to withdraw, or liquidate) and their perception of what is happening, thus enlarging the expected eventual reaction.

To facilitate new educational pathways, scholars have begun to develop ideas of nudging diversity and behavioural friction. Behavioural friction refers to instances of intentionally inserting time or cognitive effort before deciding to act on high-stakes decisions, to limit reflexivity and favour deliberation. Nudging diversity refers to the platform nudging users

towards relative behavioural diversity, versus the instinct to nudge the user towards average behaviour. Both notions resonate with Gigerenzer's (2015) concept of ecological rationality, which calls for decision environments to be configured not for efficiency but for decisions that promote resilience and adaptiveness.

This section has contended that while AI nudging effectively steers individual behaviour, there are inherent risks when these behaviours are normalised and scaled through a market environment. These systems present threats to structural diversity (the pillars of market stability) by structuring behavioural convergence, reinforcing emotional synchrony, and advancing reflective loops. The following section will outline empirical instances of AI behaviours in the form of case studies relevant to AI discursively playing a role in the acceleration of behaviours that have already led to moments of volatility and disruption.

3.4 Case Studies on Market Volatility

While in theory, the risks of AI-assisted nudging are becoming clearer, the ramifications of their systemic impacts are perhaps best situated within the context of instances of market disruption. In these episodes, the behavioural and algorithmic architectures discussed in this thesis demonstrate their potential to destabilise, not in abstract form, but in the empirical language of liquidity shocks, price cascades, and surges in volatility. This section considers three representative instances: the Flash Crash of 2010, the cryptocurrency boom-bust cycles, and the GameStop short squeeze of 2021. Each instance demonstrates how emerging examples of digital nudging, behavioural synchrony, and algorithmic reflexivity combine to destabilise markets in instances with high volume, real-time, AI-mediated behavioural signals.

The 2010 Flash Crash

The Flash Crash of 2010, while not caused by AI nudging, is crucial for understanding how autonomous digital agents can have systemic behavioural consequences when they interact at scale. The U.S. equity market experienced gaffes, arguably among the most rapid and catastrophic dislocations, on May 6, 2010. The Dow Jones Industrial Average fell nearly 1,000 points, and as quickly as it fell, it regained those losses. While many initially attributed the dislocation to a "fat finger" error, deeper investigation, including a seminal analysis by Kirilenko (2017), found that the cause of the crash was autonomous algorithmic trading machines reactively responding to one another, combined with market participants withdrawing liquidity.

While this incident shows how elements of high-frequency reflexivity can develop, in this case, trading algorithms reacted not to price movements but to anticipated reflexive responses from other algorithms. The perturbation the algorithms were witnessing in the market was manifesting as downward momentum and therefore provided a behavioural signal to exit the system, causing a self-reinforcing feedback loop. Individually, there was likely rationality to each decision, but collectively, the uncertainty that developed at each decision point destabilised the market. Even at the time of the crash, there was little evidence of AI nudging in retail and technology-mediated environments that would compare, so they provide a structural parallel. It illustrates, at the level of behaviours, Soros's (1994) concept of reflexivity and Shiller's (2003) discussion of the influence of emotions and momentum on the

price of an asset.

If such systemic instability can emerge from autonomous machine-to-machine interactions in professional trading, risks may be more pronounced if AI nudging is used to influence millions of users at once, creating synchronised, amplified behavioural patterns of contagion that could be destabilising at the market level.

Cryptocurrency Markets

The territory of cryptocurrency provides a contemporary example of behaviourally amplified algorithmic systems. Assets like Bitcoin and Ethereum have experienced astronomical highs followed by extreme price declines, often tied to sentiment, virality, and speculative reflex, rather than fundamentals. This behaviour is heightened by platforms that use nudging techniques powered by AI (e.g., push notifications about price movements; entertaining trading prompts), where the platforms are engineered to attract users' attention and motivate them to act. What differentiates this realm from others is the lack of traditional valuation anchors. With no cash flows or intrinsic estimates, the price of a cryptocurrency is oriented solely to perceived momentum and collective belief. Tirole (1982), in his exploration of speculative bubbles, showed how asset prices can become detached from fundamental value when driven primarily by expectations of future resale. This dynamic renders markets vulnerable to instability. Later, Abreu and Brunnermeier (2003) expanded on this by demonstrating how coordination problems and delayed arbitrage among rational traders can sustain bubbles and lead to informational cascades. Nudging in this regard takes on an almost dramaturgical form. Binance and Coinbase, among other platforms, frequently provide nudges laden with availability bias and suggest herd behaviour by spotlighting the "most traded," "most volatile," or "most held" tokens instead of giving suggestions. These nudges are typically presented not as suggestions to buy tokens, but as neutral informational nudges. However, as Bikhchandani (2000) notes, even negligible signals can trigger cascades of coordinated actions when individuals lack independent signals or adequate time to reason about their actions.

AI-based nudging increases this phenomenon by providing prompts to the individual (customised nudges based on individual behaviour). For example, a user who habitually engages when trends are upwards may receive prompts that trigger that behaviour, while a user showing signs of indecision might receive emotionally supportive nudges, prompting the individual to feel more confident participating. Over time, these feedback cycles embed users into behavioural grooves that are commensurate with the platform's incentives (e.g., prompts that cause frequent trading/engagement), as opposed to better outcomes for the investor in the longer term.

As a result, the transition to a market prioritising salience, emotional state, and momentum over strategy occurs. Oftentimes, there are peaks and falls in cryptocurrency prices that result from not just sentiment/excited refrains from social media, for example, but also coordination prompts from the platform/device, and macroeconomic signals are weakly, if ever, linked to price adjustments. In these markets, AI-Nudging does not filter out bias - it amplifies it - by directing attention to assets that will likely induce action on one's part.

GameStop and the Meme Stock Phenomenon

An especially powerful instance of behavioural convergence over time is the GameStop short

squeeze in January 2021. The events started relatively innocuously, developing as an attempt on Reddit to challenge the institutional short-sellers, before quickly morphing into a global financial spectacle. Retail investors - acting together through the social media platform r/WallStreetBets – were able to drive GameStop shares from $17.25 to $483 in three weeks - an increase of roughly 2,700%. Many of the characteristics of this convergence were motivated not simply by the markets in the traditional sense but by a developed collective identity derived through digital communities: a sense of rebellion, active participatory belonging, and an emotional experience.

Platforms like Robinhood affected this convergence behaviour, simplifying, gamifying, and reducing transaction friction so that everyone could trade transparently and feel expressive behaviour over the trades. More importantly, the algorithmically tailored nudges, reflecting the top-traded stocks, surfacing social sentiment, and revealing price momentum, were firmly back-feeding on attention and user action. These capabilities, derived from AI-driven environments, either in engagement or information exchange with content users, were curated along their behaviours, empowering mimesis and a collective sentimental reaction beyond a rational valuation.

The importance of the GameStop episode rests on dynamics of price volatility but also revealed how retail trading could be linked to emotionally charged collective identity (which is difficult to identify) and the prominence of AI-personalised environmental nudging, which can result in a synchronisation of behaviours and activity that could be destabilising. Herding can be thought of as any cognitive and affective behaviours along a collective digital narrative reinforced in real-time through the nudging systems that have taken direction from their user interaction experience.

Robinhood's decision to limit trading at the peak of the squeeze because they could not make enough liquidity offended many and opened a whole other administrative oversight process. It posed interesting questions about the governance gap in AI-mediated environments: what responsibility do platforms have when AI-mediated systems help create, elevate, and emotionally fuel mass financial behaviour? The incident illustrates the potential of AI nudges beyond simply stabilising on task performance, but allowing for systemic shifts in behaviour with no general accountability.

Conclusion: From Case to System

Collectively, the case studies showcase the predictable consequences of the mechanisms discussed throughout the chapter. Whether in terms of the millisecond reflexes of a Flash Crash, the emotive cycles of cryptocurrency, or the mimetic mania of meme stocks, there is a consistent pattern: AI Nudging interacts with the naive and faltering functionality of behavioural heuristics to produce highly coordinated and often volatile outcomes. The theoretical consequences of convergence, reflexivity, and homogenisation become more materially realised in liquidity constraints, price bubbles, and social-financial contagion.

They are not merely anomalies - they also manifest the emergence of a financial architecture with algorithmic persuasion, behavioural psychology and digital architecture as stakes that increasingly mediate market behaviour. They require not only ethical reflection but institutional supervision and regulatory imagination. Ultimately, as the last chapter will discuss, the future of AI nudging will depend upon transforming into frameworks to promote transparency, behavioural variance and systemic resilience.

CHAPTER 4 – CONCLUSION

5.1 Reaffirming the Research Question and Contribution

This thesis was intended to answer a deceptively simple research question: How does nudge theory, in a world where AI mediates nudging, shape investor behaviour and the architecture of the financial system? As this work progressed, it became apparent that the question was far more complex, as it relates not just to behavioural economics and digital finance, but also to the very architecture of decision-making within a computationally based world. What began as a simple examination of financial nudging became a broader interrogation of how we understand and distribute autonomy, risk, and responsibility when decision-making gets obliterated by computational optimisation.

This thesis is multidisciplinary. It draws on behavioural economics, financial systems theory, and AI ethics, and provides an integrative lens for our increasingly present technological reality. By theoretical analysis, literature critique, and conceptual synthesis, this thesis has demonstrated that AI nudging should not be understood as a new version of earlier choice architecture; rather, it is a qualitatively distinct form of behavioural governance. One that is contingent, highly personalised, and largely inscrutable.

In doing so, this research has provided several conceptual tools to help understand this transition. Terms like second-order nudging, behavioural monoculture, and algorithmic reflexivity offer some vocabulary to characterise an environment where behavioural influence can extend at multiple levels: individual, institutional, and systemic. By taking a classical behavioural theory perspective, extending these theories into modern digital environments, and demonstrating how these behaviours scale into vulnerabilities at market-level, this thesis contributes to an emerging discussion on the ethics and structure of AI-mediated finance.

5.2 Synthesis of Key Findings

This thesis has followed a coherent trajectory, beginning with the theoretical basis of behavioural nudging and moving toward the technological change, empirical demonstration, and systemic implications of nudging. The chapters have each represented an element of this progression, ultimately revealing how AI-supported nudging is reshaping financial behaviour and the fabric of modern market systems.

In Chapter 1, the conceptual groundwork of nudge theory was established, beginning with the theoretical evolution of nudge theory from behavioural economics. It began with the recognition of Simon (2010), Kahneman & Tversky (1979), and later Thaler & Sunstein (2008) that cognitive limitations, heuristical biases, and affective framing bound human decision-making. Nudging in its classical form was intended to be a light-touch corrective, a way to structure choice architectures to position individuals to obtain better outcomes without coercion.

As described in Chapter 1, artificial intelligence has changed the framework, scale, and nuances of nudge interventions significantly. Nudging was once thought of as static choice architectures and now represents a shift to dynamic, adaptive, and hyper-personalised

systems. Nudges now operate in systems (often through 'as-a-service' models) that are continuously refined with machine learning systems that update based on user behaviour. This has developed into what Yeung (2019) and Nyman (2023) call 'hypernudging', where the distinction between nudging (behavioural support) and behaviour engineering becomes less defined. Chapter 1 provided the theoretical lens that frames this thesis and provides a conceptual tool for analysing AI-driven nudges' empirical dynamics and ethics.

Chapter 2 transitioned from theory to practice and examined how AI nudging works individually, specifically in retail finance, personal investing, and trading platforms. These technologies often carry with them nuance and the potential for genuine benefit. Nevertheless, while they purport to address well-documented behavioural biases among retail users, such as present bias, loss aversion, and overconfidence, they frequently operate by subtly exploiting those very same vulnerabilities. Technologies, such as nudging and gamification, through real-time nudges, gamified interfaces, and emotionally calibrated messaging, reinforce behavioural proclivities, those narratives that may serve financial firm interests more than user welfare. In a way, these technologies put users in a position of dependency, not empowerment.

The chapter also showed that AI systems personalise nudges based on financial history and psychological dispositions built up using inferences. Therefore, the behavioural environment presented to the user is specific, affective, and persuasive, but also opaque. While the intention may be benevolent, user autonomy, literacy, and self-determination outcomes are profound.

Chapter 3 followed Chapter 2 with a systematic analysis of AI nudging and how it works to reshape financial markets through behavioural synchronisation. As platforms share the same types of nudges with all their users, retail and institutional users tend to converge homogeneously. This chapter conceptualises this idea into a ‘behavioural monoculture’, whereby market agents respond reflexively, creating the risk exposures of herding, volatility and contagion.

Case studies of the 2010 Flash Crash of the US stock market, the cryptocurrency cycles 2020, and the GameStop phenomenon demonstrated the real-world and social consequences of reflexive markets, which are technically nudged through algorithms. These occurrences should not be seen as anomalistic, but symptomatic signs of the systemic, structural shifts in behaviour and highly automated, reflexive transactions that digital nudges, emotional states, and algorithms create to function as consolidated systems that ingest and deploy behaviours that can be seen as efficient, seductive, and susceptible to failure.

5.3 Ethical and Regulatory Implications

As AI nudging turns from an accent to a meaningful, definitional characteristic of digital financial service platforms, ethical questions become a primary concern. In the past, the design of behaviour in interfaces and services was more akin to behaviour governance, where platforms not only track and predict user behaviour, but more importantly, fundamentally intervene and shape user choice through a myriad of finely tuned nudges. This thesis has detailed that the incidences of AI nudging exist on a continuum, ranging from positive

corrective impacts to profoundly negative exploitative influences. The question is not whether AI nudges work, but what the end goal is and what values are respected or disregarded.

One ethical issue is the lack of transparency and asymmetrical influence. Users are frequently unaware, in real time, that they are being nudged, let alone how the nudging is explicitly designed for their individual psychological and emotional predicates. An extreme contrast exists in the information available to platforms with innumerable behavioural datasets and high-powered optimisation capabilities to construct decision environments at remarkable resolutions. The degree of transparency and imbalance of information is concerning, as autonomy is viewed as intact in structural terms but essentially eliminated in substance.

The areas of moral challenge become deeper when the nudging practices are consistent with platform revenue models. As a result, behavioural nudging may prioritise engagement over consumer dignity, and decision architecture may become a form of soft manipulation instead of a source of behavioural support. The behavioural literature shows that cognitive biases and inappropriate behaviours are consistently present. However, when these same behaviours are stitched to increase app usage, market high-fee products with substantial consumer impact, or synchronise responses for the broader marketplace, the architecture of influence crosses a line for which there may be no ethical criteria.

The underlying behaviours of nudging practices underline the need for rethinking regulation. Regulatory bodies have historically dealt with behaviours regulated through disclosures, risk assessments, and appropriate investor protection. However, the architecture of AI-driven nudging demands behaviourally informed regulation, beyond simply identifying the choices available on a platform at any given moment. It must also account for how those choices are presented within a uniquely constructed choice architecture, where emotional framing can significantly shape user perception and decision-making. Regulators should take this further by requiring platforms to disclose their nudging mechanisms, detailing when, how, and why behavioural interventions are initiated.

A potentially radical idea would be to see frameworks that promote behavioural diversity by design. This would also help provide choice architecture that does not over-optimise individual preferences, limit ubiquitous synchronisation of user bases, or create excessively coherent private behaviour. Introducing friction in the form of confirmed seconds, or time to reflect or deliberate when engaging with the technology, should also be considered.

Ethical evaluation frameworks like FORGOOD (Fairness, Openness, Respect, Goals, Outcomes, and Dignity) (Sunstein, 2014) provides a behavioural ethics lens through which AI-enabled nudging can be evaluated. Rather than viewing these values as abstract ideals, they can be operationalised as design and oversight principles that help platforms assess whether nudging strategies support or undermine user autonomy. Integrating such models would shift regulatory conversations from formal compliance to normative alignment, encouraging systems prioritising user welfare in intention and measurable outcomes.

AI ethics or responsible practices in finance, beyond passive acceptance of compliance, require individual and collective commitments to behaviours that respect human and behavioural autonomy. This requires building systems designed to optimise for better outcomes, honour human dignity in the context of choice, and pivot to consumer wellness, especially where human financial, emotional, and social well-being are at stake.

5.4 Limitations of the Thesis

While this thesis aims to deliver a comprehensive and interdisciplinary understanding of AI-powered nudging in financial contexts, certain limitations must be noted. First, the thesis is ultimately theoretical and conceptual. Although rooted in extensive and current literature, no empirical testing had been conducted. The behaviour and dynamics the thesis investigates are captured conceptually through theoretical synthesis, case study analysis, and existing literature rather than through primary experimentation, observation, or user-based studies. Second, the thesis examines Western financial infrastructures and digital markets, specifically the United States and Europe. Nudging strategies, behavioural reactions and regulatory environments can differ significantly between geographical, cultural, socio-economic, legal and political contexts. While there are similarities between how AI and algorithms are regulated and analysed across the Global North, a global comparative lens would provide further nuanced opportunities for understanding, especially in places where financial digitalisation occurs from very different behavioural underpinnings. Third, the technology itself is rapidly changing. AI systems are becoming increasingly autonomous and effective with each iteration and release. Therefore, the conclusions drawn in the thesis are also temporal, and the implications of future developments - whether in affective computing or decentralised finance - could mean these conclusions will need to be revised in the future. These limitations do not detract from the value of the thesis but highlight potential opportunities for future research.

5.5 Directions for Future Research

The issues presented in this paper represent not a termination but an initiation of inquiries. The meaning of this research, particularly in terms of the more expansive "nudge", requires ongoing and interdisciplinary engagement as the context in which AI orienting becomes embedded in our digital architecture for finance. One important aspect of that engagement resides in empirical testing. While a theoretical engagement to nudge has been presented in this thesis, future researchers can experiment and longitudinally study how AI-nudges may work overtime to change financial behaviours, specifically, with user demographic characteristics, such as literacy, emotional preparedness, and autonomy.

Another possible area is considering the psychological after-effects of long-term exposure to algorithmically mediated outcomes for decisions. For example, how might users learn, develop or appropriate the logic of nudges? Does reliance on AI undo critical reflection of financial decision-making or fritter away the confidence to act independently? These are important questions as nudges transit from infrequent and exploratory to ubiquitous forms of behavioural infrastructure.

Finally, one of the most important aspects is collaboration and conversation between behavioural economists, AI ethicists, legal scholars, and technologists. If those conversations continue, future systems are indeed likely to be designed in ways that build on behavioural insight and respect for agent integrity.

5.6 Final Reflection: Behavioural Autonomy in an Age of Algorithms

At its foundation, this thesis has been an inquiry study not only in digital finance or behavioural design but in an even longer-lasting inquiry: What is autonomy within an age of algorithmic influence? As platforms become increasingly adept at predicting and installing human behaviours, the conditions for deciding begin to alter. The interface of mediation - now with a voice, quiet yet firm; computational yet effective.

AI nudging offers an opportunity to rethink freedom. When moments of choice are continually nudged, as soft, invisible and pervasive, which obliterates autonomy as any interesting form of simulation, a choreography of building preferences, as opposed to ones found, nudging embeds outcomes for an individual. Nevertheless, the moment also opens a way - design a system that may assist in choosing instead of replacement; correct for biases and delete deliberations; and remind us that dignity and guidance need not be mutually exclusive.

As this thesis has articulated, the projects ahead matter not only for problems of savings but also for solving societal and ethical challenges; as we consider the value of intelligent or automated systems, we cannot lose sight of designing a system that reflects our better values - transparency, restraint, variability, and honouring the fine art of good choices.

BIBLIOGRAPHY

- Abreu, Dilip, and Markus K. Brunnermeier. “Bubbles and Crashes.” Econometrica, vol. 71, no. 1, Jan. 2003, pp. 173–204.
- Alemanno, A., 2015. Nudge and the Law: A European Perspective. Hart Publishing.
- Banerjee, A. V. “A Simple Model of Herd Behaviour.” The Quarterly Journal of Economics, vol. 107, no. 3, 1 Aug. 1992, pp. 797–817.
- Barber, Brad M., and Terrance Odean. “Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors.” The Journal of Finance, vol. 55, no. 2, Apr. 2000, pp. 773–806.
- Bikhchandani, S. and Sharma, S. “Herd Behaviour in Financial Markets.” IMF Staff Papers, vol. 47, no. 3, July 2000, pp. 279–310.
- Binns, Reuben. “Algorithmic Accountability and Public Reason.” Philosophy & Technology, vol. 31, no. 4, 24 May 2018, pp. 543–556.
- Cai, Cynthia Weiyi. “Nudging the Financial Market? A Review of the Nudge Theory.” Accounting & Finance, vol. 60, no. 4, 28 Mar. 2019, pp. 3341–3365.
- Caliskan, A., Bryson, J.J. and Narayanan, A., 2017. Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), pp.183–186.
- Danielsson, Jon, and Andreas Uthemann. “On the Use of Artificial Intelligence in Financial Regulations and the Impact on Financial Stability.” 2023.
- Felsen, G., Castelo, N. and Reiner, P.B., 2013. Decisional enhancement and autonomy: Public attitudes towards overt and covert nudges. Judgment and Decision Making, 8(3), pp.202–213.
- Financial Stability Board. The Financial Stability Implications of Artificial Intelligence. 2024.
- Gigerenzer, G., 2015. Simply Rational: Decision Making in the Real World. Oxford University Press.
- Hull, C.L. “The Goal-Gradient Hypothesis and Maze Learning.” Psycnet.apa.org, Psychological Review, 39(1), 1932, pp. 25–43.
- Kahneman, D., 2011. Thinking, Fast and Slow. Farrar, Straus and Giroux.
- Kahneman, Daniel, and Amos Tversky. “Prospect Theory: An Analysis of Decision under Risk.” Handbook of the Fundamentals of Financial Decision Making, vol. 47, no. 2, 11 June 1979, pp. 99–127.
- KIrilenko, A. “The Flash Crash: High-Frequency Trading in an Electronic Market.” The Journal of Finance, vol. 72, no. 3, 21 Apr. 2017, pp. 967–998.
- Laibson, David. “Golden Eggs and Hyperbolic Discounting.” The Quarterly Journal of Economics, vol. 112, no. 2, 1 May 1997, pp. 443–478.

- Leal, Cristiana Cerqueira, and Benilde Oliveira. “Choice Architecture: Nudging for Sustainable Behaviour.” Sustainable Management for Managers and Engineers, 11 Dec. 2020, pp. 1–17, https://doi.org/10.1002/9781119804345.ch1.
- Loewenstein, George, and Richard H Thaler. “Anomalies: Intertemporal Choice.” Journal of Economic Perspectives, vol. 3, no. 4, Nov. 1989, pp. 181–193.
- Loewenstein, George, et al. “Projection Bias in Predicting Future Utility.” The Quarterly Journal of Economics, vol. 118, no. 4, 2003, pp. 1209–1248.
- Nissenbaum, H., 2009. Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford University Press.
- Nyman, Stig. “The Birth of AI-Driven Nudges.” Research.cbs.dk, Hawaii International Conference on System Sciences (HICSS), 2023.
- Odean, T. “Volume, Volatility, Price, and Profit When All Traders Are Above Average.” The Journal of Finance, vol. 53, no. 6, Dec. 1998, pp. 1887–1934.
- Rebonato, Riccardo. “A Critical Assessment of Libertarian Paternalism.” Journal of Consumer Policy, vol. 37, no. 3, 18 Aug. 2014, pp. 357–396.
- Sadeghian, Armindokht H, and Ali Otarkhani. “Data-Driven Digital Nudging: A Systematic Literature Review and Future Agenda.” Behaviour & Information Technology, 29 Nov. 2023, pp. 1–29.
- Samuelson, W. and Zeckhauser R. “Status Quo Bias in Decision Making.” Journal of Risk and Uncertainty, vol. 1, no. 1, Mar. 1988, pp. 7–59.
- Shefrin, H. and Meir S. “Behavioural Portfolio Theory.” The Journal of Financial and Quantitative Analysis, vol. 35, no. 2, June 2000, p. 127.
- Shiller, R. J. “From Efficient Markets Theory to Behavioural Finance.” Journal of Economic Perspectives, vol. 17, no. 1, Feb. 2003, pp. 83–104.
- Simon, H.A. A Behavioural Model of Rational Choice. 2010.
- Soros, G. The Theory of Reflexivity. 1994.
- Sunstein, Cass R. “The Ethics of Nudging.” Papers.ssrn.com, 20 Nov. 2014.
- Sunstein, Cass R. Why Nudge?: The Politics of Libertarian Paternalism. Yale University Press, 2014.
- Thaler, R.H. and Sunstein, C.R., 2008. Nudge: Improving Decisions About Health, Wealth, and Happiness. Yale University Press.
- Thaler, Richard H., and Shlomo Benartzi. “Save More Tomorrow: Using Behavioral Economics to Increase Employee Saving.” Journal of Political Economy, vol. 112, no. S1, Feb. 2004, pp. S164–S187.
- Tirole, J. “On the Possibility of Speculation under Rational Expectations.” Econometrica, vol. 50, no. 5, Sept. 1982, p. 1163.
- Wagner, B., 2021. Ethics as an escape from regulation: From ‘ethics-washing’ to ethics-shopping? In Being Profiled: Cogitas Ergo Sum. Amsterdam University Press.

- Yeung, K. “Hypernudge”: Big Data as a Mode of Regulation by Design.” Social Science Research Network, 23 Oct. 2019, pp. 118–136.

- Zuboff, S. “The Age of Surveillance Capitalism.” Social Theory Re-Wired, 15 Jan. 2023, pp. 203–213.