



**IMSIS**  
International Master  
Security, Intelligence  
& Strategic Studies



**Erasmus  
Mundus**

# **Mindful Maneuvers:** A Normative Framework for Disinformation Defence Using Cognitive Security

August 2023

University of Glasgow: 2686693  
Dublin City University: 21109478  
Charles University: 81164422

**Presented in partial fulfilment of the  
requirements for the Degree of International  
Master in Security, Intelligence and Strategic  
Studies**

Word count: 20,013

Supervisor: Dr. Vit Stritecky, M.Phil., Ph.D.

Date of Submission: 02-08-2023



University  
of Glasgow

**DCU**



CHARLES  
UNIVERSITY

## Abstract

In an era marked by digital interconnection, the phenomenon of disinformation has evolved into a critical challenge to both individual and collective cognitive security. This thesis identifies disinformation as not merely a byproduct of the information age but as a tactical weapon, wielded by various state and non-state actors to influence, distort, and destabilize, with the potential to sway democratic processes. The current global landscape, characterized by polarization and hybrid warfare, has elevated the role of disinformation in exploiting vulnerabilities to erode societal trust and credibility. The thesis acknowledges the limitations of conventional security measures in countering disinformation, and instead advocates for translating proactive strategies from cybersecurity such as encryption, threat modelling, and constant monitoring into the cognitive security domain. A comprehensive, pioneering framework is proposed that integrates artificial intelligence, machine learning, cognitive psychology, and other disciplines, aiming to provide robust protection against disinformation's insidious effects. The proposed framework emphasizes privacy by design and insists on a strict data trail to mitigate abuse. Potential constraints, such as practical implementation hurdles, consent overhaul, resource allocation, and bias in data collection, are critically examined. Recommendations are outlined for continuous refinement, focusing on streamlined structures, proactive data collection approaches, explainable AI applications, and real-world testing. The thesis ultimately serves as a foundation for an ethically informed, proactive approach to safeguarding cognitive security in our interconnected world, recognizing the need for ongoing adaptability and refinement in the face of technological advancement and evolving disinformation tactics.

## Introduction

As we find ourselves living in a world which, through its information superhighway, has become increasingly interconnected, the phenomenon of disinformation has emerged as a grave challenge to both individual and collective cognitive security. It is crucial to understand that disinformation is not just a byproduct of our information age but a tactical weapon that can be used to influence individuals and groups, distort public opinion, destabilize societies, and even sway the course of democratic processes. The current geopolitical landscape, rife with conflict, polarization, and an increasing propensity for hybrid warfare, has made disinformation a tool of choice for various state and non-state actors. These actors aim to manipulate the cognitive environment of target populations, exploiting social, cultural, and political vulnerabilities to sow discord and mistrust, erode social cohesion, and undermine institutional credibility. The stakes are high. Our collective and individual cognitive security, which refers to protecting our minds from manipulation and deception, is at risk. However, the multifaceted and complex nature of this challenge renders conventional security measures and reactive solutions inadequate to address the impact and instability caused by disinformation.

In the realm of cybersecurity, substantial strides have been made towards creating normative frameworks and policy initiatives that proactively address the digital threats our society faces. Consequently, this thesis posits that these established cybersecurity paradigms can inspire a similarly proactive approach to combat disinformation, a concept we henceforth refer to as 'cognitive security.' Cognitive security refers to the practices and methodologies to defend against social engineering attacks that aim to disrupt cognition. In the field of cybersecurity, it refers to the use of Artificial Intelligence and Machine Learning technologies modelled on human cognition to detect possible security threats (COGSEC, 2017). The premise of this thesis lies in the conceptualisation of cognitive security as an analogous field to cybersecurity, thereby allowing us to extrapolate and adapt the strategies used in the latter to address the former. The existing body of cybersecurity frameworks has created an ecosystem of defence

mechanisms which operate not just reactively but proactively. They are designed to anticipate potential threats and neutralise them before they cause harm, rather than merely patching up the damage post-incident. This research aims to explore the potential of this preventive stance in the context of cognitive security, with an aspiration to fortify society against the onslaught of disinformation. The goal is identifying risks and anomalies that might elude conventional, algorithmic security protocols. The use of cognitive security to defend against disinformation attacks is stemmed from the growing concord of wariness towards the threat posed to national security by information warfare (Vasu et al., 2018). The endeavour to construct a proactive normative framework for cognitive security raises many complex questions and challenges. How do we balance the quest for cognitive security with the imperative to preserve individual liberties, most notably, data privacy? How can we combat the impact of disinformation on fragile societies? What technologies are utilised in such a process, and what risks do they pose to society? Furthermore, finally, how can we mitigate those said risks? This thesis seeks to grapple with these questions, arguing for the urgent necessity of a proactive approach whilst remaining cognisant of the need for this approach to be both robust and ethically sound.

In cybersecurity, we have seen how effective proactive measures have helped secure our digital environment. Pre-emptive software updates, encryption, threat modelling, and constant monitoring are just a few examples of the tools and strategies employed to avoid potential threats. Therefore, this thesis aims to translate these strategies to fit within the cognitive security paradigm to provide a similar degree of protection against disinformation and its impact. In conclusion, the thesis aims to provide a roadmap for the development of a normative framework for cognitive security, founded on the belief that just as we have shielded our digital worlds from the perils of cyber threats, so too can we shield our minds from the pervasive, insidious onslaught of disinformation. It is hoped that the insights generated here will serve as a basis for constructing a practical, proactive, and ethically informed approach to safeguarding cognitive security in our information-rich, digitally interconnected world. In the following sections, we discuss the threat of disinformation and how it has risen

to affect national security. We explore the current efforts in combating this phenomenon and address the lack of a robust system that facilitates an end-to-end risk analysis of disinformation especially in today's polarised society. We then go on to propose a novel methodology for such a contraption and lay out a framework that provides systematic guidelines for each of potential pitfalls associated with various processes in this novel contraption in later stages. Finally, we also discuss the limitations of this method and potential areas of improvement in future.

## Background

The adage 'Sticks and stones may break my bones, but words do not hurt me' is far from the truth in today's day and age. Social media penetration has radically reinvented how information is accessed, shaped, and spread. People credit the advent of the internet as being revolutionary regarding communication. While it is true, it was within the line of how the world perceived the introduction of the telephone, radio, and television. The internet, however, paved the way for the development of social media, which has ruptured the fabric of reality in many ways. Suddenly, the political discourse was being boosted by social media platforms' participatory environment, where diplomacy has become less private and more performative. However, the information diffusion on these platforms is driven by the popularity of the content. Thus, conflicts of popularity and perception in the virtual world began to take effect in the real world (Singer and Brooking, 2018). The participatory nature of these platforms that were mentioned before introduced a means for everyone to participate in this fight for attention and outrage that is easily accessible regardless of the physical borders. Countering voices were ever present, leading to an endless cycle of global information conflict that did not exist before (Singer, and Brooking, 2018).

Amongst these cycles, few have leveraged information diffusion to their benefit, sometimes at the expense of democratic values that went unnoticed for a while. Some of the well-known instances of such occurrences include the political

campaigns and rise to power of Donald Trump and Narendra Modi, and the use of the internet by the Islamic state during the invasion of Iraq. Trump and Modi's rebranding process involved drawing the masses' attention and nudging their participation into their narratives. The invasion of Iraq noticed the use of hashtags where the planned attack was being masqueraded as a brand campaign that used virality on social media to conflagrate the violence, thus leading to false perceptions of the Iraqi army. This tactic was not unique, though. The Germans used radio to bombard the French people with a continuous barrage of propaganda, causing fear and doubt and leading to confusion about the troop movements. This method of targeting the enemy through other means was not always successful. War propaganda during the German Blitz failed to be effective in Britain and ended up being a joke during World War 2 despite this radio channel being the most popular.

Conversely, tens of millions of leaflets air-dropped by the United States in Vietnam were used as toilet paper. Social media, in a decade, changed that. Attacking the adversary's spirit, i.e., the spirit of the people, no longer requires military action and can be done with a smartphone. It is very easy to communicate with people you conflict with, debate, and persuade them apart from stalking their digital lives. These interactions can lead to further hate when more sympathisers on either side join this metaphorical tug of war. As such, social media has become a weapon in light of these newfound arms that can be salvaged on the internet. Singer and Brooker (2018) argue that just as the internet has reshaped war, war is now radically reshaping the internet. Progress in technology has always found itself to the ends of war despite the original creators' optimism for a social medium and world peace. Internet reshaping war followed the same patterns as the telegraph, telephone, printing press or radio that came before but fundamentally capable of performing all of its predecessors' strengths. The internet is neither limited to one-on-one linkages like telephone or telegram nor print media and radio broadcast capabilities.

However, the key merit highlighted by each of these is being able to communicate within the period of time to when the interventions alter the outcome. This proved to be a double-edged sword in recent times. In

contemporary discourse, this kind of propaganda took the label of 'fake news', with its surge influenced by social media and the emergence of data analytics firms like Cambridge Analytica that caused widespread commotion worldwide.

The 2018 Cambridge Analytica scandal significantly altered discussions around digital privacy and data usage, highlighting their potential impacts on democratic structures. This controversy centred on accusations against Cambridge Analytica, a former political consultancy, claiming misuse of about 87 million Facebook users' personal data. This was reportedly used to devise intricate and personalised advertising during two key events: the 2016 US presidential election and the Brexit referendum (Daswani et al., 2021). This case underlines the profound societal implications of unregulated data harvesting. This unauthorised data acquisition, executed via a seemingly innocuous personality quiz application, marked an unprecedented violation of user privacy. This underlines the digital era's inherent instability regarding personal data and the possible manipulation of such information by malign forces for political advantage. The scandal underscored the urgent need for robust regulatory frameworks to govern the gathering, storage, and utilisation of personal data in the digital realm. Furthermore, it highlighted the power dynamics at play in this context, where corporations and political entities can manipulate public sentiment and democratic processes through the strategic application of personal data. Since the incident, the world has grown more concerned about fake news, the spread of which in social networks poses threats to social stability, economic development, and political democracy (Guo et al., 2021).

Cambridge Analytica's influence on people's minds was predicated on a detailed, algorithmically driven approach to microtargeting. The complex strategy under discussion pertains to a refined form of psychographic profiling, which involves gathering, processing, and applying a multitude of personal data from varied digital sources, focusing on platforms like Facebook. Sophisticated machine learning tools were used to build models of psychological characteristics based on the 'Big Five' or 'OCEAN' personality framework. This includes openness, conscientiousness, extraversion, agreeableness, and neuroticism, illuminating the inherent tendencies of individuals. Armed with

this knowledge, political advertisements were personalised to evoke certain emotions and microtargeted towards fence-sitters unsure about their standing to influence political perspectives in favour of a particular side (Omelianenko, 2017). The critical technique in this process exploited cognitive biases, notably confirmation bias, which refers to the tendency of people to agree with and trust information that aligns with their pre-established beliefs (Kim and Dennis, 2019). In essence, Cambridge Analytica's methodology enabled the strategic steering of public opinion. This showcases a fresh but morally controversial aspect of the intricate intersection of technology, psychology, and politics. Figl et al. (2019) underscore the persistent challenge of mitigating the dissemination of misinformation and the concomitant cognitive dissonance engendered by it, despite the concerted efforts to combat these phenomena by deploying sophisticated fake news and malicious bot detection mechanisms.

As the digital age continues to evolve, the proliferation of disinformation presents an increasingly significant societal concern. Several conceptual and operational frameworks have been introduced to address this challenge, each with varying degrees of success. Among the most widely employed frameworks in combating disinformation is fact-checking, represented by PolitiFact and Snopes. This approach leverages the scrutiny of professional fact-checkers to authenticate content and refute false claims (Lazer et al., 2018). While fact-checking initiatives contribute significantly to debunking disinformation, they grapple with multiple limitations. The fact-checking process is often time-consuming and lacks the scalability to keep pace with the real-time dissemination of false information (Pennycook & Rand, 2020). Furthermore, the 'backfire effect', where debunking attempts inadvertently reinforce false beliefs (Lewandowsky et al., 2012), poses a significant challenge.

The second major approach to mitigate disinformation involves algorithmic content moderation, which is extensively used by social media platforms. This automated method employs artificial intelligence and machine learning (AI/ML) to identify and mitigate the spread of disinformation (Zhou & Zafarani, 2020). Its primary strength lies in its scalability and ability to rapidly process vast volumes of data. However, algorithmic approaches are not without

limitations. False positives and negatives persist due to disinformation's complex and ever-evolving nature.

Further, the 'black box' problem – where algorithms operate in opaque ways – raises transparency and accountability issues (Pasquale, 2015). Media literacy programs represent another critical framework that empowers individuals with the skills to distinguish credible from false information (Hobbs, 2010). The strength of this approach lies in its focus on developing individuals' resilience against disinformation, offering a proactive strategy. However, the effectiveness of media literacy programs can be hampered by cognitive biases such as confirmation bias and the Dunning-Kruger effect, leading people to overestimate their ability to detect false information (Pennycook & Rand, 2020).

Lastly, there are legislative and regulatory frameworks designed to suppress disinformation. This approach is embodied in Germany's 'NetzDG' law, where hefty fines are imposed on social media platforms that promptly remove 'illegal' content (Schulz & Pukropski, 2019). The strength of regulatory measures lies in their potential to effect systemic changes. However, they raise complex issues around freedom of speech, and their effectiveness can be undermined by difficulties in defining and adjudicating 'truth' in a diverse society. The existing measures to counteract the surge of misinformation online have proven to be insufficient, which can be attributed, to a degree, to the inherent limitations of solutions rooted in algorithms. This lack of sufficiency arises from a conundrum where the algorithms can unintentionally magnify the echo chamber effect, leading to increased polarization among users. This inadvertent consequence may lead to the suppression of free speech as algorithms work to filter out what they perceive as 'unwanted' content, thus censoring certain viewpoints, which might be otherwise integral to fostering diverse, well-rounded conversations. To better understand the complexities involved in tackling this issue, we must examine the interests of social media corporations, which reveal an intriguing dichotomy upon closer inspection. On one side of the spectrum, these corporations are committed to reducing the spread of misinformation. As they present it, such a commitment aligns with the broader objectives of maintaining the fabric of society intact and promoting informed public discourse. However,

at the same time, these companies are unquestionably motivated by their financial bottom lines. They often resort to strategies to maximise user engagement to fulfil their economic ambitions. This engagement is, at times, driven by the spread of sensationalized misinformation which, paradoxically, can promote discord rather than foster informed discussion (Del Vicario et al., 2016). This seemingly contradictory stance – with ethical commitments to society on one end and profit-making strategies on the other – poses a tension within the operational frameworks of these social media corporations. They find themselves attempting to balance the ethical responsibility of maintaining a truthful digital space against their commercial obligations to shareholders. Google's Jigsaw project presents a case in point for this dichotomy. Designed as a solution to tackle extremist content on the internet, it finds itself grappling with the dual nature of its goals.

On the one hand, it is a tool aimed at addressing social and geopolitical issues; on the other hand, it is an arm of a technology company that naturally benefits from power consolidation, creating an unequal digital landscape in the process (Powles, 2016). Despite these inherent challenges, the Jigsaw project has been able to carve out a unique space for itself. By actively participating in the discourse around misinformation and power distribution in digital spaces, it has successfully allied itself with academic institutions, media houses, and non-governmental organizations. While this brings them credibility and support, it simultaneously allows them to act as funder, partner, beneficiary, and agent of soft power, furthering their influence and reach (Powles, 2016).

Meanwhile, as a psychological phenomenon, the intricate and manifold nature of cognitive dissonance stemming from disinformation further complicates the discourse. Cognitive dissonance, characterised by the mental discomfort arising from holding contradictory beliefs, attitudes, or values, permeates almost all facets of human cognition and behaviour (Jordan, 2011). The prevalence and pervasiveness of disinformation potentially make it resistant to mitigation merely through technological interventions, thereby adding to the failures of technology saviourism often campaigned by large technology companies. As such, the question of cognitive dissonance and misinformation on social media

transcends the boundaries of any single discipline or organisation, requiring a concerted effort from multiple sectors of society. Moreover, the increasing prevalence of information operations in the contemporary geopolitical landscape has also engendered heightened vigilance among national governments, cognisant of the potential for foreign interference to exploit the innate cognitive vulnerabilities of their citizenry. In this context, human cognition has emerged as a critical element of national security, necessitating the development of robust countermeasures to protect against the deleterious effects of weaponised information (Burt & Geer, 2018). The omnipresence of social media platforms and their pervasive influence on public discourse have rendered societies susceptible to attacks predicated on the manipulation of collective consciousness, ultimately undermining the stability and integrity of the state. The massive explosion of behavioural data through shared content, interactions with others' content and the platforms made available by social media has made this crusade possible (Waltzman, 2017). This wealth of information affords intelligence agencies invaluable insights into target populations' behavioural patterns and cognitive biases, enabling the development of targeted countermeasures to mitigate the impact of adversarial information operations. However, little attention has been paid to anticipating and evaluating the potential intelligence and national security responses to the threats posed by adversaries on the internet.

The European Union (EU) has been a forerunner in championing legal frameworks for better user privacy and data protection in recent years, starting with the General Data Protection Regulation (GDPR) and further enforcing strict regulations through the Digital Services Act (DSA) and the Digital Markets Act (DMA) (Vergnolle, 2021). However, their current responses to the threat of disinformation could be boiled down to four approaches. These approaches, namely self-regulation, co-regulation, direct regulation, and audience-centred solutions, each carry unique characteristics, applications, and potential implications (Durach et al., 2020). Firstly, self-regulation refers to the actions undertaken voluntarily by digital platforms themselves. This approach is grounded in the concept of corporate social responsibility, where platforms

are encouraged to act responsibly in their operations without the need for external enforcement. Examples of self-regulation may include developing and enforcing internal policies against hate speech, misinformation, or other forms of harmful content. For instance, companies such as Facebook and Twitter employ this method to an extent, implementing their own guidelines to combat various forms of inappropriate conduct on their platforms. Secondly, co-regulation involves a cooperative framework established between multiple stakeholders. This can include European Union-level and national-level authorities, internet platform companies, media organizations, researchers, and other relevant parties. Under this model, digital platforms work in concert with these entities to develop and enforce regulations that uphold content and conduct standards. A prime example of co-regulation can be seen in the formation of the Global Internet Forum to Counter Terrorism (GIFCT), a cooperative initiative involving major tech companies and governmental bodies. Thirdly, direct regulation refers to governmental bodies imposing legal measures and sanctions. This could involve passing legislation that directly influences the operations of digital platforms, with consequences for non-compliance typically enforced via legal sanctions. Direct regulation has been manifested through laws such as the General Data Protection Regulation (GDPR) in the EU, which imposes strict data protection and privacy rules. Finally, audience-centred solutions represent an approach that emphasizes the end-users of digital platforms. This method involves enhancing the digital literacy of users and providing tools for fact-checking and verification. Such measures are intended to empower individuals to better discern the credibility and reliability of online content, fostering a more informed and discerning user base. An instance of this approach is the proliferation of fact-checking organizations and media literacy programs.

Mitigating the proliferation of disinformation necessitates a continuous commitment to the observation, examination, and critical appraisal of the strategies executed by entities such as internet platform corporations, global institutions, or sovereign nations. Dittrich (2019) explores the potential advantages and hazards associated with the European Union's (EU) co-

regulatory framework, juxtaposed against the model of self-regulation. The effectiveness of self-regulatory measures instated by platforms cannot be measured reliably, and there is no procedure to monitor the effects of algorithms deployed to curate and restrict content. This introduces an opaque layer to the effects of new features or changes introduced to existing ones. Decisions like Facebook's transition towards a video-centric newsfeed can have substantial implications for private individuals and business entities utilizing social media platforms. For instance, when Facebook modified its newsfeed algorithms to prioritize video content over text and de-emphasize political material, many news organizations witnessed a precipitous decline in online revenues virtually instantaneously. This minor technical adjustment, therefore, yielded a significant influence on the information consumption patterns of millions of European citizens (Madrigal and Meyer, 2018). Some forms of direct regulation like the German *Netzwerkdurchsetzungsgesetz (NetzDG)* or the updated Indian Intermediary Guidelines and Digital Media Ethics Code under the Information Technology (IT) act could prove dangerous to free speech without the right checks and balances where they could be abused to remove content on social media that is not favourable to the government. The audience-centred solutions, while instrumental, have extremely diminishing results and can only be implemented in conjunction with other approaches. Relying solely on the end-users to make educated mental choices is not viable, the same way one cannot be expected to determine if a product on a supermarket shelf is safe to consume. The shelves are regularly restocked with products that are safe to consume and not expired while also showing the expiry information to the shoppers so they can make a conscious choice. A co-regulation approach strikes a good balance while maintaining participation and dialogue between various levels of stakeholders.

However, such an approach would necessitate governments, platforms, and other infrastructure supporters to work in tandem towards a common goal of combating disinformation that proves challenging. The vague nature of information operations, which straddle the liminal space between overt acts of aggression and more insidious forms of subversion, necessitates a

reconceptualization of conventional approaches to national security. The dynamic nature of the digital landscape requires a fluid and adaptive approach to security that can anticipate and respond to the evolving tactics and strategies employed by adversaries. The blurring boundaries between state and non-state actors, between foreign and domestic threats, and even between reality and perception in the digital arena necessitate a more sophisticated and nuanced understanding of national security. A recent 2018 United States Special Operations Command (USSOCOM) analysis on grey zone conflicts asserts that the uncertain nature of information warfare necessitates embracing cognitive security to bolster national psychological resilience. In simple terms, cognitive security involves implementing a mix of attacking and defending strategies to combat the spread of false information, public opinion manipulation, and the decay of faith in democratic bodies (Rajtmajer and Susser, 2020). This new security perspective recognizes that the defence against these information operations should not be solely reactive but rather anticipatory, building robust psychological defences that can withstand the corrosive effects of disinformation campaigns. It calls for the development and implementation of innovative strategies and technologies, informed by an in-depth understanding of the psychosocial dynamics of the target population, to detect and neutralize potential threats in the information environment.

Addressing digital threats from adversaries requires a shift from the currently passive cognitive security approach to a more aggressive one. However, the ethics and policy considerations involved in this move towards an assertive stance are still not adequately discussed or defined. As the interventions devised by the security community become more intrusive and impactful, the ethical boundaries of these actions become increasingly blurred (Rajtmajer and Susser, 2020). The absence of a comprehensive ethical framework to guide these interventions results in a regulatory vacuum, wherein the permissibility of these actions is largely determined by their effectiveness rather than their conformity to any ethical standard. This poses a potent threat to civil liberties and personal privacy, warranting an urgent re-evaluation of our ethical stance towards cognitive security. Discussing these issues and deliberating on the ethical

framework is important as they pose different challenges and the inherent scale of impact from traditional military or national security activities. The challenges posed by cognitive security interventions are further exacerbated by their reliance on machine learning models. These models are designed to extract and analyse behavioural data to classify personality traits. While this capability has significant potential in predicting and mitigating threats, it poses significant ethical and privacy challenges. The opacity surrounding the functioning of these models, commonly referred to as the 'black box' problem, further compounds these challenges, as it impedes our ability to scrutinise and regulate their behaviour.

Using user data in policy making is an approach that is becoming more prevalent as our society becomes increasingly data driven. It involves the analysis of data about people's behaviours, preferences, and needs to inform the development and implementation of policy. Matz and Netzer (2017) point out that not only are academics now deeply engaged in this realm, but so too are businesses and governments, leveraging such data to inform a broad spectrum of decisions and policies. This broad swath of data, derived from social media interactions, search histories, e-commerce transactions, and beyond, provides a multi-layered, in-depth understanding of an individual's life. Ramon et al. (2021) elucidate that these digital footprints can offer highly intimate insights into users' lives, going beyond mere behaviour patterns encompassing several domains of human existence. The scope of this digital revelation is quite profound. Kosinski et al. (2012) emphasise that digital footprints can extend as far as determining an individual's personality traits, intelligence, and political orientation. As elucidated by Matz et al. (2017), psychological profiling seeks to leverage digital footprints for automated assessment of psychological traits. This approach pivots on proactive strategies in its implementation. For instance, Brdiczka et al. (2012) engaged in a proactive profiling method applied to the World of Warcraft (WoW), a widely popular massively multiplayer online game. They created a predictive model to determine the likelihood and timing of a player leaving their guild, which can detrimentally impact the group's efficacy and stability. Whilst the primary focus was on insider threat detection, the

possibility of adapting such models to discern threats arising from disinformation attacks offers a promising and relevant application. The potential for creating a proactive defence mechanism against disinformation attacks is notable. Such a deployed system could examine psychometric profiles to detect individuals or collective groups susceptible to cognitive dissonance, a psychological conflict arising from inconsistent beliefs and behaviours spurred by disinformation attacks. The potency of this psychological state can lead to impaired judgement, making it a prime target for disinformation campaigns. Thus, building a predictive model that proactively identifies such targets would be a significant breakthrough in combating disinformation. In this context, psychometric profiles serve as valuable tools for predictive modelling. These profiles can help identify cognitive biases, personality traits, behavioural patterns or sociodemographic factors that make an individual or group more likely to experience cognitive dissonance when exposed to disinformation. A predictive model incorporating these factors could potentially detect these vulnerable targets proactively, offering an opportunity to take preventative measures.

Bargar et al. (2019) suggest that this predictive modelling could aid authorities in formulating and executing targeted counter-information operations. A study by Tappin et al. (2022) explored political micro-targeting by comparing naïve approach and the single best message approach crafted to target the attitude of the receivers. The results show that microtargeting performs better than traditional broadcasting with a sizable persuasive advantage. Similarly, focused operations could potentially prove more effective in mitigating disinformation, as they address specific individuals or groups rather than adopting a blanket approach. Consequently, this surgical approach to counteract disinformation could enhance the efficiency and effectiveness of such operations, primarily by minimising the scattergun effect associated with broader strategies. Undeniably, significant complexities and ethical conundrums are associated with utilising these methodologies. Both data mining and artificial intelligence (AI), as fundamental pillars of these pursuits, invariably intersect with a constellation of critical issues – namely, the acquisition of informed consent, the safeguarding

of personal privacy, the preservation of data security, and the sustenance of fairness.

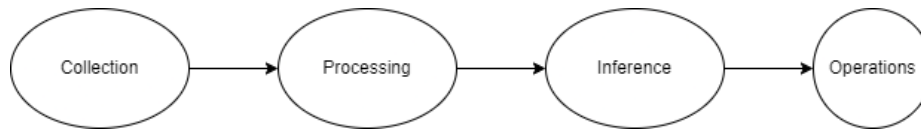
As we delve deeper into the digital footprints of individuals and groups to predict and counteract disinformation, we inevitably find ourselves grappling with whether and how to obtain consent for these data-driven interventions. This goes beyond the traditional understanding of consent as a one-time agreement and introduces a dynamic, ongoing process that must respect the autonomy and dignity of data subjects. Parallel to the concern of informed consent is the equally pivotal issue of personal privacy. The large-scale data collection and processing required for proactive cognitive security measures may infringe upon the privacy rights of individuals, leading to potential misuse or abuse of sensitive personal information. Therefore, a robust privacy-preserving framework is essential, ensuring that the data collected is used strictly for its intended purpose and is protected against unauthorized access or disclosure. Also linked to privacy is the issue of data security. As data repositories swell with an influx of personal information, they increasingly become attractive targets for potential cyber threats. A breach in data security could expose individuals to substantial material and psychological harm. Hence, there is a pressing need to bolster data security measures, integrate state-of-the-art encryption and access control mechanisms, and foster a culture of security within organizations. As we leverage AI algorithms for predictive modelling and decision-making, we must remain vigilant to the risk of inadvertent biases creeping into these systems. Such biases could result in unfair outcomes, further marginalizing vulnerable populations or perpetuating existing social inequalities. Thus, investing in techniques for detecting, mitigating, and eliminating bias in AI systems is imperative. These factors necessitate rigorous scrutiny, underlining the imperative for stringent regulations and ethical guidelines in applying these advanced technological tools. AI, in particular, presents complex challenges relating to explaining ability, fairness, and trustworthiness that are yet to be fully explored. As such, it is incumbent upon researchers and practitioners to maintain a careful balance between harnessing

the potential of these techniques for counteracting disinformation while ensuring ethical considerations and individuals' rights are not infringed upon.

## Makings of a new bulwark: A co-regulation approach

This thesis envisions a robust system wherein end-user data forms the core basis for algorithmic processing to identify groups of individuals who are particularly susceptible to disinformation. Nevertheless, the architecture of such a mechanism is far from straightforward, considering the complex and multifaceted nature of disinformation. Notably, disinformation does not exist in isolation. Instead, it interacts dynamically with various factors, including the source of the information, the medium of dissemination, and the cognitive and behavioural characteristics of the target population. In this process, data is akin to raw material; its potential is unlocked only when carefully processed and interpreted. To this end, end-user data must be systematically collected, accurately categorized, and rigorously analysed to yield actionable insights. However, extracting relevant and meaningful information from this data requires deploying advanced machine learning and artificial intelligence algorithms capable of handling high-dimensional, multimodal data. These algorithms must sift through large volumes of data to discern patterns and adapt to the evolving landscape of disinformation. Equally complex is how the target population receives and processes disinformation. The susceptibility of individuals to disinformation can be influenced by a myriad of factors, including cognitive biases, emotional states, societal influences, and pre-existing beliefs. Therefore, to develop a holistic understanding of vulnerability to disinformation, it is necessary to adopt an interdisciplinary approach that combines insights from cognitive psychology, sociology, politics, data science, and cybersecurity. Such a mechanism should also include the efforts of intelligence agencies that constantly track external influence to understand the objectives and thereby provide a deeper context for the prediction model to infer the vulnerable groups of people. The proactive nature of such a mechanism raises concerns that are already widespread in the circles of cybersecurity and geopolitics.

The design of the proposed framework is predicated on a 'Layered Model', a structured approach which compartmentalises the process into four distinct but interconnected stages: Data Collection, Data Processing, Data Inference, and Operations. This model mirrors a linear progression wherein the subsequent layers are heavily influenced by their preceding layers, but the influence does not extend in the reverse direction (refer to Fig.1).



*Figure (1): Flow of data across the layers*

This one-way directional influence offers a rigorous structure to the model, prohibiting any potential feedback loops that might distort the accuracy and objectivity of the process. The foundational layer is Data Collection. In this initial stage, end-user data is systematically gathered from various sources. The quality and depth of the collected data determine the effectiveness of the subsequent layers. This data could encompass various information, from individuals' social media activity and online behaviour patterns to more sensitive demographic data, including age, gender, and location. Since the nature of disinformation is variable across different sources and how the target population perceives it, a comprehensive and diverse dataset is vital. However, this data collection is performed adhering to stringent ethical guidelines, thereby ensuring the respect for privacy rights and informed consent of the users. Once the data is collected, it ascends to the next layer - Data Processing. The collected raw data is processed using advanced machine learning and artificial intelligence algorithms. These algorithms are designed to manage high-dimensional, multimodal data and extract meaningful patterns from the colossal information. This processing layer is influenced by the type and quality of data collected in the preceding layer. For instance, the effectiveness of a specific machine-learning model might vary depending on the characteristics of the input data. However, it is important to note that the Data Collection stage is independent of the needs or outcomes of the Data Processing layer. The purpose is to maintain the integrity of the data and prevent any potential biases that could

be introduced by tailoring data collection to fit pre-determined processing needs. The third layer, Data Inference, transforms processed data into actionable insights. Leveraging the patterns and trends identified during data processing, this layer infers potential disinformation vulnerability among the target population. It can predict individuals or collective groups susceptible to cognitive dissonance spurred by disinformation attacks. The effectiveness and accuracy of this layer depend heavily on the quality of data processing, which in turn is contingent on the data collected.

The final layer is Operations. Based on the insights derived from the data inference stage, specific countermeasures and proactive strategies are developed to combat the identified disinformation threats. These could range from tailored counter-information operations by intelligence agencies to policy recommendations and digital literacy programs for vulnerable communities. Again, the operations layer is strictly bound by the data inference, processing, and collection layers. Overall, the layered model offers a structured, step-by-step approach to disinformation threat identification and countermeasure development. By clearly defining one-way directional bounds between the layers, the framework ensures the process maintains its integrity, objectivity, and respect for user data privacy throughout. This linear progression eliminates the potential for abuse and the introduction of biases, thus providing a robust foundation for the fight against disinformation.

Since the internet is vastly influenced by private entities, most of them being for profit, makes it extremely hard for most democratic governments to rely on their infrastructure to feed into the mechanism solely. The authority, platforms possess over end-user data and data published on their webpages further challenges in sourcing the data directly. Though governments around the world have started to circumvent these roadblocks through regulation and other collaborative efforts when it comes to restricting and removal of content, a larger data collection effort such as the one required to bring the proactive mechanism to life requires closer working ties between the government agencies and platforms, advertising providers. Figure (2) elucidates various situational issues or concerns specific to the processes in each layer of the model, along

with the stakeholders involved in the data flow. The stakeholders include profit-seeking individuals or organizations, platform, or technology stack providers, the state, and civil or non-profit bodies. They are matched against the layers of the framework to highlight the pitfalls at their intersections. Importantly, this figure separates the technology stack from the stakeholders to underscore the inherent issues associated with employing the specific technologies used in this process. It should be noted, however, that while the technology stack is represented separately, it should ideally be under the jurisdiction of a state institution that is subject to external oversight from various stakeholders to ensure accountability and transparency.

	Individual	Platforms/ Profit-seeking	Technology stack	State	Civil bodies/ Non-profit
Collection	Privacy, Consent	Consent mechanism, Data procurement		Regulation	Advocacy
Processing		Data protection	AI bias, Anonymization, Overfitting,		
Inference			Abuse	Usage, Policy	Participati on
Operations		Transparency		Oversight	

*Figure (2): Layer-Stakeholder: Situating issues or concerns*

In the subsequent sections of this thesis, we delve deeply into each layer of the proposed framework, clarifying the roles and responsibilities of each stakeholder within their respective domains. This comprehensive analysis is not only essential for understanding the operational mechanics of the framework but also crucial in identifying the potential pitfalls that may arise in the execution of each layer's functions. The focus of this investigation spans both the processes and the technologies utilised at each stage. Particular attention will be paid to best practices that can mitigate these inherent pitfalls, thereby improving the effectiveness of the overall framework. These best practices encompass a range of areas, from ethical data collection and analysis to ensuring the secure and responsible use of AI and other advanced technologies. We will examine strategies to foster collaboration and coordination among the diverse

stakeholders involved while maintaining a steadfast commitment to user privacy and data security. The decisions taken to craft the guidelines in this work are inspired from existing discourses in different cases where these technologies are partly employed. Through a survey of existing implementations and analysis, these methodologies are translated to better fit this mechanism where psychometric profiling can be safely used to create a positive impact by addressing the risk of disinformation. At this juncture, it is important to highlight that the operations layer, which marks the culmination of the framework's processes, primarily interfaces with the end-users through various forms of media. This includes traditional outlets such as newspapers, television, and radio and more contemporary platforms such as digital news outlets, social media, and government-backed information broadcasting channels. Through these channels, the countermeasures developed by the authorities are disseminated, aiming to bolster cognitive security and safeguard the integrity of the public discourse. However, while the operations layer plays a crucial role in the framework, the scope of this thesis does not extend beyond this stage. In particular, the subsequent steps of monitoring and evaluating the impact of these countermeasures on the end users are not within the purview of this thesis. Instead, this thesis focuses on employing cognitive security for the development of efficient counter-information operations, by providing a robust foundation for further research and exploration in this novel and increasingly crucial field.

## Collection Layer

In the past few years, gathering data on the internet has become increasingly widespread among various platforms, encompassing social media networks, search engines, and online shopping websites. These platforms have adopted multiple methods to amass a vast array of user information. The data they gather encompasses a wide range of details, including personal data, demographic information, geographical coordinates, and individual preferences. This vast collection of user data serves multiple purposes for the platforms, including enhancing the quality of their services, refining their advertising strategies, and acquiring valuable insights into consumer behaviour patterns (Mäihäniemi,

2022). By utilizing this data, these platforms strive to tailor their offerings more effectively to their user base's unique needs and desires, resulting in a more personalized and tailored online experience. Part of that tailored experience is targeted advertisements. The election interference caused by Cambridge Analytica used the same strategy to reach out to specific users with tailored content to sway their voting (Kanakia et al., 2019). It started with a professor from Cambridge University who created a captivating Quiz App specifically designed for Facebook, with the primary goal of delving into the intricacies of users' psychological profiles. During that period, Facebook's terms of service explicitly acknowledged that the app developer possessed the authority to gather user data to conduct research. However, the terms did not explicitly clarify whether the developer was permitted to extend this data collection to include information about the users' friends. Upon signing up to participate in the study, users were presented with an engaging survey that demanded their attention and input. A convenient Facebook Login Button was seamlessly integrated into the survey to facilitate their engagement with the app. This button prompted users to log into the app using their Facebook credentials, granting the app access to their personal data.

Interestingly, as users innocently authorized the app to access their user data, they unknowingly and unintentionally authorized the app to collect data about their friends. This unforeseen consequence meant that not only was the app able to gather information such as the users' names, genders, locations, ethnicities, and educational backgrounds, but it also inadvertently acquired comparable data on their friends. This vast data collection extended beyond basic demographics, delving into intricate details such as the pages users liked and even the specific brands of clothing they favoured. The scale of this data-gathering endeavour is truly astonishing. Starting with a modest network of 250,000 nodes, the professor and their affiliated entity, Cambridge Analytica (CA), to whom the sold the data, managed to amass a staggering amount of data encompassing approximately 75 million nodes (Kanakia et al., 2019). This extensive dataset provided an unparalleled glimpse into millions of individuals' lives, preferences, and behaviours, empowering the professor and Cambridge Analytica with

unprecedented insights. The current data protection standards around the likes of GDPR, which was introduced after this fiasco, would not allow such a thing to happen. The inherent clause within the GDPR requires the platform to specify the end users of the usage and intended processing of their data while obtaining their consent to collect it (Goddard, 2017). User consent acquisition in digital platforms often hinges on accepting terms of service and privacy policies.

Nonetheless, the transparency of such consent procurement processes, particularly regarding data collection, has come under scrutiny among activists and policymakers. While these agreements detail the kind and use of data to be collected, many users may overlook or misinterpret these details, which brings into question the true comprehension of consent. Furthermore, third-party data collection and tracking, whereby user data is shared without explicit consent, exacerbate this concern. In order to improve transparency, platforms ought to ensure users understand the nature of the data collected and its application whilst offering clear opt-out options for those opposed to data sharing. Transparency should also extend to third-party data tracking, allowing users to decline such practices.

Our contemporary world relies on big data, making data protection and consent mechanisms paramount for safeguarding individual privacy and personal agency (Human et al., 2022). In recent years, global efforts have been directed towards this end, with the European Union establishing foundational standards through the EU General Data Protection Regulation (GDPR). GDPR, in theory, transfers data control from organizations to users, permitting individuals to regulate how their data is accumulated and applied. Consent is a crucial component of the GDPR and one of the few legal grounds for processing app user data (Utz, 2019). The GDPR specifically protects EU citizens and residents, while non-EU individuals remain outside its purview under which, organizations must detail data processing procedures and request consent before data handling. Individuals retain the right to withdraw consent, after which any data collected can no longer be processed.

However, data amassed under prior consent can still be processed (Arfelt, 2019). Some organisations have started utilising 'dark patterns' to acquire consent for

data collection. For instance, consent increased by 22-23% when the opt-out button was not presented on the initial page of a study by Nouwens et al. (2020). Some websites resort to persistent and intricate user tracking, even tracking those who have rejected cookies (Papadogiannakis et al., 2021). Despite the GDPR's proclaimed commitment to transparency, the reality of the 'Notice and Consent' model does not fully endorse the user-centric principle of transparency (Spagnuolo et al., 2019). The 'Notice and Consent' framework, widely accepted for data collection on the internet, informs users about the types of data collected and the intended uses, thereby acquiring user consent. This typically manifests as a privacy policy outlining the website's data practices and a consent mechanism such as a checkbox for user agreement. This process empowers users with control over their personal data and promotes transparency and accountability. However, relying on 'Notice and Consent' as the epitome of transparency without further scrutiny is potentially harmful and futile (Bietti, 2020). The large quantities of data collected can pose risks not immediately discernible when users consent.

The 'Notice and Choice' model aims to ensure that individuals can provide informed and free consent, thereby achieving a balance between privacy and the benefits of information processing. However, this approach has its detractors, as users often lack the information necessary to make informed decisions and usually do not peruse the provided notices. Despite these criticisms, policymakers and privacy advocates continue to emphasise adherence to 'Notice and Choice' (Sloan and Warner, 2014). Thus, ethical data collection should extend beyond terms of service (Fiesler et al., 2020) to a more tangible concept for users at the point of interaction. In his book 'Privacy and Freedom' (1967), Alan Westin defined informational privacy as the individual's ability to control when and how their information is collected and used. However, true informational privacy requires a broader ability to grant or withhold consent. The 'Notice and Choice' model, used widely for online data collection, attempts to provide this control by presenting terms and conditions and a means for user acceptance.

Nonetheless, this model faces substantial hurdles, as users typically lack an understanding of online business data practices, making informed decisions difficult (Sloan and Warner, 2014). 'Notice and Consent' attempts to remedy this by presenting information through a standard contract such as a privacy policy. However, most individuals do not read these notices and cannot make informed decisions.

In the legal realm, hypothetical knowledge is often invoked. As long as users have adequate opportunity to understand a notice, they are assumed to be aware of its contents, even without reading it. This approach applies to 'Notice and Consent', yet it rests on the assumption that individuals will indeed take the opportunity to read and understand the notices, an assumption that may be unrealistic. Companies often claim that users consent to comprehensive data collection by agreeing to their Terms of Service (TOS). Recent incidents have, however, demonstrated that consumers may not meaningfully consent to the TOS. This raises significant privacy concerns, as the contractual concept of consent seems insufficient to protect citizens' privacy rights (Kim and Telman, 2015). The complexity of presenting data flow across the internet at the point of consent collection due to the interconnected nature of user tracking systems further complicates the situation. Thus, while courts may treat a simple acceptance click as informed consent, users may not fully comprehend the rights they relinquish in exchange for using a specific website or service. Notable instances like Facebook's manipulation of users' news feeds to study user behaviour and mood changes have raised eyebrows. Facebook defended its position by stating that users had given consent through agreement with their TOS (Goel, 2014).

Similarly, Google maintained that Gmail users, by accepting their online TOS and Privacy Policy, consented to data usage for various objectives. However, many users remain unaware that actions as seemingly innocent as clicking an icon or hyperlink can result in legally binding agreements (Sloan and Warner, 2014). This was humorously illustrated by video game retailer Gamestation, which included a "soul" claim clause in their contract, which users could void

by selecting an option; many did not, indicating that users do not thoroughly review such agreements (Fox News, 2016).

Yet, it is not entirely fair to lay the blame at the user's doorstep. The digital documents that serve as contracts can be extraordinarily lengthy, making it challenging for the average user to read and comprehend them. The iTunes TOS, for example, spans a staggering 32 pages, a considerable commitment for what might be a minor transaction. Adding to the confusion, these terms are frequently updated (Taylor, 2003). A key issue lies in the current digital consent model's simplicity. On the surface, a user clicking to agree to a set of terms seems to represent informed consent. However, in reality, the user is often not fully cognisant of the rights they forfeit in exchange for a website or service's usage. Dark patterns employed by websites to subtly encourage users to agree to data collection further complicate matters. The representation of data flow across the internet at the point of acquiring consent is also a complex issue. This complexity stems from the intricately interconnected system that monitors user activity and delivers targeted advertisements. In this ecosystem, profilers categorise consumers based on their likelihood to purchase, drawing upon data such as age, gender, ethnicity, marital status, profession, and internet activity (Steel, 2010). This data is then used to optimise text and display advertising. Advertising exchanges, such as Google's AdSense, distribute advertisements to hosting websites. Upon a visitor's arrival, an advertising exchange matches the visitor's profile with their current website activity to deliver highly targeted advertisements (Davis, 2006). Businesses then compete through an auction process to present their tailored advertisements, a process that takes mere milliseconds.

Given this lightning-fast and complex process, the advertising ecosystem cannot present a comprehensive snapshot of data flow and its utilisation at the point of interaction where consent is collected. Therefore, we must recognise the limitations of current consent models and explore more robust, user-friendly, and transparent solutions. Not only do we need to ensure that users understand the implications of their consent, but we must also advocate for their right to control how their data is used. Addressing this issue requires a concerted effort

from policymakers, businesses, and digital platform providers to champion the principles of transparency, user autonomy, and privacy. The lack of user comprehension regarding consent acquisition demands greater transparency and ease of use. It suggests the need for a user-centric, effective, and easily navigable system prioritising transparency and user autonomy. This system should be designed to inform the user precisely what data is being collected, how it will be used, and by whom. It should also elucidate the consequences of data sharing and the user's rights in controlling their data without requiring them to comb through lengthy and complex TOS.

Moreover, the process for revoking consent should be as streamlined and clear as the process for providing it. Current practices often make revoking consent an onerous process, contributing to user frustration and disillusionment. Improved transparency in this area would enhance users' trust and engagement. Furthermore, consent should be a continuous process rather than a one-off event. Users should be periodically reminded of their consent status and given opportunities to adjust their preferences. This approach would respect users' evolving preferences and their right to alter their decisions over time. Ultimately, consent for data collection should not be an 'all or nothing' proposition. Users should be able to provide granular consent, deciding which types of data they are comfortable sharing and which they are not. To facilitate this, businesses and platforms must develop more refined data collection practices that respect user choices and privacy.

Following is an eight-point guideline envisioning a new interface for obtaining consent for data collection to increase transparency, foster trust and keep the web experience as seamless as possible. This guideline is created with best practices in mind by current standards and should be continuously updated as we move forward.

### 1-Intuitive Onboarding

Upon their initial visit to the website, visitors would be promptly met with a concise summary detailing the fundamental characteristics of the platform. This compact introduction not only outlines the platform's core purpose and mission but also provides an early overview of the platform's commitment to user

privacy. This initial stage plays a pivotal role in establishing the overall atmosphere and tone of the user's subsequent experience and exploration of the website. It marks the beginning of a journey, setting the stage for the visitor's interaction with the site. At this point, the website does not yet seek any permissions from the users. Instead, the primary focus is constructing a preliminary foundation of trust and mutual understanding between the platform and its users. The purpose of this approach is to subtly convey the platform's dedication to ensuring its users' privacy, a commitment that underscores its overall mission. By providing this information up front, the site seeks to foster an environment where users feel secure and informed. At the same time, users are not put in a situation where they agree to the terms for a quick entry without deeper inspection. The detailed and careful introduction works to anticipate and address the potential concerns of visitors even before they engage with the platform more significantly. The objective is not to immediately ask for access permissions but rather to cultivate a sense of confidence and trust. At this early stage, the user would be allowed to familiarize themselves with the platform's values, ethos, and approaches to user privacy. This way, as they navigate deeper into the site, they are equipped with a solid understanding of the platform, making their user experience more comfortable and well-informed.

In summary, this crucial stage would set a positive tone for the users' journey through the website. The platform provides a user-friendly, secure, and understanding environment by not asking for any permissions at this point but instead building an initial layer of trust and comprehension. This method demonstrates significant respect for user privacy, contributing to an atmosphere of trust, understanding, and optimal user experience.

## 2-Contextualised Consent Requests

When individuals engage with different components or features of the platform, each demanding unique data, they come across clearly expressed, non-disruptive requests for their consent. These requests should be designed to explain the need for the specific data being requested, its planned application, and the benefits the user can expect from its utilization. The primary goal of this approach is to offer a comprehensive understanding of how users' data will be

handled, thus making the consent process more transparent and user-friendly. Consider a scenario involving an algorithmic content recommendation feature to give a concrete example. In such a case, users might be asked: "For us to provide personalized recommendations, we need access to your browsing behaviour while you are using our platform. Would you be comfortable with us collecting this information? This procedure will significantly assist us in creating a more customized and enhanced content experience for you." This specific dialogue is formulated to explain the data collection process to the user and seek their approval. The user's browsing history on the platform is instrumental in making content suggestions that align with their preferences and interests, leading to a more tailored and satisfying user experience. This information request is not made as a vague requirement but is rather tied directly to the benefits the user stands to gain from allowing access to their browsing data. The platform strives to cultivate an environment of trust, respect, and mutual benefit by ensuring clarity and transparency in data usage. Similarly, certain characteristic data like the users' likes and search queries could be accompanied by a message stating the efforts to inform and protect them from the impact of disinformation. While the complex mechanism behind these processes need not be presented, some information, like flow and access to that data, keep the process transparent to the end user.

The overall aim is to balance the need for data to improve platform functionality with respect for user privacy and control over personal information. By making these consent requests lucid and specific, the platform helps users understand the value of their data and the benefits they can accrue from sharing it. This approach fosters a sense of collaboration and informed consent rather than imposition and ambiguity, thus leading to a more engaging and enriching user experience with clear value exchange.

### 3-User-Centric Consent Management

To ensure that all users are aware and in control of their participation, any solicitation for consent should be accompanied by an easily comprehensible and straightforward method for users to indicate their agreement or disagreement. This process could be realized through the use of a binary system such as a

toggle switch or checkbox, which could explicitly demonstrate the user's willingness or lack thereof to participate in the data collection process. To elaborate further, this binary system, whether it takes the form of a toggle or a checkbox, is a representation of the user's consent. It is intended to mirror the user's explicit approval or denial regarding the data collection process. In simple terms, by selecting or unselecting the box or switch, the user effectively communicates their decision about the data collection request. While a toggle This selection represents their acquiescence or rejection of the request. The platform must respect and adhere to the choice made by the user. This is a fundamental principle in maintaining user trust and fostering a respectful interaction between the user and the platform. While this is now a common practice in certain parts of the world, having many toggles at the entry point of the webpage fails to convey the exact purpose. Instead, localised pop-ups that do not obstruct the user can present more granular requests that are easy to follow (more in Informative Interjections). In no circumstances should the platform take measures to degrade the user's experience due to their decision regarding data collection.

The user's decision, whether it is to grant or refuse consent, must not be used as a basis to degrade or depreciate the quality of service they receive from the platform. To put it another way, the user's choice to grant or deny consent should not influence their ability to fully utilize the platform's services, and the user's experience should be maintained at a high quality regardless of their decision. Therefore, it is the responsibility of the platform to uphold this principle of respect towards user consent. The user's decision should be honoured, and no negative consequences should be imposed on the user as a result of their choice. In this way, we ensure that users maintain control over their data, fostering a more transparent and respectful interaction between platforms and their users.

#### 4-Dynamic Consent Interface

The envisioned platform should ideally include a feature known as a 'Privacy Dashboard'. This is a dedicated area within the platform interface that is always accessible to users, designed specifically for reviewing and understanding the nature of the data that has been gathered about them. This interactive hub would

not only enable the users to examine the specifics of their personal data that has been collected but also provides detailed insights regarding why this particular information was necessary to collect and how it is being utilized within the system and by their partners. This transparency is pivotal as it is meant to shed light on the complex processes of data collection and usage, an aspect that often remains opaque to most users. It offers them a clear overview of what kind of personal information the platform holds and for what reasons. It would delineate the specific data points collected, the intent behind the collection of each piece of data, and the methods of applying this data within the platform. These detailed insights help to elucidate the often complex logic behind data collection and use, thus demystifying these processes for the users in a separate and accessible area that does not interrupt the user experience.

Furthermore, the Privacy Dashboard would also serve as a dynamic control centre where users can adjust their consent preferences at any given moment. This feature reinforces users' autonomy over their data, empowering them to make informed decisions about their privacy based on the information provided by the dashboard. The ability to modify consent choices in real-time is an essential feature, providing users with the flexibility and control necessary to manage their data in a manner that aligns with their comfort levels and privacy concerns.

#### 5-Incremental Data Procurement

Instead of attempting to procure all data from users concurrently, it is recommended that digital platforms pursue a more staggered methodology, distributing the process of data acquisition throughout the course of the user's interaction with the system. This alternative approach implies that data would be collected from users in a gradual, segmented way as they navigate and interact with the platform, as opposed to overwhelming them with a single, simultaneous request for all their data. The strategy of phased data collection serves multiple important functions. One primary advantage is that it can prevent users from feeling overwhelmed or 'inundated' by numerous data requests all at once. For many individuals, the experience of being asked to provide a vast amount of information in a single instance can be quite

disconcerting, possibly leading to a feeling of unease or anxiety. The potential for such feelings may increase if the user cannot immediately understand why each piece of data is needed or if the data requests seem excessive or intrusive in their context. By spreading data collection points throughout the user journey, users are not only given a chance to become more familiar with the platform before being asked to share more personal or sensitive information but the requests themselves can also be more specifically tailored to the context of the user's current interaction with the platform. This means that each data request can be made at the most relevant point in the user's journey and in the context most likely to make clear the need for that specific piece of data. This approach enhances the relevance and comprehensibility of each data request. It allows for each request to be fully contextualized, both in terms of what is being asked and why it is being asked. This strategy ensures that the user can comprehend why certain information is required from them at each point, thereby increasing the likelihood that they will feel comfortable providing it.

#### 6-Assisted Consent Flow

In the design of systems that require critical operations, such as this application of cognitive security to combat disinformation, it is crucial to obtain the clear and explicit consent of the user. These operations should be orchestrated within the context of guided workflows, also referred to as assisted flows. By utilizing this architectural framework, it becomes possible to provide the user with a detailed, step-by-step elucidation of the process, emphasizing the sequence of events that will take place preferably in the consent interface as described above (refer to Dynamic Consent Interface). Each step in the guided workflow should be structured in a clear and sequential order. This is particularly useful in complex procedures, where the understanding of each task and the necessary permissions required for their execution can be daunting or confusing for the user. By breaking down these processes into individual steps, it becomes easier for users to comprehend what they are consenting to, thus making informed decisions. In other words, assisted flows play an important role in explicitly outlining the permissions required at each stage of the process. This detailed outline would allow the user to fully understand the implications of their

consent. It guides them through the process, highlighting the permissions that will be required at each juncture. Thus, the user is given an insight into what they are authorizing, which in turn provides them with greater control and understanding over the process.

This practice is particularly vital for critical operations, where there could be significant consequences if a user unintentionally approves an action without fully understanding its implications. These complexities are broken down into smaller, more manageable parts with guided workflows. Therefore, users are better equipped to make informed decisions about their involvement and the permissions they grant. In summary, architecting critical operations that demand explicit user consent in the form of assisted flows is a best practice. It creates a more transparent and comprehensible environment for the user, ensuring that they have all the necessary information at their disposal to understand each step and the corresponding permissions required in the sequence of events.

#### 7-Informative Interjections

As a user traverses the various facets of the platform, their journey should be intermittently dotted with instructional and informative pop-ups or sidebars. These elements, embedded strategically within the user interface, act as additional layers of communication aimed at enhancing the user's comprehension and awareness of their privacy rights and how their personal data is managed and utilized. In essence, these informative interjections, placed strategically throughout the user's journey, would help illuminate the intricacies of data privacy and usage. It is worth noting that platforms are the only place the end user is exposed within this process involving cognitive security, so these interjections offer insights into the practices employed by the platform concerning the user's data, making it easier for the users to grasp the implications and exercise their rights accordingly. Their purpose is not just to inform but also to empower the users, giving them the knowledge and understanding required to navigate the platform with informed consent. However, despite being crucial for imparting vital knowledge, these notifications need to be subtly blended into the user's browsing journey. They need to be easily discernible and identifiable, but their presence should not

intrude upon or disrupt the seamless experience of navigating the platform. The user's browsing journey should not be marred by a bombardment of endless notifications; rather, it should be scattered in a way that augments their understanding without imposing on the fluidity of their platform interaction.

It is a delicate balancing act, integrating vital informational content into the user interface while maintaining the browsing experience's uninterrupted flow. The goal is to create a harmonious symbiosis where the platform not only serves its primary function but also stands as an educative tool that guides the users about their data privacy rights and usage policies, all the while ensuring that their journey is not hampered in any way.

#### 8-Translucent Linguistic Practices

The practice of employing dense and complex legal jargon ought to be consciously avoided in favour of the application of language that is transparent, precise, and readily comprehensible to its intended users. This shift in communication strategy is essential, particularly in the fields where the detailed operations of data gathering, and application are involved. The aim is to ensure that these complex processes are not obscured or made unnecessarily convoluted by the individuals who would interact with them. By adopting an approach that advocates clarity, the content is made readily comprehensible even to those without specialized knowledge. By doing so, it is possible to avoid reducing the information into a collection of opaque and potentially misleading phrases that create a disconnect between the parties involved - the data collectors and the data providers or end-users. By embracing a more straightforward and intelligible form of language, the users would be better equipped to understand the implications and the consequences of their decisions in terms of the use and sharing of their data. More importantly, the clarity offered by such a language enables the individuals to fully understand the implications of their actions, thus allowing them to make well-informed decisions, particularly in terms of their consent.

Consequently, this proposed approach to communication, which emphasizes accessibility and comprehensibility, is instrumental in empowering users to give truly informed consent. This is of utmost importance as informed consent is not

merely about ticking off a box. However, it is a process that requires full comprehension and awareness of the implications and potential outcomes. Therefore, clear, concise, and user-friendly language should be the standard in such matters, ensuring that individuals are completely aware of their choices and the potential consequences thereof.

To ensure compliance, there may also be a need for stronger regulation and enforcement of transparency and consent practices. Regulatory bodies should continue to adapt and improve regulations to keep pace with rapidly changing technology and data collection practices. They should enforce penalties for companies that employ deceptive practices, fail to properly inform users about data collection, or disregard users' consent preferences. Further, civil bodies and humanitarian organisations should be included in the dialogue of regulation to endure participation and collective interest from all the different sections of society. Finally, an interdisciplinary approach would be helpful in this overhaul of how individuals interact with webpages and further developing these guidelines to better fit the final product. Incorporating insights from fields such as behavioural economics, social psychology, and user experience design can help develop consent practices that are user-friendly, transparent, and respectful of privacy.

## Processing Layer

By drawing from the voluminous pool of collected data, Cambridge Analytica meticulously crafted unique user profiles constructed for each individual and classified with precise care according to a range of demographic and psychographic indicators. This utilization of data to curate personalized user profiles was an extensive process that depended heavily on the geographic locale of the user within the United States. This factor in turn contributed significantly towards the formation of the user's political perspective and corresponding voter profile (Kankia et al., 2019). For instance, residents near U.S. border regions demonstrated concerns primarily rooted in immigration matters. These individuals were thus assimilated into a category denoted as anti-

immigration voter profiles, reflecting their principal apprehensions. The interpretation of the data in this manner facilitated the generation of tailored profiles that accurately mirrored the political ideologies of the individuals concerned. In stark contrast, individuals living within the nation's hinterlands demonstrated different primary concerns. Issues of declining manufacturing jobs and the controversial construction of oil and gas pipelines through Native American settlements seemed to be at the forefront of their preoccupations (Kankia et al., 2019). Such concerns were thus utilized as central determining factors during the creation of their distinct user profiles.

Furthermore, ultrahigh net-worth individuals residing in upmarket suburban and downtown localities expressed particular interest in tax breaks. Thus, these financial concerns became an essential component in forming their profiles, emphasizing the user-centric approach to profile creation. The brand of clothing users preferred was another unconventional but insightful data point used to gain a deeper understanding of their potential political leanings. As an illustrative example, denim brands like Wrangler and L.L. Bean have traditionally been linked with a more conservative demographic. In contrast, Kenzo, another denim label, tends to resonate with a more liberal audience (Halliday, 2022). This nuanced correlation allowed for the further refinement of user profiles, specifically categorizing them into broader ideological divisions, such as conservative or liberal, as per their clothing preferences.

To advance this process of user categorization, an array of classification methods was utilized, particularly focusing on the psychological metric known as the OCEAN score. This acronym stands for five key personality dimensions: Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism. Each individual user's score along these dimensions was used to add another layer of personalization to their profile (Grassegger and Krogerus, 2017). To harness the potential of this vast dataset, a Machine Learning model was trained using regression techniques. This allowed for the effective prediction of a new user's political affiliation based on the model's knowledge acquired from existing datasets. This algorithm not only took into account the user's OCEAN score and their psychographic data, but it also considered their activities on

social media platforms like Facebook, such as likes, posts, and shares (Kankia et al., 2019). Moreover, the model was designed to take into account the user's specific policy concerns and their desires for governmental change. The integration of such diversified data sets enabled the model to deliver more nuanced and accurate predictions of political affiliations, ultimately augmenting the understanding of the complex interplay between various demographic, geographic, and psychographic factors in shaping political affiliations.

In a bid to elucidate the intricate connection between individual online behaviour and inherent psychological constructs, a series of intricate and meticulously designed experiments were carried out by Omelianenko (2017). The primary objective of these experiments was to explore the existence, if any, of a consequential correlation between an individual's psychometric scores, as delineated by the comprehensive O.C.E.A.N. model, and the collection of Facebook 'likes' directly associated with said individual. This comprehensive investigation incorporated the application of both rudimentary and advanced machine learning algorithms. Yet, irrespective of the complexity of the algorithmic model employed, the prediction accuracy for almost all O.C.E.A.N. personality traits were, disappointingly, dismal (Omelianenko, 2017). This led to the inferential conclusion that, in its current technological and analytical capacity, leveraging machine learning models to accurately estimate an individual's psychometric profile solely grounded on their Facebook 'likes' appears, regrettably, to be an unfeasible endeavour, as noted by the researchers. The weak correlation between these two disparate variables, psychometric scores and Facebook 'likes', could be potentially attributed to the inherently nuanced and multi-faceted nature of personality traits. Such complex traits necessitate a more robust and intricate model of individual behavioural inclinations and preferences than can be readily inferred from the relatively superficial insight provided by a person's Facebook 'likes'. This finding, therefore, highlights the need for continued research in this captivating domain, with a specific emphasis on complementing Facebook 'likes' with additional data points that could potentially enrich the prediction model and thereby substantially enhance prediction accuracy. Shifting the focus to demographic

profile estimation through the lens of machine learning models, an interesting paradigm emerges. Unlike the psychometric data, a robust correlation was unearthed between key demographic traits, namely Age, Gender, and Political Views, and Facebook activity as quantified by 'likes' (Omelianenko, 2017).

The findings from the rigorous experiments demonstrated the plausible feasibility of deploying advanced machine learning methods to construct an accurate demographic profile of an individual solely grounded on the repository of collected Facebook 'likes'. Similar attribution was found between social media use and depression under the This inference underscores that demographic characteristics, inherently more observable and considerably less complex than psychological traits, hold the potential to be accurately predicted through discernible patterns embedded within Facebook 'likes'. Another effort at predicting user personality in terms of the O.C.E.A.N model was made through the text and linguistics of user activity on the internet, where the model classifies social media text to the pre-determined personality facet from the Big Five personality traits, maps the knowledge to the ontology, and uses machine learning to predict personality traits. The ontology-based model for measuring human personality is still rare, and the proposed model enriches the current methodology to measure human personality by observing writing and linguistics (Alamsyah et al., 2020). Concerning the selection of the appropriate machine learning model, a compelling case emerged for the Shallow Neural Network architecture. Among all machine learning models tested by researchers, this specific architecture outshone the rest by delivering the highest overall prediction accuracy (Omelianenko, 2017). Here is an analogy to help understand the workings behind the opaque processes. Let us say we must get a ball through a hoop. A simple tool, like a ramp, might be the most effective way to get the job done. There is only one element in the working of this mechanism. In the same way, the Shallow Neural Network, which is a simple model with only one layer for making predictions, would be the most effective for the task. When trying to add more layers to the shallow architecture model (like making a multi-stage ramp or conveyor belt system) to allow for more complex neural architecture that might be better at predictions, it resulted in poorer results

(Omelianenko, 2017), like making a ball get through a hoop more complicated than it needs to be. This is likely due to overfitting (when the model is too complex and performs well on training data but poorly on new, unseen data), which means that the more complex neural architecture was able to comprehend the variable parameters to an extent where there is little to no room for fuzzy logic and therefore works very well on training data but creates an inherent bias within the model to categorise new data to resemble the training data. On the other hand, the end user data has many proxies that contribute to the result with higher dimensionality than what is possible in a shallow neural network. To accommodate this, Singular Value Decomposition or SVD can be used to reduce the data to its most essential parts (Loan, 1976), similar to simplifying a big complex list of movies someone has watched down to key information like their preferred genres, actors, or directors. The background process works in a way where one has a really big, complicated jigsaw puzzle. Trying to understand the entire picture all at once can be overwhelming. So, they break it down into smaller parts, or groups of pieces, to make it easier to understand. That is essentially what SVD does with data. It takes a complex dataset and breaks it down into a set of simpler, smaller pieces. These pieces (technically called matrices) can then be used to recreate the original data, but they also allow us to see patterns and structures in the data that might not be obvious otherwise. For example, in the context of movie recommendations (like we discussed earlier), an SVD might be used to reduce a large dataset of user ratings for different movies into a smaller set of 'taste profiles'. Based on their previous ratings, these profiles could then be used to predict what other movies a user might like.

To use Singular Value Decomposition (SVD) in conjunction with the OCEAN model, the raw data (such as likes and activity) would first need to be mapped onto variables that relate to each of the five personality traits. For instance, a user's 'likes' on social media could give us a glimpse of their openness to experience - liking a variety of different topics might indicate high openness, while liking only a few closely related topics might indicate low openness. Activity level and the types of interactions could be used to infer traits like

extraversion and agreeableness. This process would involve domain knowledge in psychology and potentially also machine learning to identify patterns. Once the raw data is mapped to the OCEAN variables, we would have a large matrix where each row represents a user, and each column represents one of the OCEAN traits. However, if we have many users and detailed breakdowns of the OCEAN traits, this matrix could still be quite large and complex. This is where SVD can be used to simplify this large, complex matrix into a set of smaller parts (matrices) that are easier to understand. The result of this is that each user could be represented by a small set of values that capture the most important information about their personality, according to the OCEAN model. One important thing to consider for cognitive security is to do data analysis by anonymizing the data and focusing on aggregate trends rather than individual behaviours to respect user privacy. Rather than analysing each row (each user) separately, users could be grouped based on specific traits. For instance, all users who have a high score for extraversion can be grouped together or based on who shows a particular pattern of likes and activity. By focusing on these groups rather than individual users, we can maintain privacy while reducing the complexity of your data. This approach allows us to conclude patterns and trends within these groups while preventing the identification of individual users. For example, you could find that the group of users with high extraversion scores tend to like certain posts or engage in certain activities without revealing anything about what any specific user likes or does. However, it is important to note that anonymization should be done carefully to avoid the risk of 're-identification'. This is where seemingly anonymous data can be combined with other information to identify individuals (Hay et al., 2008). A study by Kosinski et al. (2016) suggests that significant improvements in prediction accuracy could be realized through the construction of individualized advanced machine-learning models for each dependent variable, such as individual OCEAN parameters. An area to explore would be an advanced model for each parameter, taking into account the likes, activity, and text inputs from the ontological approach (Alamsyah et al., 2020). However, further research is necessary to make any significant claims regarding this process.

At the outset, raw social media data, primarily text-based, is transformed into a format interpretable by our algorithmic model through data preprocessing. This process employs Natural Language Processing (NLP) techniques, starting with tokenization, where strings of text are broken down into meaningful units or tokens (Webster and Kit, 1992). Simultaneously, it is crucial to eliminate stop words, which are frequently used words with minimal informational value, such as prepositions and articles. This step streamlines the data set, enhancing processing efficiency and focusing on words that offer richer semantic information. The next step involves converting these tokens into numerical vectors, a process known as embedding. This process uses techniques like Word2Vec, GloVe, and transformer models like BERT, which are pivotal in transforming raw text data into a usable format for analysis. In traditional language models, text is processed either from left to right or right to left. These models have a high chance of missing out on important context as they do not consider the full scope of the sentence. BERT, which stands for Bidirectional Encoder Representations from Transformers, analyses text in both directions at once, capturing a deeper understanding of the context. It does so by looking at each word within the context of all the words that come before and after it in a sentence (Shen and Liu, 2021). Let us take the example of political discussions on social media. If the model were to analyse a post that says, "I do not support this policy", a left-to-right model might misinterpret the meaning of the word "support" because it has not yet seen the word "do not", which significantly changes the sentiment of the statement. BERT, on the other hand, considers the entire context and can understand the real sentiment behind the statement. Moreover, BERT can recognize that the word "support" has different implications based on the words around it. This context awareness enables a more precise representation of the text, which is essential when understanding and analysing polarized viewpoints. This procedure effectively turns qualitative language data into structured, numerical data that machine learning models can understand. These numerical representations of the words maintain the context and semantic richness of the original text, enabling us to perform a meaningful analysis of the text data.

The pre-processed data is then sorted into distinct categories using unsupervised learning techniques such as clustering. This segmentation could be based on a range of shared characteristics, from geographical location to clothing brand preferences. However, care should be taken to avoid oversimplification or incorrect categorization, which could lead to inaccurate representation of the user base. The segmented data undergoes a deep exploration to identify trends, patterns, and correlations using descriptive and inferential statistics, correlation analysis, and advanced data analysis techniques. The insights obtained are used to create nuanced user profiles encapsulating their political affiliation, policy concerns, and psychographic traits indicated by their OCEAN scores. This methodological approach also includes stemming or lemmatization, reducing words to their root form, thereby homogenizing different morphological variations of a single word (Khyani et al., 2021). This stage allows the grouping of related words, facilitating a more streamlined and accurate analysis. These procedures, though intricate, help to unlock the potential of user data, ensuring the highest quality inputs for subsequent profiling and model training. Through sentiment analysis and topic modelling, we can uncover emotional tones within words and hidden topics within the text data (de Arruda et al., 2015). These insights help to illuminate the user's personality traits, such as their level of Openness, and inform the development of individualized OCEAN profiles.

Once the data is properly formatted, segmented, and analysed, the machine learning (ML) model training begins. This is an iterative process where the model learns the underlying relationships between the input features (user characteristics and behaviours) and the associated output labels (like political affiliation). Techniques like regression are often used for training, which quantifies the relationships between variables. The ML model should undergo continuous evaluation and adjustment until it can make accurate predictions on unseen data. The trained ML model would make predictions on new users' potential political affiliations based on their attributes. As the context and data change, the model continually refines its predictions, ensuring adaptability to new data and changing environments. Moreover, unsupervised learning techniques, including k-means clustering and Gaussian Mixture Models, are

used to group users based on their extracted feature similarities, creating categories of personality profiles (Farrukh et al., 2022). This process leads to trait Assignment, where the dominant OCEAN traits that define each cluster's character are discerned and tagged. A significant part of the process involves calculating a vulnerability score for each user and assessing their susceptibility to disinformation. This procedure relies heavily on the OCEAN personality profiles curated for each user, utilizing them as key indicators of susceptibility. Certain personality traits can be assigned a specific 'vulnerability weight', quantifying their influence on the individual's susceptibility to disinformation. The comprehensive vulnerability score for each user is calculated by adding the products of each trait's score and its corresponding vulnerability weight. This aggregate score provides a quantifiable measure of the user's potential susceptibility to disinformation, bridging the gap between theoretical constructs of personality traits and practical tools for gauging vulnerability. This detailed approach allows us to build a coherent and comprehensive profile of end-users, helping to predict their behaviours and susceptibility to disinformation. This information should always be displayed in the aggregated format without access to granular data unless required for deep inspection of the model.

This process, however, is not immune to the many pitfalls discussed in academia and industry. AI and other computational algorithms suffer from an array of problems, some of which are highlighted below.

**Understanding of Context:** AI models, even with advanced natural language processing capabilities, still lack a complete understanding of human context. They might struggle to understand sarcasm, humour, or cultural nuances in the text, which could lead to false positives or false negatives when using the extracted data to compute risk scores.

**Dealing with Ambiguity:** Ambiguity is a common characteristic of human language. AI might struggle to correctly interpret information that could be understood in multiple ways. This can be particularly challenging when dealing with highly involved debates on social network platforms, where the intent is often hidden beneath layers of ambiguity.

**Data Bias:** AI models are trained on data, and hence, they reflect the biases present in the training data. If the input data over-represents or under-represents certain types of behavioural profiles, the trained model may not perform well in detecting risk for all types of personality traits.

**Dependence on Quality of Annotated Data:** The effectiveness of AI models relies heavily on the quality of the annotated data used for training. If this data is incorrectly annotated or not representative of real-world scenarios, the models may not perform effectively.

**Manipulation of AI Systems:** There is a risk that actors spreading disinformation could manipulate the AI system itself. They could attempt to "poison" the training data or launch adversarial attacks to decrease the system's effectiveness (Khurana et al., 2019).

**Ethical and Privacy Concerns:** Using AI for creating psychological profiles involves monitoring and analysing data, which can raise serious ethical and privacy concerns. Balancing the need for effective data processing and actionable results with the need to respect privacy is a major challenge.

**Accountability and Transparency:** AI decision-making processes, especially those involving deep learning models, can often be opaque. This lack of transparency makes it difficult to understand why certain groups of users are considered at risk of disinformation or what led to that inference.

#### **Cracking the black box open: A long way to go**

In an analytical perspective, Explainable Artificial Intelligence (XAI) possesses significant potential to tackle the complexities associated with utilizing end-user data to construct psychometric profiles for evaluating susceptibility to disinformation. The implementation of XAI delivers clearness and accountability in decision-making processes, thereby validating that the system operates as anticipated, complies with regulatory norms. Furthermore, XAI fosters confidence in end-users, enables model scrutiny, and facilitates beneficial application of AI, while concurrently diminishing compliance, legal, security, and reputational risks associated with AI in production environments (Braynen, 2022). This is of highest importance considering the rapidly

approaching AI regulation in the EU, UK, USA, and Brazil. However, there is no universally accepted technical definition for the term "explanation" (Gilpin et al., 2022). This gap necessitates establishing a criterion for effective explanations, which includes evaluating the intrinsic quality of explanations, gauging user satisfaction, and determining the depth of user comprehension about the AI process explained. Key performance indicators for XAI can encompass explanation efficacy, user satisfaction, mental models, curiosity, trust, and human-AI performance. Further, XAI can aid developers to ascertain system functionality, maintain compliance with regulatory standards, and ensure the opportunity for affected individuals to question or modify the consequent outcome (Braynen, 2022). A look into how the AI interprets the annotated data, can help spot errors or inconsistencies in the annotations, leading to improvements in the quality of the training data (Gunning et al., 2019) therefore the robustness of annotated data is a critical determinant of XAI model quality. Synthetic data can be employed to verify explainable AI models and adhere to data privacy norms. Employing synthetic data to broaden XAI across teams can stimulate collaboration, thereby enabling continuous and large-scale implementation of XAI (Tjoa and Cuntai, 2022).

The endeavour to facilitate XAI models is however facing many roadblocks. Firstly, XAI models can be intricate and perplexing, creating hurdles even for experienced data scientists and machine learning experts. There exists a significant challenge in affirming the correctness and exhaustiveness of the explanations offered by XAI. While the initial insights might be relatively straightforward, the audit trail becomes increasingly complex as the AI engine continuously interpolates the data. It can be computationally demanding, posing a considerable challenge when scaling for large AI datasets and real-world applications. This can be due to various reasons involving the need for multiple computation rounds and real time analysis. While AI models have been heavily optimised for training and inference through years of research and development, XAI is relatively new with little to no optimisation that has been done so far (Arrieta et al., 2020). These models also might not offer explanations that generalize effectively across diverse situations and contexts, limiting their

universal applicability. Additionally, there is often a trade-off between explainability and accuracy. XAI models may need to relinquish some level of precision to boost transparency and understandability which is a major complication for deployment in critical infrastructure such as this framework. Lastly, integrating XAI models with pre-existing AI systems can be challenging, necessitating substantial alterations to established processes and workflows. Therefore, these factors collectively contribute to the complexity of implementing XAI and highlight the need for further advancements in this field. It is reasonable to explore a model-specific method suited for this framework instead of a model-agnostic XAI technique to ensure robustness of this process from end to end (Rosenfeld, 2021) necessitating a significant research investment into realising this framework.

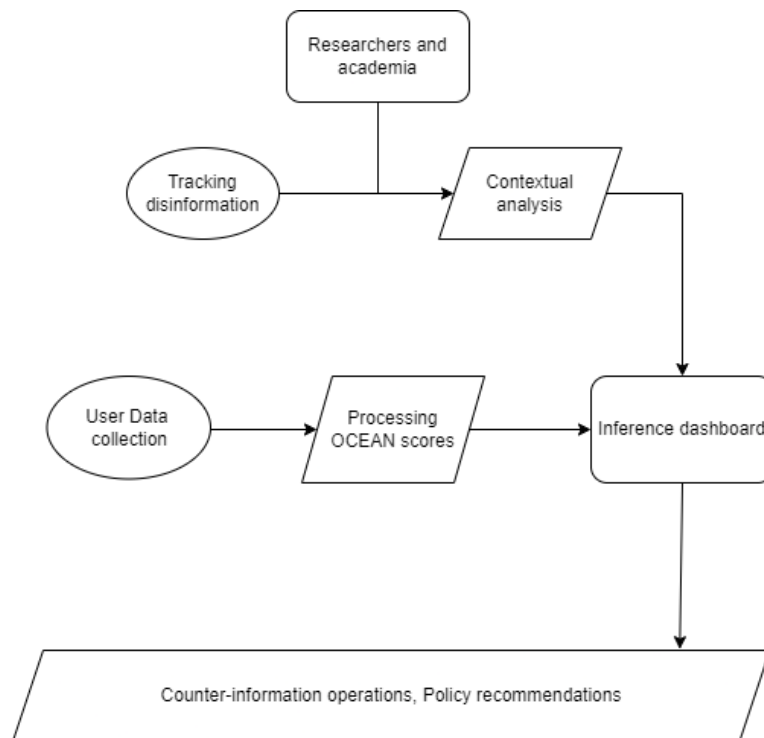
## Inference Layer

When it comes to inference, the predictions of the model about the individual's vulnerability might be off the mark, depending on the situation. How people perceive disinformation can vary widely depending on several factors, including their existing beliefs, political alignment, media literacy, trust in authority, and social network influence, among others. To produce more precise results, an external parameter can be introduced that would improve the result accuracy. This parameter concerns the context of a specific disinformation campaign which can be obtained by deploying a parallel monitoring system capable of tracking disinformation across multiple platforms. This could involve keyword tracking, sentiment analysis, and other NLP (Natural Language Processing) techniques to identify potential disinformation campaigns. The rapid development of Large Language Models (LLMs) in the past year also shines a light on greater possibility in contextual analysis (Zhang et al., 2023). It might also involve the use of AI algorithms that can detect anomalies in communication patterns, such as sudden increases in activity around a particular topic or sentiment. Analysts would study their target populations' historical, cultural, and political landscapes to better understand the potential effects of disinformation campaigns. This contextual analysis would assist in

understanding the specific societal and political contexts in which disinformation could spread.

Further, the integration of disinformation context into vulnerability prediction would involve a multi-disciplinary approach. Psychologists, data scientists, and intelligence analysts would need to work together to develop a model that considers various factors, including individuals' psychometric profiles, their exposure to disinformation, and the societal context in which they are operating. The model could then be enhanced with crisis scenarios, such as a separatist movement, pandemic, religious sentiment, or environmental issues. This would involve creating hypothetical situations and analysing how these could affect individuals' vulnerability to disinformation. Machine learning algorithms could be used to predict outcomes based on past data and the specific variables of each scenario. Below is a possible scenario describing the working of this mechanism.

*Scenario: Imagine a country experiencing a separatist movement in one of its regions. The intelligence organization predicts that a foreign adversary might exploit this situation by launching a disinformation campaign aimed at fuelling the movement and creating further division. They identify the key themes that might be used in such a campaign (such as historical grievances, ethnic identity, or economic inequality), and they feed these into their predictive model. The model, which has been trained on a vast amount of data from online platforms, would predict which aggregated groups of individuals are most likely to be influenced by such a campaign. It identifies, for example, that individuals with certain psychometric profiles (derived from the OCEAN score) who live in economically disadvantaged areas are particularly susceptible. Figure 3. Illustrates the flow of data and the process described in this scenario. With the information obtained at the inference dashboard, the intelligence organization can take proactive steps to counter the disinformation campaign, such as launching an information campaign aimed at these vulnerable individuals, alerting local authorities and community leaders, or increasing online monitoring in the affected areas. They could also prepare for potential consequences, such as an increase in protests or violence, based on the outcomes predicted by the model. This proactive approach could potentially prevent or mitigate the effects of the disinformation campaign, ensuring the country's stability during a potentially volatile situation.*



*Figure (3): Flow of data in the framework*

The inference layer includes stakeholders where people (both holding authorities and external members) have access to process data that holds valuable insights. The ability to predict vulnerability to disinformation is undoubtedly powerful. Yet, such power comes with significant ethical implications and the potential for misuse if not safeguarded correctly. Primarily, this information can be manipulated to exploit the very vulnerabilities it identifies. For instance, unscrupulous actors could use such insights to develop more targeted and effective disinformation or propaganda campaigns, exacerbating the issues the predictive model sought to mitigate. Similarly, access to such data could result in a form of "digital discrimination". Knowing the areas or individuals more susceptible to disinformation, certain groups might be stigmatized or treated unfairly, with their increased vulnerability used to justify invasive surveillance or punitive actions (Criado and Such, 2019). There is also a risk that these data could be employed for commercial exploitation. Marketing firms or corporations might seek to use the insights to manipulate consumer behaviour, promoting products or ideas to those most susceptible to persuasion (Hirsh et al., 2012). Such practices, often termed "psychographic targeting", can lead to the exploitation of vulnerable individuals and an erosion

of autonomous decision-making. Further, the data itself can become a target. Cybercriminals or foreign adversaries could also seek to breach the intelligence organization's systems to gain access to this valuable dataset. If successful, they could exploit the data or use the knowledge of the system's operation to develop countermeasures, negating the agency's predictive capabilities. Finally, the use of such data raises significant privacy concerns. The vast quantities of online data required for the model's operation could involve substantial intrusion into individuals' digital lives. If not appropriately anonymized or secured, this data collection could violate privacy rights or expose individuals to undue scrutiny. Therefore, while vulnerability prediction offers significant potential for proactively combating disinformation campaigns, it must be handled with care, transparent processes, and robust data protection measures to prevent abuse and maintain public trust. Following is a guideline involving some best practices that can help mitigate the risks mentioned above.

#### Access Control

Role-Based Access Control (RBAC) is a method used widely in corporate intranets where data access permissions are based on the roles of individual users within an organization. This ensures that only authorized personnel can access the data they need to perform their tasks. It also limits the risk of data breaches since a person with malicious intent has limited access (Ferraiolo et al., 1999). Furthermore, to maintain accountability, a system of detailed activity logs should be established to monitor and record each data access request, granting the ability to conduct audits and identify any unauthorized attempts. In this context, where government agencies have access to data regarding individual vulnerability to disinformation, a robust implementation of Role-Based Access Control (RBAC) would be critical to safeguard the privacy and rights of individuals. The system can help ensure that sensitive data is accessed only by authorized personnel for legitimate purposes.

**Data Analysts:** They are involved in the process of training the predictive models, performing data cleaning, and statistical analysis. Their access would be primarily on the 'data layer'. They would need access to the raw and processed data. However, they might not necessarily need to know the identities

of specific individuals in the data set when data anonymization methods are used.

**Investigative Officers:** Their role involves using the output of the predictive models to identify potential threats and initiate investigations. They would typically have access to the predictive models' outputs, dashboards, and alerts. They would have a higher level of access to individual context-based risk data than data analysts, particularly when an investigation is underway.

**Policy Makers:** They are involved in decision-making based on the overall trends and not the individual data points. Their access would be mainly to the key information, which consists of aggregated and anonymized reports, trends, and patterns identified by the system. They would not need access to the raw data or individual results.

**System Administrators:** They manage and maintain the system, ensuring its security and smooth operation. They would have extensive access to the 'system layer' but should ideally not have access to the raw or processed data unless necessary for system maintenance or troubleshooting. Their primary role is to control who has access to what rather than to work directly with the data or the system outputs.

The implementation of RBAC in this context would ensure that each role only has access to the data and system components necessary for their job, reducing the risk of data breaches or misuse. The system must maintain a record of who accesses what information and when to provide an audit trail and further enhance security.

### Data Anonymization

Anonymization is the process of removing personally identifiable information from the data. It aims to protect user privacy by ensuring that individual users cannot be identified from the data. This process is often complex due to the myriad ways data can be cross-referenced to identify individuals. Anonymization techniques can vary from simple data masking to more complex methods, such as differential privacy. Differential privacy involves adding statistical "noise" to the data, maintaining overall accuracy while ensuring

individual data points cannot be re-identified (Shrivastva et al., 2014). To protect individuals' privacy, we must adopt an anonymization process to ensure that the data is used responsibly and ethically. Here is how it might work:

**Data Masking:** The first step is to eliminate or mask directly identifying information. This might include names, addresses, social security numbers, or any other details that could be used to immediately identify an individual. Techniques used could include pseudonymization, data scrambling, or data blurring.

**Generalization and Bucketing:** Next, certain data points like age, income, or location might be generalized. For example, exact ages might be replaced with age ranges (e.g., 20-30), or specific locations might be replaced with larger areas (e.g., Southwest region rather than a specific city). This process further obfuscates individual identities while retaining enough information to conduct meaningful analysis.

**Noise Addition/Differential Privacy:** To provide an even higher level of privacy, the government might employ differential privacy techniques. This process involves adding statistical "noise" or randomness to the data. This noise does not significantly affect the overall patterns and trends in the data, but it does make it impossible to accurately identify any individual. For example, to a dataset about susceptibility to disinformation, random "noise" might be added to the OCEAN scores, slightly altering them but maintaining their relative values.

**Continuous Monitoring and Auditing:** It is critical to regularly audit the anonymization methods for efficacy and compliance with privacy regulations. As new potential privacy risks emerge (such as more sophisticated re-identification techniques), the government must respond and evolve its anonymization methods to ensure ongoing privacy protection.

### Data Usage Policy

A clear, comprehensive, and enforceable data usage policy is key to maintaining the ethical use of inferred data. This policy must outline what the data can be used for, who can use it, and in what circumstances. It should prohibit any

unauthorized use of data and lay out the penalties for violations (Zuiderwijk and Janssen, 2014). The policy must be made accessible to all stakeholders and be regularly reviewed and updated as necessary. Below is an example of a comprehensive and enforceable data usage policy that could be implemented for government agencies using data on the vulnerability of people to disinformation:

**Policy Development:** The first step involves creating the policy document. This should be done by a collaborative team that includes data privacy experts, legal advisors, IT experts, government representatives, and potentially also ethicists or representatives from the public to ensure a wide range of perspectives. The policy needs to delineate:

- What data will be collected: Specifically identify the types of data that will be collected and why these specific data types are necessary for the desired purposes—for example, demographic data, social media usage, education levels, etc.
- Who can use it: Define the specific government agencies that will have access to the data. Justification for each agency's access should be explicitly stated.
- In what circumstances: Outline the specific scenarios in which the data can be used. This might include, for instance, monitoring for potential disinformation campaigns, developing public awareness initiatives, or informing policy decisions.
- Unauthorized use of data: Clearly define what constitutes unauthorized use of data and prohibit it explicitly. This could include using the data for political advantage, selling it, or sharing it without consent.
- Penalties for violations: There should be strict penalties for violations, which could range from fines to job termination or even legal prosecution for severe infractions.

**Dissemination and Training:** Once the policy has been developed, it should be disseminated to all stakeholders, including all employees of the relevant government agencies, policymakers, and potentially also the civil bodies. This

may involve workshops or training sessions for employees to understand the policy and its implications. The policy should also be made publicly available, such as by posting it on government websites.

**Enforcement:** The policy must be enforced effectively to ensure compliance. This could involve regular audits of data used by an independent body or internal department. Any breaches of policy should be acted upon swiftly, with penalties applied as per the terms of the policy. There should also be channels for reporting potential breaches and protections for whistleblowers.

**Regular Review and Update:** The policy should not be static; it must be regularly reviewed and updated to reflect changes in technology, societal attitudes, legal frameworks, and the evolving understanding of disinformation. This might involve annual or biennial reviews, with amendments made as necessary. Updates to the policy should be communicated to all stakeholders in the same way as the initial policy was disseminated.

**Transparency and Accountability:** Throughout all these processes, there should be a strong emphasis on transparency and accountability. The government should be open about its use of this data, providing regular updates on how it is being used and the impacts it is having. There should also be opportunities for public input, such as consultations or feedback channels, to ensure the policy continues to meet the needs and expectations of the community. This will help to build trust and ensure the ethical use of the data.

**Third-Party Access and Sharing:** If any third parties, such as research institutions or other government departments, need access to the data, this should be outlined in the policy. It should detail under what circumstances data can be shared, what data can be shared, and the procedures that third parties must follow to ensure data privacy and protection.

**Redress Mechanisms:** The provision for redress should be laid out in the policy. If an individual feels their data has been misused or their privacy infringed upon, they should have a clear and effective means to lodge a complaint and seek resolution.

**Disaster Recovery and Breach Response:** The policy should clearly state how the government will respond in the event of a data breach or disaster, including immediate response steps, public communication, and measures to prevent future occurrences.

### Data Audit

Regular auditing of data usage helps to ensure that all stakeholders are adhering to the data usage policy. These audits can identify violations, rectify mistakes, and confirm the effectiveness of the existing data safeguards. They can also be used to highlight areas for improvement. The audits should be conducted by independent auditors for unbiased results and should be mandatory at frequent intervals. The auditors will review the data access logs, verify that data is being used in accordance with the policy, identify any potential violations or data breaches, and assess the effectiveness of the current safeguards. They may also conduct interviews with stakeholders, test the security systems, and review incident response procedures. For example, if the government agencies are using this data to implement strategies to counter disinformation, the auditors would check if these strategies were indeed aligned with the policy, whether the access to the data is limited to authorized personnel only, and whether the data has been securely stored and handled. After the audit, a report should be compiled detailing the findings, including any policy violations or security issues discovered. The report should also include recommendations for improving the current systems and procedures based on the auditors' expertise and the industry's best practices. The government agencies should then take action to address the identified issues, such as rectifying mistakes, strengthening data safeguards, retraining personnel, or even updating the data usage policy if necessary. The effectiveness of the changes should be monitored, and the data usage policy should be reviewed regularly to ensure it remains relevant as technology and societal norms evolve. This includes a periodic review of the policy, systems, and procedures in light of new risks or challenges.

### Security Measures

Protecting the data from breaches and leaks is crucial. This can be achieved through various security measures, such as data encryption, secure data storage

and transmission protocols, intrusion detection systems, and regular security audits. Encryption protects the data by converting it into a code that can only be decoded with a specific key. This means that even if someone manages to gain unauthorized access to the data, they would not be able to read it without the decryption key. In addition, a breach notification system should be in place to alert stakeholders in the event of a security breach. Apart from encryption, data is often transmitted between various departments or individuals within the agency. These transmissions must be secure and encrypted to prevent interception and unauthorized access. The agency should employ intrusion detection systems to identify any suspicious activity or attempts to breach the system. These systems can help detect and mitigate threats before they cause a data leak. Despite all these precautions, there is always a risk of a breach. That is why it is important to have a robust breach notification system in place. If a breach occurs, this system will alert the necessary stakeholders immediately, allowing them to take swift action to limit the damage. The agency's staff with access to the data should be updated about the best data protection practices. Regular training sessions and awareness programs should be organized to ensure that every staff member understands their role in protecting this sensitive data.

## Operations Layer

Intelligence agencies can actively map the susceptibility of various demographics to disinformation. By using risk data, they identify potential targets, themes, and channels of disinformation attacks. With this knowledge, they could develop counter-narratives that are released through social media, traditional media, and other platforms to neutralize the effects of disinformation. The operation would aim to correct misconceptions, emphasize the importance of validated information sources, and strengthen the public's media literacy skills. As this is a delicate operation, it would require a careful balance between being proactive and maintaining the trust and privacy of the public. Here are some ways this endeavour could be realised.

**Social Media Micro-targeting:** By using vulnerability data, social media

campaigns can be more effectively targeted to the individuals most susceptible to disinformation. Different platforms cater to different audiences, and the vulnerability data can guide decisions on which platform to use for each specific group.

**Content Customization in Traditional Media:** The content disseminated through traditional media outlets like newspapers, television, and radio can be customized according to the vulnerability data. Localised messages designed to address the concerns or misconceptions prevalent among vulnerable groups can help in increasing the effectiveness of the counter-information operation.

**Website and Blogs:** Government or organizational websites can serve as a hub for accurate information. Blogs and websites like chatrooms and forums can feature articles and resources tailored to address the disinformation narratives that have been found to resonate with those identified as vulnerable. These platforms can serve as an authoritative source of information for these individuals, who often spend many hours interacting with others in these spaces.

**Education Interventions:** Based on the vulnerability data, educational programs can be developed to address specific misconceptions or knowledge gaps that make certain groups of individuals more susceptible to disinformation. These interventions can take place in schools, colleges, and universities.

**Targeted Email Campaigns:** Using the vulnerability data, email campaigns can be targeted to groups of those most susceptible to disinformation, such as employees of a private organisation. These emails can provide detailed counterpoints to disinformation narratives that have been found to resonate with the targeted group.

**Community Engagement:** Public meetings and community events can be organized in areas identified as highly susceptible to disinformation. These provide a forum to directly engage with the community, disseminate accurate information, and discuss their concerns.

**Influencer Engagement:** Influential figures who can reach the demographic identified as vulnerable can be partnered with to spread counter-information

messages to their followers. Social media analytics tools can be used to identify these influencers based on their reach, relevance, and resonance.

**Partnerships with Civil Society Groups:** Civil society groups often have strong connections within their communities. By utilizing the vulnerability data, these groups can be effectively leveraged to distribute counter-information messages on a grassroots level, especially in communities identified as particularly vulnerable to disinformation.

Alternatively, the insights from vulnerability data can provide the basis for intelligence agencies to recommend new policies or amend existing ones.

## Limitations

Our understanding of disinformation is continually evolving, and it is difficult to fully grasp its multifaceted and dynamic nature. New forms of disinformation continue to emerge, often exploiting novel communication channels and technologies before we have a chance to understand and respond to them. This limitation could lead to potential blind spots in our defensive strategies. While counter-information operations can be effective, they also run the risk of escalating information warfare. In the case of government advertisements and influencers, their messages may be perceived as propaganda, leading to further scepticism and backlash. The creation and maintenance of effective oversight mechanisms are crucial yet complex in nature. Involving multiple stakeholders, such as policymakers, intelligence agencies, and data analysts, can lead to differing perspectives and potential conflicts of interest.

The process also requires significant human and computational resources to collect, process and analyse data, formulate strategies, and implement and oversee operations. As this framework draws on expertise from diverse fields, including artificial intelligence, machine learning, cybersecurity, cognitive psychology, sociology, politics, and digital communication, the interdisciplinary form presents a number of practical and theoretical limitations despite the absolute need to combat the multifaceted nature of disinformation. Coordinating efforts and facilitating productive collaboration across such a wide array of disciplines can be logistically challenging. Each of these disciplines has its

terminologies, methodologies, and epistemologies, and bridging these differences to achieve a common understanding and coherent action can be a considerable undertaking. Implementing this framework could require substantial resources, including the expertise and time of professionals from a variety of fields, the necessary technology and computational power to perform complex data analysis, and funding to sustain these operations. These resources may not be readily available or may need to be redirected from other initiatives. The role of private sector entities, such as social media companies, is crucial in combating disinformation. Their cooperation is required in data sharing and implementing countermeasures. However, their business interests might not always align with those of public security agencies, leading to potential conflicts. Any strategy to combat disinformation must secure public trust and demonstrate transparency in its operations. This can be difficult to achieve given the sensitive nature of this work, and any perceived misuse of personal data or infringement of privacy rights can undermine public trust. In the end, this implementation is limited in training data based on end-user consent which might skew the predictions based on the individual consent preferences of various groups of people. Languages are also a big barrier for various technological implementations discussed in this process. Many efforts are in a way to adapt NLP techniques for regional languages, which has proven to be a slow and difficult process.

Further, the integration of theoretical constructs and models from different disciplines can also increase the complexity of the framework. For instance, the integration of psychometric profiles into predictive models may require a deep understanding of both psychological theory and machine learning algorithms, which may not be readily available in a single analyst or even a single team. Since this is a novel and evolving area, there are few established research results and none with this implementation. The qualitative insights drawn in this work are based on discourses on existing case studies pertaining to individuals or parts of the socio-technological implementation, theorised to work in tandem and thus might not be exhaustive in the recommendations or guidelines provided for each section.

## Conclusion

The intersection of disinformation and national security, as evaluated in this investigation, presents a pressing challenge. As our world becomes increasingly digitized, the risk of disinformation stirs not just confusion but poses a significant threat to the integrity of nations. It can fissure societies, depreciate faith in institutions, and even precipitate conflicts. This makes the establishment of cognitive security an urgent matter, akin to safeguarding physical critical infrastructure of a state. Substantiated by various incidents, it is indisputable that weaponized disinformation has the potential to distort public sentiment, incite turbulence, and tarnish democratic processes. In turn, this places disinformation squarely within the realm of national security matters. As we navigate the complexity of such threats, a holistic understanding of, and strategy against, disinformation is non-negotiable. Similar to the integral nature of physical infrastructures for national stability, cognitive security becomes imperative for upholding the sanctity of our information and human thought ecosystem. The current research at hand illustrates that reactive measures such as fact checking and digital literacy are inadequate in confronting the convoluted and fast-paced evolution of disinformation. Drawing inspiration from cybersecurity, this thesis underscores the need for an anticipatory strategy that can promptly identify groups of people at risk from disinformation and help devise counter disinformation campaigns. It points towards the use of sophisticated analytics, predictive modelling, and continuous monitoring while highlighting the distinctive challenges presented by cognitive security. It is also worth noting that this effort is grounded in a co-regulatory approach which promises a balance of involvement and said infrastructure creation among public, private and state entities.

The comprehensive framework proposed in this thesis is a pioneering effort to proactively steer through the labyrinth of disinformation and its impact on cognitive resilience which weaves various disciplines into a cohesive model, including artificial intelligence, machine learning, cybersecurity, cognitive psychology, sociology, politics, and digital communication. The layered

approach of the presented framework promises privacy by design and strives to be shielded against abuse by both domestic and foreign actors by following a strict data trail from start to finish. However, in the brilliance of this novelty, we should not lose sight of potential constraints and areas for improvement. The hurdles associated with practical implementation, including resource allocation, are significant. The overhaul of consent and data collection practices specifically would require an unprecedented joint effort between platforms, data handlers, data processors, regulatory bodies and researchers. Owing to this overhaul, the mechanism becomes reliant on end-user consent for data collection, which may introduce bias through the under or over-representation of certain personality traits. However, It is a crucial step to ensure users' trust in the process and should be further studied for improvement. The clear recommendations included in this work for various stages of this endeavour, from data collection to processing, inference and finally, operations that help realise our end goal of combating disinformation deliberate on the inherent challenges situated within the various processes.

Further iterations should strive for streamlined structures, well-defined proactive data collection approaches, and more accessible explanations. As we step into the future, honing the applications of Explainable AI within this context, with an aim to reveal biases and provide insights, is a worthy endeavour. Rigorous testing of the model in real-world scenarios will yield invaluable insights into its adaptability and efficiency. The framework also has certain limitations due to the resources and capital required to create this mechanism, as well as the difficulty in involving different stakeholders from broad areas of society to agree on their respective roles, thereby creating a well-oiled machine. It is crucial to understand that refining this work should be viewed as a continuous process, ever ready to address new challenges in the fast-paced world of technology and disinformation. The unwavering commitment to ethical practices and public trust preservation should underscore all future efforts as well. In the words of P.N. Howard, if disinformation campaigns are "lie machines," then our aim should be to counteract them with "truth machines," a role that this framework strives to fulfil. This thesis could

chart a beginning point for deliberations in the battle against disinformation, equipping us with valuable insights and a framework for progress. The quest for cognitive resilience against disinformation is ongoing, and with the knowledge gained and groundwork laid here, we could stand better prepared to face and overcome the disinformation challenge.

## Bibliography

1. Adadi, A. and Berrada, M., 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6, pp.52138-52160.
2. Alamsyah, A., Widiyanesti, S., Putra, R.D. and Sari, P.K., 2020. Personality measurement design for ontology based platform using social media text. *Adv. Sci. Technol. Eng. Syst*, 5(3), pp.100-107.
3. Alexis Madrigal and Robinson Meyer, 2018. How Facebook's Chaotic Push Into Video Cost Hundreds of Journalists Their Jobs, *The Atlantic*. Available at: <https://www.theatlantic.com/technology/archive/2018/10/facebook-driven-video-push-may-have-cost-483-journalists-their-jobs/573403/>
4. Arfelt, E., Basin, D. and Debois, S., 2019. Monitoring the GDPR. In *Computer Security—ESORICS 2019: 24th European Symposium on Research in Computer Security, Luxembourg, September 23–27, 2019, Proceedings, Part I 24* (pp. 681-699). Springer International Publishing.
5. Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58, pp.82-115.
6. Bargar, A., Pitts, S., Butkevics, J. and McCulloh, I., 2019, May. Challenges and Opportunities to Counter Information Operations Through Social Network Analysis and Theory. In *2019 11th International Conference on Cyber Conflict (CyCon)* (Vol. 900, pp. 1-18). IEEE
7. Bietti, E., 2020, January. From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency* (pp. 210-219).

8. Braynen, A., 2022. Towards More Task-Generalized and Explainable AI Through Psychometrics (Doctoral dissertation, University of South Florida).
9. Brdiczka, O., Liu, J., Price, B., Shen, J., Patil, A., Chow, R., Bart, E. and Ducheneaut, N., 2012, May. Proactive insider threat detection through graph learning and psychological context. In 2012 IEEE Symposium on Security and Privacy Workshops (pp. 142-149). IEEE.
10. Burt, A. and Geer, D.E. (2018) "DATA PROTECTION FOR THE DISORIENTED, FROM POLICY TO PRACTICE," Flat Light. Hoover Institution, 20 November. Available at: <https://www.hoover.org/research/flat-light> (Accessed: November 27, 2022).
11. Carthy, S.L., Doody, C.B., Cox, K., O'Hora, D. and Sarma, K.M., 2020. Counter-narratives for the prevention of violent radicalisation: A systematic review of targeted interventions. *Campbell Systematic Reviews*, 16(3), p.e1106.
12. Choucri, N. and Clark, D.D., 2019. International relations in the cyber age: The co-evolution dilemma. MIT Press.
13. COGSEC (2017) What is cognitive security?, Cognitive Security & Education Forum. Available at: <https://www.cogsec.org/what-is-cognitive-security> (Accessed: November 27, 2022).
14. Criado, N. and Such, J.M., 2019. Digital discrimination. *Algorithmic regulation*, pp.82-97.
15. Daswani, N., Elbayadi, M., Daswani, N. and Elbayadi, M., 2021. Facebook Security Issues and the 2016 US Presidential Election. *Big Breaches: Cybersecurity Lessons for Everyone*, pp.97-130.
16. Davis, H., 2006. Google advertising tools: Cashing in with AdSense, AdWords, and the Google APIs. " O'Reilly Media, Inc."
17. de Arruda, G.D., Roman, N.T. and Monteiro, A.M., 2015, November. An annotated corpus for sentiment analysis in political news. In *Anais do X Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana* (pp. 101-110). SBC.

18. Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G. and Quattrociocchi, W., 2016. Echo chambers: Emotional contagion and group polarization on facebook. *Scientific reports*, 6(1), p.37825.
19. Dittrich, P.J., 2019. Tackling the spread of disinformation Why a co-regulatory approach is the right way forward for the EU. Bertelsmann Stiftung Policy Paper 12 December 2019.
20. Dowling, M.E., 2022. Cyber information operations: Cambridge Analytica's challenge to democratic legitimacy. *Journal of Cyber Policy*, 7(2), pp.230-248.
21. Durach, F., Bârgăoanu, A. and Nastasiu, C., 2020. Tackling disinformation: EU regulation of the digital space. *Romanian journal of European affairs*, 20(1).
22. European Commission, 2018. Tackling online disinformation: a European Approach (2018) GDPR.eu
23. Farrukh, M.U., Wainwright, R., Crockett, K., McLean, D. and Dagnall, N., 2022, December. Building Actionable Personas Using Machine Learning Techniques. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 463-472). IEEE.
24. Ferraiolo, D.F., Barkley, J.F. and Kuhn, D.R., 1999. A role-based access control model and reference implementation within a corporate intranet. *ACM Transactions on Information and System Security (TISSEC)*, 2(1), pp.34-64.
25. Fiesler, C., Beard, N. and Keegan, B.C., 2020, May. No robots, spiders, or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. In *Proceedings of the international AAAI conference on web and social media* (Vol. 14, pp. 187-196).
26. Figl, K., Kießling, S., Rank, C. and Vakulenko, S., 2019. Fake news flags, cognitive dissonance, and the believability of social media posts.
27. Fox News (2016). 7,500 Online Shoppers Unknowingly Sold Their Souls. [online] Fox News. Available at: <https://www.foxnews.com/tech/7500-online-shoppers-unknowingly-sold-their-souls> [Accessed 18 Jul. 2023].

28. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M. and Kagal, L., 2018, October. Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) (pp. 80-89). IEEE.
29. Gilpin, L.H., Paley, A.R., Alam, M.A., Spurlock, S. and Hammond, K.J., 2022. " Explanation" is Not a Technical Term: The Problem of Ambiguity in XAI. arXiv preprint arXiv:2207.00007.
30. Goddard, M., 2017. The EU General Data Protection Regulation (GDPR): European regulation that has a global impact. *International Journal of Market Research*, 59(6), pp.703-705.
31. Goel, V., 2014. Facebook Tinkers with Users' Emotions in News Feed Experiment, *Stirring Outcry*, N.Y. TIMES (June 29, 2014), Available at: <http://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html>.
32. Grassegger, H. and Krogerus, M., 2017. The data that turned the world upside down. *Vice Motherboard*, 28.
33. Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. and Yang, G.Z., 2019. XAI—Explainable artificial intelligence. *Science robotics*, 4(37), p.eaay7120.
34. Guo, B., Ding, Y., Sun, Y., Ma, S., Li, K. and Yu, Z., 2021. The mass, fake news, and cognition security. *Frontiers of Computer Science*, 15(3), pp.1-13.
35. Gupta, G.K. and Sharma, D.K., 2022, March. A review of overfitting solutions in smart depression detection models. In 2022 9th International Conference on Computing for Sustainable Global Development (INDIACom) (pp. 145-151). IEEE.
36. Halliday, R., 2022. 'Cyber warfare' in style: Cambridge Analytica and a mediatized ethics of fashion. *International Journal of Fashion Studies*, 9(1), pp.131-149.
37. Hasen, R.L., 2022. *Cheap speech: How disinformation poisons our politics—and how to cure it*. Yale University Press.

38. Hay, M., Miklau, G., Jensen, D., Towsley, D. and Weis, P., 2008. Resisting structural re-identification in anonymized social networks. *Proceedings of the VLDB Endowment*, 1(1), pp.102-114.
39. Hirsh, J.B., Kang, S.K. and Bodenhausen, G.V., 2012. Personalized persuasion: Tailoring persuasive appeals to recipients' personality traits. *Psychological science*, 23(6), pp.578-581.
40. Howard, P.N., 2020. *Lie machines: How to save democracy from troll armies, deceitful robots, junk news operations, and political operatives*. Yale University Press.
41. Human, S., Alt, R., Habibnia, H. and Neumann, G., 2022. Human-centric personal data protection and consenting assistant systems: towards a sustainable Digital Economy.
42. Jordan, N.S., 2011. Cognitive Dissonance as a Potential Mediator of the Misinformation Effect.
43. Kanakia, H., Shenoy, G. and Shah, J., 2019. Cambridge Analytica—a case study. *Indian Journal of Science and Technology*, 12(29), pp.1-5.
44. Khurana, N., Mittal, S., Piplai, A. and Joshi, A., 2019, October. Preventing poisoning attacks on AI based threat intelligence systems. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1-6). IEEE.
45. Khyani, D., Siddhartha, B.S., Niveditha, N.M. and Divya, B.M., 2021. An interpretation of lemmatization and stemming in natural language processing. *Journal of University of Shanghai for Science and Technology*, 22(10), pp.350-357.
46. Kim, A. and Dennis, A.R., 2019. Says who? The effects of presentation format and source rating on fake news in social media. *Mis quarterly*, 43(3), pp.1025-1039.
47. Kim, N.S. and Telman, D.A., 2015. Internet Giants as Quasi-Governmental Actors and the Limits of Contractual Consent. *Mo. L. Rev.*, 80, p.723.
48. Kosinski, M., Stillwell, D. and Graepel, T., 2013. Private traits and attributes are predictable from digital records of human behavior.

- Proceedings of the national academy of sciences, 110(15), pp.5802-5805.
49. Mäihäniemi, B., 2022. The role of behavioural economics in shaping remedies for facebook's excessive data gathering. *Computer law & security review*, 46, p.105709.
  50. Matz, S.C. and Netzer, O., 2017. Using Big Data as a window into consumers' psychology. *Current opinion in behavioral sciences*, 18, pp.7-12.
  51. Matz, S.C., Kosinski, M., Nave, G. and Stillwell, D.J., 2017. Psychological targeting as an effective approach to digital mass persuasion. *Proceedings of the national academy of sciences*, 114(48), pp.12714-12719
  52. Nissenbaum, H., 2011. A contextual approach to privacy online. *Daedalus*, 140(4), pp.32-48.
  53. Nouwens, M., Liccardi, I., Veale, M., Karger, D. and Kagal, L., 2020, April. Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1-13)
  54. Omelianenko, I., 2017. Applying Deep Machine Learning for psychodemographic profiling of Internet users using OCEAN model of personality. *arXiv preprint arXiv:1703.06914*.
  55. Papadogiannakis, E., Papadopoulos, P., Kourtellis, N. and Markatos, E.P. (2021). User Tracking in the Post-cookie Era: How Websites Bypass GDPR Consent to Track Users. *arXiv:2102.08779 [cs]*. [online] doi:10.1145/3442381.3450056.
  56. Powles, J., 2016. Google's Jigsaw project has new ideas, but an old imperial mindset. *The Guardian*. [www.theguardian.com/technology/2016/feb/18/google-alphabet-jigsaw-geopolitical-games-technology](http://www.theguardian.com/technology/2016/feb/18/google-alphabet-jigsaw-geopolitical-games-technology).
  57. Rajtmajer, S. and Susser, D., 2020, September. Automated influence and the challenge of cognitive security. In *Proceedings of the 7th Symposium on Hot Topics in the Science of Security* (pp. 1-9).

58. Rosenfeld, A., 2021, May. Better metrics for evaluating explainable artificial intelligence. In Proceedings of the 20th international conference on autonomous agents and multiagent systems (pp. 45-50).
59. Shen, Y. and Liu, J., 2021, November. Comparison of text sentiment analysis based on bert and word2vec. In 2021 IEEE 3rd International Conference on Frontiers Technology of Information and Computer (ICFTIC) (pp. 144-147). IEEE.
60. Shrivastva, K.M.P., Rizvi, M.A. and Singh, S., 2014, November. Big data privacy based on differential privacy a hope for big data. In 2014 International Conference on Computational Intelligence and Communication Networks (pp. 776-781). IEEE.
61. Singer, P.W. and Brooking, E.T., 2018. LikeWar: The weaponization of social media. Eamon Dolan Books.
62. Sloan, R.H. and Warner, R., 2014. Beyond notice and choice: Privacy, norms, and consent. *J. High Tech. L.*, 14, p.370.
63. Spagnuolo, D., Ferreira, A. and Lenzini, G., 2019, March. Accomplishing Transparency within the General Data Protection Regulation. In ICISSP (pp. 114-125).
64. Steel, E., 2010. Exploring ways to build a better consumer profile. *The Wall Street Journal*.
65. Sunstein, C. R. (2017). *#Republic: Divided democracy in the age of social media*. Princeton University Press.
66. Tappin, B.M., Wittenberg, C., Hewitt, L. and Rand, D., 2022. Quantifying the persuasive returns to political microtargeting.
67. Taylor, H., 2003. Most people are “privacy pragmatists” who, while concerned about privacy, will sometimes trade it off for other benefits. *The Harris Poll*, 17(19), p.44.
68. Tjoa, E. and Cuntai, G., 2022. Quantifying explainability of saliency methods in deep neural networks with a synthetic dataset. *IEEE Transactions on Artificial Intelligence*.
69. USSOCOM. 2018. *Defining Gray Zone Challenges*. Technical Report. US Special Operations Command.

70. Utz, C., Degeling, M., Fahl, S., Schaub, F. and Holz, T., 2019, November. (Un) informed consent: Studying GDPR consent notices in the field. In Proceedings of the 2019 acm sigsac conference on computer and communications security (pp. 973-990).
71. Van Loan, C.F., 1976. Generalizing the singular value decomposition. *SIAM Journal on numerical Analysis*, 13(1), pp.76-83.
72. Vasu, N., Ang, B., Teo, T.A., Jayakumar, S., Raizal, M. and Ahuja, J., 2018. Fake news: National security in the post-truth era. S. Rajaratnam School of International Studies.
73. Vergnolle, S., 2021. Enforcement of the DSA and the DMA-What did we learn from the GDPR?.
74. Waltzman, R., 2017. The weaponization of information. RAND, URL: [https://www.rand.org/content/dam/rand/pubs/testimonies/CT400/CT473/RAND\\_CT473.pdf](https://www.rand.org/content/dam/rand/pubs/testimonies/CT400/CT473/RAND_CT473.pdf), accessed, 10, pp.11-18.
75. Webster, J.J. and Kit, C., 1992. Tokenization as the initial phase in NLP. In COLING 1992 volume 4: The 14th international conference on computational linguistics.
76. Westin, A.F., 1968. Privacy and freedom. *Washington and Lee Law Review*, 25(1), p.166.
77. Zhang, W., Deng, Y., Liu, B., Pan, S.J. and Bing, L., 2023. Sentiment Analysis in the Era of Large Language Models: A Reality Check. arXiv preprint arXiv:2305.15005.
78. Zuiderwijk, A. and Janssen, M., 2014. Open data policies, their implementation and impact: A framework for comparison. *Government information quarterly*, 31(1), pp.17-29.