Cognitive Strategies for Curbing Online Misinformation: The Potential of Nudges

Pedro Pinilla Plaza

Research Master's Thesis (30 EC)

Supervisor: Prof. Dr. Jack Vromen

Advisor: Dr. Conrad Heilmann

34,001 words

Erasmus Institute for Philosophy and Economics

Erasmus University Rotterdam

July 2023

Contents

Acknowledgement

I want to thank my parents, Pedro and Victoria, for their unconditional love, and Chiara and Paloma for filling my days with the best memes. The four of you make my life worth living.

I would also like to thank Jack Vromen for his patience and guidance. Without his invaluable comments, this thesis would not exist, but also make me wish the writing process had occurred under different circumstances, ones that allowed me to explore all theoretical possibilities open by them.

Chapter 1

Introduction

Misinformation and fake news have crowded social media in the last few years. While the spread of misinformation and its use as a political weapon is hardly a new phenomenon (Pennycook and Rand, 2021; Waldman, 2018; Hundley, 2017), recent events have clearly shown that social media platforms offer a new way to accelerate their diffusion and reach. Thus, for example, regarding the case that sparked the new interest in researching misinformation, the 2016 US elections, a profusely cited study (Allcott and Gentzkow, 2017) "estimated that a particular set of news stories that are known to be false were shared on Facebook at least 38 million times in the 3 months leading up to the 2016 election (30 million of which were for news favoring Donald Trump)" (Pennycook and Rand, 2021, p. 389). A trend that has continued in events like the 2016 Brazilian presidential election, the COVID-19 pandemic, the 2020 US election, and the subsequent storming of the Capitol, or the war in Ukraine just to mention some examples.

Following the development of the spread of misinformation online, researchers quickly jumped to the study of the cognitive mechanisms that make people fall for – and share – misinformation. Fortunately, researchers did not have to start completely anew since they could draw from the findings made in fields like decision theory or behavioral economics (i.e., Kahneman and Tversky, 1979; Starmer, 2000; Kahneman, 2003) as well as in the literature on conspiracy belief, rumors, or bullshit receptivity (Pennycook and Rand, 2021; Sunstein and Vermeule, 2008; Lindeman and Aarnio, 2007; Berinsky, 2017; Pennycook, Cheyne, Barr, Koehler, and Fugelsang, 2015). Thus, even though the cognitive science of misinformation is still in its infancy, we can already see it featured in some compendiums of political epistemology (Greifender, Jaffé, Newman, and Schwarz, 2021; Hannon and de Ridder, 2021; Edenberg and Hannon, 2021). There, we can appreciate that, although the results might not be robust enough to deem them conclusive, there is a rich variety of theories that would explain the spread of misinformation: from those that blame it on the use of heuristics (i.e., source, familiarity, or political party signaling heuristics) to that which points at the effects of motivated reasoning.

However, social media platforms, governmental institutions, and other researchers have not stayed expectant and have proposed multiple kinds of interventions that aim at curtailing misinformation online. A common strategy social media companies follow is relying on professional fact-checkers to determine news veracity. Once fact-checkers' work is done, platforms have different options for implementing the feedback, from merely warning users about the news' quality to downranking the false and misleading articles, or directly deleting them (Walter, Cohen, Holbert, and Morag, 2020; Nieminen and Rapeli, 2019). Given the impossibility of upscaling human fact-checking to all the dubious news and the problems derived from only covering some of them (i.e., the implied-truth effect, by which people may think that all the news that has not been flagged as false is therefore true, even though most of them could have not been checked at all), social media platforms have started using tools like machine learning and natural language processing to automate the process. Nonetheless, other approaches to curtailing misinformation avoid fact-checking altogether and employ techniques like 'inoculation' – whereby people learn how to identify misleading news – or redesigning the platforms so users interact with news differently.

When it comes to selecting which measure to introduce the decision can be justified, naturally, on many grounds. The one that may come to mind more straightforwardly is to compare different interventions' efficacy, that is, how successful they are in curtailing misinformation, which in turn can be measured in different ways, like people's engagement with the article, time spent reading it, or influence on people's posterior beliefs and behavior. Another way would be to evaluate the interventions' normative credentials. Here, again, there are multiple possibilities depending on the normative value that we want to focus on, from interventions' paternalist components to analyzing how interventions meddle with people's free speech rights or their impact on people's privacy, just to name a few. Yet, another approach to evaluating interventions would be to assess their relationship with the findings made in the cognitive sciences. That is, whether, and if so to what extent, the interventions work as they are intended to at the cognitive level. In other words, it is an assessment of the *coherence* between the (often implicit) cognitive mechanism assumed by the proposed interventions and the state-of-the-art findings of the mechanisms that make people fall for and/or share misinformation online[1].

This thesis takes the third route and aims to provide the conceptual tools that make possible a *coherence evaluation* of interventions in *the context of online misinformation*. In particular, I will focus on *nudges*, a type of intervention that, following a *libertarian paternalist approach*, aims at helping decision-makers to make better choices (Thaler and Sunstein, 2009). Nudges have been implemented in a wide variety of contexts since Thaler and Sunstein

[1] For the sake of brevity, in the remaining of the thesis, I will refer to his type of evaluation that looks at the alignment between the cognitive assumptions made by interventions and particular cognitive theories as 'coherence evaluation'.

created them over a decade ago although they have also been heavily criticized, especially from a normative point of view (i.e., Bovens, 2009; Hausman and Welch, 2010; Grüne-Yanoff 2012). Importantly for the purpose of this thesis, *coherence* critiques have also gone underway (Heilmann, 2014; Grüne-Yanoff and Hertwig, 2016). Thus, in the first part of this thesis (Chapter 2), I will continue this thread and propose a definition of nudge that specifies its cognitive assumptions in terms of a concrete dual-process theory of human reasoning (Pennycook, Fugelsang, and Koehler, 2015).

Once the definition of nudge has been provided, in the second part of the thesis (Chapter 3), the goal is to apply it in the context of online misinformation. In the last years, researchers and social media platforms have also thought of nudges as a solution to steer people away from misinformation (Thornhill, Meeus, Peperkamp, and Berendt, 2019; Horne, Gruppi, and Adali, 2019; Pennycook, Epstein, Moshel, Arechar, Eckels, and Rand, 2021). However, few comprehensive critiques, neither methodological nor normative, have taken off to this day, probably given their very short life or the fact that interventions on social media pop up at such a high speed that hinder careful evaluations of them. In this thesis, I will start filling this gap by analyzing whether nudges (as originally defined in the first part) are compatible with two of the theories that try to explain the cognitive reasons for the spread of online misinformation (Kahan 2013, 2016a; Tappin, Pennycook, and Rand 2020a, 2020b). Bringing all this together, the research question addressed in this thesis is: Do the cognitive mechanisms behind the spread of online misinformation allow for the introduction of nudges that aim at curtailing it?

In this introductory chapter, I will first briefly clarify the importance of conducting coherence evaluations alongside those focused on efficiency and normative credentials. In way of illustrating the case, I will introduce two interventions – labeled as nudges by their authors – that try to curtail the spread of misinformation (Section 1.1). Next, I will then formulate my research question together with the contribution of this thesis to the emerging cognitive science of misinformation (Section 1.2). Lastly, I will present the outline of the chapters of the thesis (Section 1.3).

1.1 Coherence evaluations of interventions aimed at curtailing online misinformation

Above, I mentioned that we can evaluate an intervention along (at least) three categories: its efficacy, its justification based on – and its impact on – different normative values, and its cognitive precision. While the first two can be seen as particularly straightforward – designing an intervention that has the desired impact while promoting certain normative values and avoiding violating others –, one could even question why to worry about carrying out the later type of evaluation. After all, it could be that an intervention is effective in achieving its goal of promoting a normatively desired outcome without being cognitively precise. This would be the case whenever the intervention inadvertently plays with cognitive mechanisms other than the postulated in the designing process. Given such a possibility, what would be the added value of getting the cognitive mechanisms right? In the following, I will briefly introduce two arguments in favor of coherence evaluations (Heilmann, 2014; Grüne-Yanoff, 2016).

Coherence evaluations and efficacy

Maybe the most direct response would be just to claim that while it might be possible to implement an intervention that successfully curtails the impact of misinformation online without *theoretically* getting right the mechanisms responsible for such an outcome, the safest way to ensure that the intervention is "effective, robust, persistent or welfare-improving" is to know the cognitive mechanisms behind the target behavior and how the intervention would affect them. Thus, in the words of Grüne-Yanoff (2016, p. 18),

> Mechanistic information is often highly important for assessing the effectiveness and welfare consequences of a given policy. In particular, without the right kind of information, we often cannot tell whether in a particular context, a policy is effective, is robust or persistent, nor whether it has a positive welfare effect. Yet these properties are often crucial for the justification of the policy in this context: without efficacy, we cannot justify the inductive inference from a study environment to the target environment. Without robustness and persistence, we often cannot justify the policy even from difference-making evidence obtained from the target environment itself.

According to Grüne-Yanoff, knowing the (cognitive) mechanisms would allow the policy-maker to gain some certainty about the performance of the intervention outside the laboratory (external validity) once it has been proved that it is efficacious there (internal validity). That is so because knowing the (cognitive) mechanisms would allow assessing whether the necessary background conditions for the working of the intervention are also present in the real-life setting. Without the information about the mechanisms then it would be hard to guarantee

that there are no other dynamics in the target environment preventing partially or completely the working of the intervention.

Additionally, Grüne-Yanoff also argues that even in the cases in which there is evidence about the intervention affecting the variable of interest in a real-life scenario, such evidence does not guarantee that the effect will last, that is, that the intervention is *persistent*. It might be the case, according to Grüne-Yanoff, that a repeated implementation of the intervention could alter the structural relation between the target variable and the desired behavioral effect in such a way that the latter is worn off. Thus, in Grüne-Yanoff's words, "Non-persistence is unlikely to be picked up by field experiments on target populations, as this would require very broad and long study perspectives. Instead, indications that such factors might be at work are better obtained from experiments that explicitly seek to produce mechanistic evidence" (ibid., p. 15).

It is worth mentioning that Grüne-Yanoff's study is not restricted to any particular context and aims to cover *behavioral policies*, broadly understood. To briefly see what Grüne-Yanoff's concerns look like in the context of interventions targeting misinformation, let me introduce BalancedView, an intervention conceived by Thornhill, Meeus, Peperkamp, and Berendt (2019). This proof-of-concept tool consists of a redesign of the social platform Twitter so that whenever a user posts a tweet about a political topic, the tool would detect it and present the user with articles from three trustworthy sources of different ideology commenting on the same topic. Figure 1.1 shows what BalancedView would look like.

Figure 1.1: (Thornhill et al., 2019, p. 4)

It is the intention of Thornhill and colleagues that BalancedView would counter *confirmation bias* and *fight the spread of misinformation* by *nudging* people into critically consuming the news and opinions that they find in their feeds. The hypothesis is that some people's online environments can be described as echo chambers, situations in which people are not exposed to a diversity of topics and/or opinions on certain topics, which leads them into believing and sharing news and articles regardless of their veracity and seek for information that reinforces such points of view. BalancedView, with their change of choice architecture, would therefore look for exposing people to a more diverse media diet such that it triggers in them "the set of questions that a professional fact checker would ask [when investigating the reputability and veracity of sources and stories]" (ibid., p.3). In other words, BalancedView intends to "gently steer users towards adopting fact-checking habits in their behavior online" (ibid., p. 8).

However, the precise working of the BalancedView intervention might be not so clear if we look at it following Grüne-Yanoff's approach. In particular, other understandings of confirmation bias, one that, for example, emphasized the cognitive mechanisms behind it might question that BalanceView works the way it is supposed to. As we will see in more detail in Chapter 3, if we follow Kahan (2016a) and define confirmation bias as a way of information processing that bases the likelihood of the new information's veracity on the alignment of such information with the beliefs previously held by the user, then it is unclear how an intervention like BalancedView can counteract *confirmation bias*. We could imagine that a user would show confirmation bias by discrediting the news articles and opinions suggested by BalancedView that go against her previous beliefs, rendering the intervention unable to interfere with the bias. That is not to say that BalancedView would not have any influence on users' ways of processing information but that in order to properly understand how the intervention may work – if at all –, either as a single-shot application or as a habit-forming one, we need more detailed descriptions of the cognitive mechanisms behind the processing of new information as well as of the routes through which the intervention affect the working of such mechanisms.

Coherence evaluations and normativity

However, improving (analysis of) interventions' efficacy is not the only perk of carrying out coherence evaluations. Normative analysis of interventions can also greatly benefit from studies of the cognitive assumptions made by the interventions. This is especially relevant for those kinds of interventions that are justified and/or praised based on their normative credentials. This second line of defense of coherence evaluations of interventions is the one put forward by Heilmann (2014) for the case of nudge policies, which will be also the focus of this thesis.

In a nutshell, nudges are implemented in situations in which people end up choosing non-preferred options only because of how the options are displayed

(Thaler and Sunstein, 2009). The classic example in the literature is that of someone choosing an unhealthy snack in the cafeteria or the cashier line just because of their prominent placement. A nudge then would change how options are shown so that the supposedly preferred option is also the *easiest* one to be selected. That is, nudges use a *rebiasing strategy* that plays with elements of the choice architecture so that people are biased toward a beneficial choice instead of biased towards a detrimental one. In the example, the nudge would consist of giving a piece of fruit the spotlight – under the assumption that people prefer to eat healthily.

Thaler and Sunstein, the authors that set nudges in motion, also coined the term 'libertarian paternalism' to characterize this kind of intervention. By defining nudges as libertarian paternalist, Thaler and Sunstein emphasized their two supposedly main appeals, people being nudged are not deprived of any option and remain in control of the decision (libertarian feature), while they are nudged towards the option that would make them better off *in their own view* (soft paternalist feature). Thus, it is not only that nudges aim at being highly effective – as they make use of the findings made in the behavioral sciences about how to influence people's choices – and cheap – since they only involve changes of the choice architecture, not, for example, educational programs –, but they are also normatively desirable. However, as was mentioned above, normative as some authors have called into question both their libertarian credentials – claiming, for example, that nudges might actually be manipulative, taking away people's agency – and the level of paternalism involved – for example, in some cases it may not be easy to ascertain people's preference and therefore nudge them toward them, while in other cases the mere fact of public and/or private institutions trying to know such preferences could be normatively problematic.

Heilmann (2014), on the other hand, takes a different route to criticize nudges and puts the focus on the *methodological conditions* for nudges to be successful. As will be shown in greater detail in Chapter 2, Heilmann's goal is to provide a framework that specifies the cognitive assumptions behind Thaler and Sunstein's original understanding of nudges. Heilmann distinguishes between four such assumptions, which specify (1) someone's initial state of mind, (2) the cognitive change that the intervention aims at bringing about, (3) the state of mind in the nudge position, and (4) that the state of mind that ultimately determines the choice. Looking at nudges in such a way allows Heilmann to i) compare nudges with other nudge-like interventions that differ from them in only some of the cognitive assumptions, ii) conclude that nudges, after all, might not be so straightforward to implement given that the assumptions could not be easy to fulfill, and iii) provide a framework with which to carefully detail the normative critiques against nudges; that is, to analyze what is normatively wrong with nudges at the cognitive level. Thus, for example, if Heilmann's framework differentiates between manipulative interventions and nudges at the cognitive level, what do the different accusations against nudges of manipula-

tion amount to? Are they all ill-headed by failing to recognize the supposed cognitive differences between nudges and manipulations? Or could (different kinds of) manipulation be defined differently and still, be applied to nudges? Helping in answering questions of this sort is one of the benefits of carrying out coherence evaluations of nudges.

To see the relevance of coherence evaluations of nudges that aim at curtailing online misinformation, let me briefly introduce Horne, Gruppi, and Adali's (2019) *Trust Nudging* intervention. We could imagine Trust Nudging as an intervention visually identical to BalancedView but that, instead of showing articles from three reputable yet ideologically different news sources, it presents the user that wants to share an article on Twitter with another article that while very similar to the original one, proceeds from a source of slightly higher quality. The authors intend that one step at a time, after a certain number of nudges, the user would consume news from sources of the highest quality (and least partisan) sources. The reason for this progressive transition until high-quality sources is that if the user, regardless of her original opinion, is directly and only presented with reputable sources, there would be a risk of the user not trusting the suggested articles, rendering the intervention useless[2]. Trust Nudging would, thus, progressively build trust in better news sources.

But how could coherence studies of an intervention like Trust Nudging impact normative evaluations? As mentioned above, one of the cognitive assumptions in Heilmann´s framework has to do with the user´s initial state of mind. In particular, nudges start from the premise that people do not act on their *real* preferences due to the choice architecture. However, there might be cases where people hold contradicting preferences. For example, consider a situation in which someone who believes in the importance of sharing only truthful information also is firmly convinced of the veracity of some piece of misinformation. If a Trust Nudging intervention is applied, an evaluation of its normative credentials in terms of manipulation would then depend on the cognitive assumption made regarding the initial state of mind: the same intervention could be seen either as a nudge or as a manipulative influence depending on whether the initial assumption involves the general preference for sharing truthful information or the specific belief on the misinformation.

[2]Grüne-Yanoff´s call for intervention´s mechanical evidence could also be applied here if, for example, Trust Nudging would play not only with trust but with confirmation bias as well.

1.2 Research question and significance

The research question addressed in the thesis is: Do the cognitive mechanisms behind the spread of online misinformation allow for the introduction of nudges that aim at curtailing it? In order to tackle this question, I will follow the next twofold strategy:

1. Developing a conceptual framework that specifies in a detailed way the cognitive requisites for the introduction of nudges.

2. Apply the framework to assess whether nudges are possible under two competing theories about the cognitive causes of the spread of misinformation.

One contribution of this thesis is to critically develop the coherence critique of nudges initiated by Heilmann (2014). In his article, Heilmann sets four conditions for nudges to be methodologically successful. That is, he sketches the cognitive mechanisms that need to be in place to introduce nudges in any context. One of the main features of Heilmann's framework is that it is spelled out using simple dual-system language. This thesis revises Heilmann's framework by pointing out the potential ambiguities entailed by its language and proposes exchanging it for another one that solves them by diving deeper into the dual-process framework (Evans and Stanovich, 2013; Pennycook, Fugelsang, and Koehler, 2015). The extension of the new framework allows for applying it to contexts in which the complexities of the cognitive mechanisms present there would be muted if we analyze them through the lenses of Heilmann's framework. For example, I would argue that the new framework offers better tools for analyzing cases of motivated reasoning, where mere references to *automatic* and *reflective* systems could obscure the subtleties involved in such mental processes. At the same time, a finer-grained framework offers the possibility of precisely locating potential problems when introducing nudges as well as a baseline platform into which different dual-process theories could be expressed. However, it is important to note that extending Heilmann's framework in such a way could rise the *methodological bar* even higher for nudges, that is, they might be even harder to implement and justify.

Another contribution of this thesis is to lay the ground for future studies about the coherence possibilities of introducing nudges but also nudge-like interventions in the context of online misinformation. The present work considers two main theories explaining the belief and sharing of misinformation on social media – politically motivated reasoning, and people's lack of attention – but there are other theories upon which *potentially* develop nudges and nudge-like interventions, like source heuristic or familiarity heuristic. The current thesis also sets the path for detailed normative analyses of (nudge and nudge-like) interventions aiming to curtail online misinformation. For example, evaluations of the impacts of interventions on people's freedom of speech could use the present framework – or others based on it – to specify the cognitive mechanisms

that preserve freedom of speech and how interventions possibly meddle with them. Finally, this thesis has also the potential to help in developing empirical hypotheses about existing interventions as well as in designing new ones.

Before moving on to the structure of the thesis, it is worth noting some of the limitations of it. As mentioned, this thesis develops a coherence analysis of nudges introduced to curtail online misinformation. It does so by following a twofold strategy, it first builds a general definition of nudge specifying the presupposed cognitive mechanisms and then applies it to the case of online misinformation. This means that the thesis brings together literatures that are rarely found together, behavioral economics, and the psychology of misinformation. However, the novelty of this interdisciplinary endeavor, the vastness of approaches within the two fields, and the limited space in the thesis impede any pretension of developing a comprehensive work. This implies, for example, that the methodological framework developed in the first part as well as its application in the context of online misinformation constitutes only a first and exploratory step, one that rests on a singular definition of what a nudge is. The definition of nudge presented in the thesis is restrictive in two senses; first, it builds on Heilmann's definition, which is, in turn, a reconstruction of Thaler and Sunstein's (2009). This already leaves out other understandings of what nudges consist of that were developed after Thaler and Sunstein's book; for example, definitions of nudges that account for interventions that steer people into socially beneficial choices (Guala and Mittone, 2015). But, additionally, the definition put forward in this thesis expands on that by Heilmann in a way that ties it to a more complex dual-process model of human cognition, something against which Heilmann might object given the care taken in his work at differentiating the assumption behind nudges and the dual-system language employed. The new nudge definition's extra commitment to the dual-process framework is, thus, a conscious decision that acknowledges that even if such a framework as a whole has been called into question in the last years (Sahlin, Wallin, and Persson, 2010), it is still guiding research lines in many fields, including those studying the cognitive aspects and epistemology of the consumption of misinformation (i.e., Brown, 2021; Pennycook and Rand, 2019, 2021, 2022; Ross, Rand, and Pennycook, 2021). Thus, it is the hope of this thesis to shed new light on this new interdisciplinary field, with the hope of being as well of practical use for the analysis and (re)design of existing and future interventions fighting online misinformation.

1.3 Structure of the thesis

The thesis is structured as follows:

In Chapter 2, I start by introducing Heilmann's (2014) conceptual framework of nudges' success conditions (Section 2.1). I put special emphasis on the differences between the cognitive assumptions *per se* and the dual-system language in which they are formulated, and on the take on the dual-system approach that can be inferred from Heilmann's depiction of the framework. Once the cognitive assumptions behind nudges and nudge-like interventions are introduced, I state the main limitations of the framework. These limitations revolve around the ambiguities distilled from a lack of clarity on the specification of the exact working of the two types of processes.

In Section 2.2, I follow Evans and Stanovich (2013) to contextualize the limitations found in Heilmann's framework. In particular, I focus on the distinction between dual-process and dual-system accounts, favoring the former. And more importantly, I trace back the framework's ambiguities regarding the working of the two kinds of processes to two main types of cognitive architecture proposed in the dual-process literature: the parallel-competitive account and the default-interventionist one.

In Section 2.3, I introduce Pennycook, Fugelsang, and Koehler's (2015) model of human reasoning. This model aims at synthesizing the parallel-competitive and the default-interventionist accounts by breaking down human reasoning into three stages that clearly specify how Type 1 and Type 2 processes are triggered and how they interact with each other. This model tackles the limitations found in Heilmann's framework and will serve as the basis for formulating the new framework in Section 2.4.

In Section 2.4, I develop a new framework that formulates the same four cognitive assumptions behind nudges as Heilmann's framework, but it does so by employing a new language that incorporates Evans and Stanovich's view on dual-process theories as well as Pennycook, Fugelsang, and Koehler's three-stage dual-process model of human reasoning. The new framework substantially modifies three assumptions compared with Heilmann's: the initial state of mind, the nudge position – which more clearly specifies the cognitive states that ensures that someone could resist the nudge –, and the choice position – which defines nudges as interventions that make people *rationalize* the option that they would choose if directly asked about it.

In Section 3, I apply the framework to the two main theories that explain the increase in beliefs in, and shares of, misinformation on social media: politically motivated reasoning and people's lack of attention to accuracy when scrolling down through their feeds. The goal in both cases is to evaluate whether any of the two scenarios met the demanding cognitive assumptions that would grant the introduction of nudges to curtail misinformation.

In Section 3.1, I focus on politically motivated reasoning (Kahan 2013, 2016a), which explains the spread of misinformation as the consequence of a way of processing information in which people weigh the new information based on its congeniality with their political predispositions. Moreover, politically motivated reasoning also seems to correlate with cognitive sophistication. After the theory is introduced in its own terms, I proceed to translate it into the dual-process language so that the framework developed in Section 2.3 can be applied. The key to the matter will be differentiating whether politically motivated reasoning is due to a rationalization of an intuitive response or if instead is the conclusion arrived at after cognitive decoupling Type 2 processes.

In Section 3.2, I focus on the theory that blames a lack of attention to accuracy for the spread of misinformation. I start this section by considering some arguments against politically motivated reasoning that in turn pave the way for introducing the theory that claims that people are good at differentiating between true and false/misleading information, and they have the genuine intention of sharing only truthful information (Tappin, Pennycook, and Rand, 2020a, 2020b). What would explain the prominence of misinformation online is the design of social media which diverts people's attention from accuracy (Pennycook, Epstein, Mosleh, Arechar, Eckles and Rand, 2021). Once the theory is introduced in its own terms, the strategy is the same as in the previous section, I first translate it into the dual-process language developed above to then apply the framework to analyze whether the demanding cognitive assumptions that would grant the introduction of nudges are met.

In Chapter 4, I conclude this thesis with a summary of (1) the new conceptual framework of the cognitive assumptions behind nudges, (2) the application of the framework in the context of the spread of misinformation online, and (3) indications for future possible research paths as well as restating the limitations of the thesis.

Chapter 2

A new coherence framework for Nudges

This thesis aims to investigate whether it is methodologically possible to introduce nudges in social media to curtail the spread of misinformation. The thesis will proceed in two steps to properly analyze such a research question. The first of them, which conforms to the present chapter, will provide a methodological definition of nudges in terms of the cognitive mechanisms involved in the pre-intervention scenario and detail how the nudge would aim at modifying them. This definition will define nudges generally, without specifying their context of application. The second step will then move to analyze whether is possible to introduce nudges to combat online misinformation in social media, given two competing explanations of the cognitive causes of the spread of misinformation. This second step will be the object of Chapter 3.

In the introductory chapter, we introduced some arguments in favor of carrying out methodological evaluations of behavioral interventions. In particular, we saw that detailing the cognitive mechanisms behind the different kinds of interventions is a great way for gaining confidence about their efficacy in the short term as well as guaranteeing that the effects would not dissipate or turn perversive after repeated expositions (Grüne-Yanoff, 2016). Moreover, we also saw that such coherence evaluations are of important relevance for interventions that are promoted because of their supposedly normative appeal. For example, if nudges intend to help people to choose their preferred options without curtailing any of the rest or being manipulative, then it is important to know how exactly the influence is done and in what ways nudges diverge from manipulations. The present chapter will carry out a coherence evaluation of nudges, offering as its outcome a framework that defines nudges – and other related interventions – according to the cognitive mechanisms behind them with a level of sophistication that renders the framework fitted to be applied in the context of online misinformation.

The framework developed in this chapter will be based on that created by Heilmann (2014). Both frameworks have a similar goal: specifying the cognitive mechanisms behind nudge interventions. However, as will be argued below, Heilmann's might have some characteristics that make it less than ideal for being applied to contexts in which the cognitive mechanisms behind the undesired behavior are formulated with a certain degree of complexity. Studying how to extend Heilmann's framework and ultimately developing a new one that can fruitfully be used to assess such contexts is the goal of the present chapter. Thus, the road ahead is as follows: In section 2.1, I introduce Heilmann's (2014) success conditions for nudge and nudge-like interventions. Such success conditions consist of four assumptions that specify -in a *minimal dual-process language*- both the individual's initial state of mind and what the interventions seek to trigger in the decision-making process. After a brief explanation of the framework, I end this section by pointing out the limitations that the minimal commitment to the dual-process perspective entails. Section 2.2 zooms out and contextualizes how Heilmann's take on dual-process theories compares with other views in the field. I do so by following Evans and Stanovich's (2013) review of the evolution of dual-process and dual-systems theories over the years. I conclude the section by stating that (1) the dual-system terminology in Heilmann's language is outdated and should be exchanged by a dual-process one, and (2) the ambiguous reading of the deliberative system in Heilmann's framework just follows the debate between *parallel-competitive* and *default-interventionist* accounts of cognitive architecture. In section 2.3, I introduce Pennycook, Fugelsang, and Koehler's dual-process model of analytical engagement as I argue that such a model (1) can address the limitations pointed out in Heilmann's, and (2) could serve as the base with which formulate the new framework. Finally, in section 2.4, I develop the new framework, which reformulates Heilmann's assumptions making use of a language that synthesizes the findings of the previous two sections.

2.1 Preliminary understanding of Nudge interventions: Heilmann's (2014) framework of nudges' cognitive assumptions and its limitations.

Nudges are a type of intervention born out of the findings made in fields like behavioral economics, bounded rationality, and psychological decision theory showing that, in real-life contexts, people's decisions deviate from what is expected according to rational choice theory. Just to mention a couple of such deviations, it has been found that people overweight low probabilities and underweight large ones because they use heuristics instead of following the rules of probability calculus, leading them to make 'biased choices'. It has also been discovered that people are more risk-averse regarding gains than when considering possible losses. Or, in the context of social media and online misinformation, there is evidence of people sharing misinformation even though they (1) have the ability to discern between true and false news and (2) claim to care about only sharing truthful information online. What makes these kinds of biased responses interesting is that they seem to be caused by how the options are presented to decision-makers. The idea behind nudges is that by changing the *choice architecture* – how options are displayed –, nudges could exploit some other cognitive mechanisms in order to make people go for the options that they would have chosen were they to carefully reflect on it. That is, nudges try to rebias people so that instead of showing a bias toward a non-desired option, the modification in the way in which options are presented makes people *effortlessly* go for what supposedly is their preferred option. Note here that it would not be nudges' intention to *debias* people, to change the choice architecture so that they do not show any biased behavior at all and instead think carefully before choosing anything; nudges still try to *facilitate* the selection of a particular option over the rest.

Seeing nudges as interventions that "facilitate the selection of the supposedly preferred option over the rest" points at the elements because of which nudges are also known as Libertarian Paternalist interventions. Libertarian Paternalism is, together with the efficacy granted by being backed by findings on the behavioral sciences, the main selling point of nudges. No other kind of intervention has previously attempted to reconcile these two seemingly opposed approaches, libertarianism, and paternalism. Nudges would do so by designing interventions that do not deprive people of any option (libertarian feature) while at the same time steering them towards the option that they would consider the best if they were to deliberate about it (*soft* paternalist feature). That is, while the *nudgee* is free to choose any of the options, the one that the nudge would make more easily accessible is not the one preferred by the policy-maker but instead that which the person nudged would choose under ideal conditions of deliberation.

Analyzing what needs to happen at the cognitive level in order for an intervention to embody the libertarian paternalistic features is the goal of Heilmann's (2014) methodological critique of nudges. As we shall see, to research it Heilmann develops a framework that, using a dual-system language (section 2.1.1), specifies four cognitive assumptions behind nudges (section 2.1.2). These assumptions clarify what is the mental state of the nudgee before the intervention is introduced and how the latter aims at modifying such mental state. Once the framework is introduced, I follow Heilmann and use it to compare nudges with other nudge-like interventions (section 2.1.3). Finally, in section 2.1.4, I reflect on what I find to be the main limitations of Heilmann's framework: the possible ambiguities that its take on dual-system theories could incur to; in particular, those regarding the exact working and relationships between the two supposed systems.

It is important to remark that what will be criticized is Heilmann's take on the dual-process view but not the assumptions of the framework. This is a possibility opened up by Heilmann, since he clearly differentiates between the dual-process view and the framework, while not fully committing to the former (i.e., "It is important to note that the above characterization depends only loosely on the general adequacy of the dual-process framework, and even when rejecting the latter, the four assumptions by themselves are still useful" (ibid., p.80)). It will be the task of sections 2.2 and 2.3 to contextualize the limitations on Heilmann's framework and introduce a new dual-process model of human cognition, to then, in section 2.4, use these findings to develop a new framework that modifies Heilmann's. It might be surprising to the reader to find out that even though Heilmann does not want to link the validity of his framework to that of the dual-process theories, the suggested modifications involve diving deeper into this kind of theory. The main reason for such a decision has to do with the fact that, as we shall in Chapter 3, dual-process theories are still of use in the field of political psychology looking at the spread of online misinformation. This, of course, renders the arguments made in the thesis dependent on the validity of the dual-process model introduced below, and therefore less flexible than those found in Heilmann's.

2.1.1 Framework's minimal language

Heilmann's minimal language consists of abbreviations and shorthand terms that make spelling out nudges' cognitive assumptions more efficient. It does not, however, form any kind of formal language, and the author is clear about that. Thus, we have that **A** stands for the automatic system, **R** for the reflective one, and **AR** for the decision-maker, and we can specify what proposition (prudent, no-prudent, or unknown) each system endorses by writing, for example, **ApRp**, **AqRp**, **AqR?**, or **A?R?**. The language also specifies that in order to signal a change of state within a system, an arrow can be written between the original and the new states.

2.1.2 The actual framework

Once the minimal language has been introduced, let's see how exactly Heilmann details the four assumptions behind *a narrowly successful nudge* (ibid., p79-80):

1. Initial state of mind: **AqRp**

   Nudges rest on a particular assumption about the initial state of mind of the decision-maker. Nudges assume that, by default, the decision-maker's reflective system endorses the 'prudent proposition' (**Rp**) while the automatic system does not (**Aq**). If this were not the case, there would either not be the need for a nudge (such as in the case of **ApRp**), or a nudge would not help the decision-maker to achieve her aims (such as in the case of **AqRq**). This is the first substantial assumption about the decision-maker that nudge make.

2. Intervention: **Aq** $\rightarrow$ **Ap**

   Secondly, nudges intervene in the choice architecture to change the initial state of mind of the decision-maker. More specifically, the intervention aims to change the state of the automatic system such that it now endorses the 'prudent proposition'. That is, nudges change the presentation of the choice such that – via well-known mechanisms established by the behavioral economics and psychological decision theory literature – the decision-maker's automatic system is triggered to support the prudent proposition (such as saving more or eating healthier[1]).

3. Nudge position: **ApRp.**

   Thirdly, nudges aim to permit the decision-maker to consider his or her position after being nudged. Indeed, nudges aim to respect the decision-making of individuals. Nudge-type interventions on the choice architecture are thus designed in a way that a decision-maker can still 'correct' for the nudge (for example, the decision-maker can still opt out of a prudent saving plan[2]).

4. Choice position: **ApRp** leads to choice according to **p**.

   Finally, nudges lead to a behavioral outcome: the decision-maker actually makes a choice based on the nudge position. The result of an ideal and successful nudge is that the decision-maker voluntarily chooses according to the prudent proposition, without the costs of deliberation that are usually associated with such a choice.

2.1.3 Nudge-like interventions

Such a characterization also serves as the basis for spelling out the assumptions behind nudge-like interventions. Here, Heilmann takes Bovens's (2009) taxon-

[1] or sharing trustful information.
[2] Or she can read and share misinformation

omy of intervention on the choice architecture and gives them the dual-process structure so that the comparison with nudges is possible:

	1. Initial State of Mind	2. Intervention	3.Nudge Position	4. Choice Position	System R Respected
Manipulation	AqRq	Aq → Ap	-	Ap(with Rq), such that p is chosen	-
Undetected Nudge	AqRp	Aq → Ap	-	Ap(with Rp), such that p is chosen	-
Classic Nudges	**AqRp**	**Aq → Ap**	**ApRp**	**ApRp, such that p is chosen**	✓
Exception Nudge	AqRq	Aq → Ap	ApRq	ApR?, such that p or q are chosen	✓
Social Benefit Nudge	AqRq	Aq → A?	conflict	open, such that p or q are chosen	✓
Social Advertising	AqRq	Aq → Ap and Rq → Rp	deliberation	open, such that p or q are chosen	✓

Table 2.1: (Heilmann, 2014, p. 86)

I will not go into the details of every type of intervention but will only sketch their main features, that later on will help in elucidating some of the implicit assumptions that Heilmann's framework makes about the dual-process approach. **Manipulations** operate with complete disregard for the initial state of mind of the reflective system and they are only concerned with changing the outcome promoted by the automatic system in a way that does not give a chance to the reflective system to even evaluate the situation. On the other hand, **undetected nudges** do work in situations of initially split minds where the automatic and the reflective system endorse different propositions but the intervention on the automatic system is 'so strong' that the reflective system is bypassed. Thus, even if the proposition chosen is the one that the reflective system would potentially endorse, this type of intervention is closer to manipulation than what nudge proponents would like to be. **Exception nudges** assume a different initial situation since here the person's automatic and reflective system point to the same proposition, with the nudge changing the automatic system's proposed outcome in a way that triggers some cognitive effort in order to make a final choice. The main problem with this type of nudge is that people's exceptional preferences are not respected. **Social benefit nudges** differ minimally from the previous kind. They assume the same initial state of mind but in this case, the social architect apparently targets the automatic system in a defying way (i.e., telling people that they are about to make a mistake). She additionally insists on the fact that the 'prudent' proposition is more socially beneficial and therefore the person should seriously reconsider her position. This may create a serious conflict within people's minds and the social planner could be seen as exerting an unduly influence on decision-makers' autonomy. Finally, **social advertising** entails a similar situation as social benefit nudges, but here the intervention intends to change both the automatic and the reflective systems, giving decision-makers more information and arguments in a more respectful way.

2.1.4 Framework's limitations

From this brief description of the four assumptions behind nudges and nudge-like interventions, I would like to highlight an important aspect of the framework that, in my opinion, obscure the understanding of how exactly these interventions work: how exactly the reflective system operates and when it has an influence on the decision process. I am aware that Heilmann does not intend to provide a "detailed description of what actually happens in nudges" but a *highly idealized* model of what nudges do. Nonetheless, I still find that it is important to highlight them since they will bring to the surface implicit assumptions in Heilmann's take on the dual-process approach that would in turn limit the applicability of the framework to *relatively complex* problems. For example, this could be the case when analyzing situations where people engage in motivated reasoning, which might require a nuanced understanding of the kinds of reflective reasoning involved, as we shall see in the next chapter.

In particular, I argue that the minimal language of the framework and its *plain* reference to 'the reflective system' merges two different conceptions of this system that deserve to be kept apart. To disentangle these two separate notions of the reflective system we can ask the next two questions:

i What exactly is meant by saying that in the initial state of mind, the reflective system endorses the prudent proposition?

ii what is implied by saying that "nudges aim to respect the decision-making of the individuals"?

There are at least, two possible answers in the description of the framework for i). The first of these would postulate that both the automatic and the reflective systems work in parallel once a decision must be made. Thus, the initial state of mind that we should find in narrowly successful nudges, **AqRp**, would simply mean that a problem has put both systems into working so that the automatic system arrives at proposition **q** while the reflective one does at **p**. This *actual* split state of mind would lead the decision-maker to go for **q**. Why is this the case is not entirely clear if we stay within Heilmann's framework description but, in the literature, has been proposed that it is because the automatic system is much faster than the reflective one so while the latter is still working out its response, the decision-maker would already have available one from the automatic system, leading her to go for it (Sloman, 1996; Smith and DeCoster, 2000).

This possible answer for i) would be the one supported by the framework if 1) we understand the present tense of the description of the initial state of mind ("...the decision-maker's reflective system *endorses* the 'prudent proposition'...") as implying that the reflective system is indeed actually working, and 2) we read the nudge position, or 'nudge respect for system R', as implying that the nudge would still grant time for the reflective system to arrive at

a proposition. Under this view of the reflective system, a nudge would work by just changing the proposition endorsed by the automatic system while still giving time to the reflective system for arriving at a proposition and/or assessing whether such response is aligned with the one promoted by the automatic system.

To bring this possibility down to Earth, we can think of the classic cafeteria example where bananas have been placed at eye level with the intention of nudging people to grab one instead of an unhealthy product. If we apply Heilmann's framework to this case, the initial state of mind (in the initial setting where an unhealthy product, I.e., a chocolate bar, is the prominent one) would be one in which seeing the chocolate bar triggers the decision-maker's 'systems' so that the automatic system quickly reaches the conclusion of falling for the bar while the reflective one takes some time (maybe because it is busy analyzing different intuitive responses and/or recalling previous experiences and commitments, e.g., the promise of heating healthily) to conclude that it would be better to forgo the tempting snack. The nudge would consist just of replacing the chocolate bar with a banana so that the automatic system (via an unspecified mechanism) endorses the healthy snack. Still, for the nudge to be narrowly successful, research should have shown that bananas are not an overwhelmingly tempting snack that can bypass the reflective system entirely.

We could also think of a toy example related to misinformation. For example, consider the case of a Twitter user who finds herself sharing a piece of misinformation that put his least favorite politician in a bad light. We could describe his initial state of mind as one in which both systems start working after encountering the tweet, with the automatic system quickly opting for believing/sharing the article, **q** (maybe because the misinformation comes from a source that he trusts or because it is not the first time that he finds such information so that he gives some credit to it), and the reflective system slowly recalling memories and knowledge that would allow her to discredit the information, **p**. Thus, **AqRp**. Since social media is a place that hampers careful reflection, the user would believe/share the information. To solve that, a nudge is implemented so that a truthful article about the same topic is placed next –but in a more salient fashion- to the original piece of misinformation. Given that the original article is still visible, we could claim that the reflective system has a good chance of not being bypassed, thus grating the 'nudge position'.

But there is another possibility in understanding the reflective system's initial position. This alternative view would hold that the reflective system has not been actually triggered by the problem at hand but rather that **Rp** in **AqRp** merely expresses *reflective –deliberative- preferences* that were formed beforehand. That is, in reality, there is no such thing as a split state of mind but a situation in which a problem triggers the decision-maker's automatic system to support a proposition that does not reflect her deliberative preferences. Under this conception of the decision-making process, we could assume that in real-life

decisions the automatic system is not just the quickest in providing a response compared with the reflective one, but it is actually the only system at work. This could be the case for a variety of reasons, for example in scenarios where the automatic response is the evolutionarily most fitted (I.e., face recognition) or where behavior has been learned to the point of automaticity (I.e., playing a musical instrument). Whether –and if so, how- the reflective system can make it into the decision-making process is not clear and Heilmann's framework would be silent about it.

This view also seems to be supported by Heilmann's framework. For example, there are explicit references to *reflective* or *deliberative preferences* when Heilmann points out the practical challenges for the introduction of nudges: "... eliciting the deliberative preferences for the individuals one wants to Nudge in the specific context at hand seems a better way to ensure one makes the right assumption about their initial states of mind" or "The assumption **AqRp** really amounts to assuming a particular kind of divided state of mind: that is, not only do we have to assume the right kind of reflective preferences, but they should also be about propositions in which the automatic system would really promote a different one" (ibid., p. 88). From these quotes, it would not be entirely clear whether the reflective system needs to actively work in the decision-making process or whether its work has already been done, and now the decision is only in the automatic system's hands.

We can further see this ambivalent position if we compare what is said in the nudge and the choice positions. Thus, on the one hand, the ´Nudge position' description states "that nudges aim to respect the decision-making of the individuals. Nudge-type interventions on the choice architecture are thus designed in a way that a decision-maker can still 'correct' for the nudge" (ibid., p.79), meaning that the reflective system needs to be somehow activated if it must override the automatic system's endorsed proposition. On the other hand, we are said in the 'Choice position' that "the result of an ideal and successful nudge is that the decision-maker *voluntarily* (emphasis added) chooses according to the prudent proposition, *without the costs of deliberation* (emphasis added) that are usually associated with such a choice" (ibid., p.80) where while '*voluntarily*' might seem to imply some sort of control exerted by the reflective system, '*without the costs of deliberation*' seems to imply that such a system does not intervene, otherwise there would be some costs associated.

To illustrate, let's reconsider the previous two examples. In the snack case, we would have the customer entering the cafeteria with the deliberative preference of eating healthily but falling for the temptation of grabbing the chocolate bar, thus **AqRp**. This is so because when making the decision, it is only the automatic system the one that is triggered and if the health-concern commitment has not been internalized, it is quite possible that the *natural* craving for sugary and fatty food has the upper hand. That is, for at least a subset of particular decisions about what to eat here and now, the abstract deliberative

preferences of eating healthily would play no role since the decision is made entirely by automatic mechanisms. Thus, the nudge of placing the banana at eye level would just try to fight that natural craving by *playing* with other features of the automatic system that leads it to endorse the banana option.

Similarly, we could think of the Twitter user as someone who claims to place truthfulness above other features (I.e., what his preferred party's position is) when it comes to valuing pieces of information. However, that deliberative preference is trumped by automatic responses displayed in the day-to-day use of Twitter. We could assume that in such a platform (with its fast-paced functioning), decisions are made entirely by the automatic system so unless the deliberative preferences are built into such a system, they have few chances of making an impact. This renders the initial state of mind **AqRp**. Thus, a nudge would consist of a modification of the platform's architecture (for example, by placing a truthful article in a prominent position on the timeline) so that the automatic system is triggered in a way that pays attention to the truthful article.

Note that in none of the examples I have mentioned the ´nudge position', the one that assures that the decision-making process is respected. I have not done so because, under this understanding of the working of the reflective system, it is not clear what respecting the decision-making process amounts to. There are no doubts that respecting the process is not just a matter of aligning the automatic and the reflective systems, manipulations and undetected nudges would do so too. But, since the proposition endorsed by the reflective system is taken for granted, we cannot really know how, if at all, the reflective system can intervene once the nudge is implemented. That is, unless it is specified how the reflective system is triggered, we cannot really assume, for example, that in the case that the misleading article is the chosen one even after the nudge has been implemented, it has been so because the reflective system jumped in and (wrongly) endorsed such an option. In other words, we cannot rule out the possibility that the nudge failed in making the automatic system endorse the truthful article, even though some automatic mechanism was triggered in that direction.

To recap, there are two different understandings of the reflective system in Heilmann's framework, the first poses that such a system works in parallel with the automatic one and the problem with that comes from the fact that we do not know how such a system could override the quicker outcome endorsed by the automatic one. The other notion sees the reflective system as having already formed some preferences but, beyond that, it is not possible to know how the system relates to the automatic one once a problem has cued a response from the latter. Both views could be defended depending on the fragment of the framework's description that we focus on. However, it seems reasonable to hold that both views cannot be true at the same time, either the reflective system works in parallel with the automatic one or it would have formed some preferences

and may intervene in the decision-making process for some unknown reason. Otherwise, in case both views were to be reconciled, it would be necessary to specify under which circumstances each mechanism is triggered.

With this critique, I do not mean to imply that Heilmann's framework is wrongheaded or useless. Nothing further from the truth. In all its simplicity, the four cognitive assumptions spelled out help us in understanding better how nudges work and they have the merit of pointing with great precision to practical (and potentially normative) issues that nudges might struggle in getting right. Indeed, a great deal of its merit is due to the use of a simple dual-process language. What I have tried to show above is that such a minimal language leaves unspecified assumptions regarding the dual-process approach that are important if we are to apply the framework to cognitively complex contexts. My contribution with the following shift in the dual-process language for spelling out the cognitive assumptions can be seen just as a deepening of Heilmann's coherence critique of nudges: they may be indeed very hard to implement.

In the next sections, I will contextualize the dual-process assumptions in Heilmann's by comparing them with the ones which Evans and Stanovich (2013) arrive at after they recapitulate the history of the dual-process approach and defend their view from critiques made to the category as a whole (section 2.2). And secondly, in section 2.3, I will introduce a dual-process model that focuses on the ways in which the reflective processes could be triggered. This will allow me to rebuild Heilmann's four cognitive assumptions with a different dual-process language (section 2.4).

2.2 The development of dual-process theories

In the previous chapter, I introduced a framework that spells out four cognitive assumptions behind nudges and nudge-like interventions. Such assumptions concern the initial state of mind of the person to be nudged, the changes in the mental state that the intervention aims at bringing about, a 'nudge position' whose goal is to guarantee that the person could resist the nudge if she does not agree with it, and a 'choice position' that tell us what the outcome should be. We have also seen that to describe the four assumptions, Heilmann uses a *minimal* dual-process language. According to this language, people's mind is formed by two different systems, one automatic and the other deliberative, each of them operating independently from the other. The need for an intervention comes from situations in which the automatic system promotes a different (and supposedly worse) option than the deliberative system. The nudge would consist of changing the choice architecture so that the automatic system endorses the deliberative option, all that while still *giving the deliberative system a chance* to resist the nudge. After introducing the framework, I remarked on what I see as potential limitations if we want to apply the framework to problems that assume more complex cognitive mechanisms. In a nutshell, I argue that the minimal language contains some ambiguity regarding the exact working of the deliberative system. In particular, it is not clear whether this system has already formed its preferences before the problem arises or whether the initial state only means that it has the potential of doing so. The framework neither specifies if the deliberative system is necessarily active throughout the decision-making or if it just has a supervisory function (and if so, how does it carry out such a task).

Since one of the important features of Heilmann's framework is that the validity of the four assumptions is independent of the appropriateness of the dual-process language, in this section, I will address the limitations and assumptions behind such a language in order to justify replacing it for a more complex one. To contextualize Heilmann's view on the dual-process account, I will follow Evans and Stanovich's (2013) recap of the development of the dual-process framework and conclude that (at least some possible reading of) Heilmann's framework might be misrepresenting the state-of-the-art dual-process theories. In particular, I will argue that contrary to Heilmann's minimal language, the dual-process language candidates for embodying the nudges' cognitive assumptions must (1) refrain from talking about systems and instead uses processes (section 2.2.1), and (2) that the ambiguities that I identified in Heilmann's framework can be traced back to the debate between the *parallel-competitive* and the *default-interventionist* accounts of cognitive architecture (section 2.2.2).

2.2.1 Dual-system and dual-process theories

Since the first dual-process theories made their debut five decades ago, multiple new theories have emerged in different fields like decision theory, psychology

of learning, and social cognition. In the beginning, these theories had no apparent connection between them beyond assuming the existence of two types of thinking, but it was not long after that some theorists tried to unify them. Those attempts of unification signified the transformation of dual-process theories into dual-system theories since their main gist was to assume that there are two systems underlying the variety of either intuitive or deliberative behavior. Unsurprisingly, as the number of dual-system theories also increased, so did the critics of both types of theories (dual-process and dual-system) (i.e., Gigerenzer, 2010; Keren and Schul, 2009; Kurglanski and Gigerenzer, 2011). In contextualizing the history of this debate -and locating the assumptions of Heilmann's model within it-, I will follow Evans and Stanovich (2013), who give a glimpse of this evolution in their defense of the dual-process approach against some of the criticisms raised against it.

The main goal of Evans and Stanovich (2013) is to assess -and refute- five lines of arguments against dual-process theories that were made popular in the last decades: vagueness and a vast number of definitions, supposed features of each system do not always occur together, there are not two differentiate types of processing but a continuum, single-process accounts can cover the same phenomena as the dual-process ones, and there is no strong evidence for dual-process theories. Here, my focus will not be on the specific refutations of each line of criticism but instead on the depiction that the authors make of what they consider the up-to-date understanding of the dual-process view and which serves them as the basis from which to mostly disregard the criticisms.

The first feature that I would like to highlight from what Evans and Stanovich consider a contemporary understanding of the dual-process approach is the commitment to talk about 'processes', dropping the unifying attempts that try to gather all the processes around two discrete systems. According to the authors, this should be so for two kinds of reasons. The first one has to do with the ambiguous use of the term *dual-systems*. While some authors have used it as a way of putting forward a *two-mind hypothesis*, whereby they add an evolutionary component, "[...] suggesting that there are two evolutionarily distinct brain systems responsible for these two types of processing [...] and evolutionarily old and animal-like form of cognition and also a recently evolved and uniquely (or distinctively) human system of thinking." (ibid., p. 224). Other authors have employed the term *dual-systems more casually*, merely as a form of distinguishing two types of processing. On which side Heilmann's position lies is not entirely clear, but I would be inclined to say that, given his intentions of minimal commitment to the dual-process framework, the casual approach may be more coherent.

In any case, regardless of this ambiguity, the term *dual-systems* is misleading since it could be seen as conveying the idea that there are only two systems behind the two types of processing. According to Evans and Stanovich, this is simply not true. There are a variety of cognitive or neural systems underlying

the two types of processing (and, for example, Stanovich (2011) has introduced the term TASS -the autonomous set of systems- to make explicit the idea that there are multiple systems responsible for autonomous processes).

The second, and more important, reason for ditching the *systems terminology* regards what Evans and Stanovich have coined as the *clustering problem*. As was mentioned above, dual-system theories were born after multiple dual-processes theories in different fields started piling up. Each of those dual-process theories attributed a pair of contrasting characteristics to two types of thinking. Thus, for example, we have that some researchers have distinguished between "implicit/explicit, associative/rule-based, impulsive/reflective, automatic/controlled, experiential/rational, nonconscious/conscious, intuitive/reflective, heuristic/analytic, or reflexive/reflective" when defining the two types of processes (Evans and Stanovich, 2013, p. 227). The leap from those individual dual-process theories to the dual-systems ones came after some authors listed those attributes together to link them to two independent systems (i.e., Stanovich, 1999). As was mentioned in section 2.1.2 and can be seen in Table 2.2, this is also the road taken by Heilmann (2014).

	Automatic System	Reflective System
Processes	Fast	Slow
	Parallel	Serial
	Automatic	Controlled
	Effortless	Effortful
	Associative	Rule-governed
	Slow-learning	Flexible
	Emotional	Neutral

Table 2.2: (Heilmann, 2014, p. 78)

The downside of these tables, and what gives rise to the clustering problem, is that some authors, both acolytes, and critics, have misused them as they have regarded them as "strong statements about necessary co-occurring features" (Evans and Stanovich, 2013, p. 228). In other words, some theorists have made every feature a defining characteristic of each system, thus giving the impression that whenever one of the systems is at work, each component of the corresponding set of features could be observed. This, in turn, has made the life of the critics very easy: they just had to show a case where not all the attributes are aligned in order to challenge the appropriateness of the whole dual-process approach (i.e., Krulangski and Gigerenzer, 2011). According to Evans and Stanovich, this is just a straw-man argument. The critics would only be right in the case that every feature is a defining characteristic of the Type of processing in question. Since this is not necessarily the case and a single pair of opposing characteristics would be necessary and sufficient for establishing the two types of processing, this type of criticism could be easily disregarded. For example, for Evans and Stanovich Type 1 processes can be defined as *au-*

tonomous and *not requiring working memory*, while Type 2 processes are those that *require working memory* and *engage in cognitive decoupling and mental simulation*. All the rest of the associated features are just typical correlates.

To conclude, Evans and Stanovich (2013) give two kinds of reasons for stopping the use of *dual-system terminology* and going back to *Type 1 and Type 2*. The first refers to the ambiguity of the term *dual system*, while some authors use it casually, others have made stronger claims, liking it to the *two-mind hypothesis*. Moreover, the term is misleading since there are more than two neural systems responsible for both intuitive and deliberative behavior. Secondly, dual-system theories tend to entail the *clustering problem*, by which each system is (unnecessarily) defined by a set of co-occurring features. For these reasons, Evans and Stanovich propose to recover the *Type 1/Type 2 processes terminology*, whereby the formers refer to autonomous processes that do not require working memory, and the latter not only requires working memory but also engages in cognitive decoupling and mental simulation. In the previous section, we saw that Heilmann (2014) makes constant use of *dual system terminology*. And while there seem to be no traces of the strongest version of it (that which links it to the two-minds hypothesis), it is beyond doubt that his dual-process language makes some commitment to the existence of two systems. One could argue that we should differentiate between the foundation of Heilmann's framework and the actual framework so that while we might question the validity of the former, we could still make use of a version of the latter that includes minor modifications such as **AqRp → T1qT2p**. In sections 2.2.2 and 2.4, I will argue that changing the theoretical foundation of the framework does not only affect the language in such a way but also what is implied by it. In other words, it is not only a matter of changing *systems* by *processes* but about better understanding how the processes work and interact with each other.

2.2.2 Parallel-competitive and Default-interventionist accounts of cognitive architecture

In the previous subsection, I have focused on defining and characterizing the two types of cognitive processes and concluded that, according to the latest developments in dual-processes theories, it is more accurate to stick with the *Type 1/Type 2 processes terminology* instead of insisting on the dual-system endeavor. In the current subsection, I will introduce the two main accounts that try to explain how the two types of processes interact: Evans and Stanovich's *default-interventionist* account of Type 1-Type 2 processes interaction, which rivals the so-called *parallel-competitive* account (Evans, 2009). Comparing both accounts and looking at their limitations will help us to contextualize the ambiguous assumptions of Heilmann's framework regarding the working of the two types of processes. Additionally, this subsection will set the ground for section 2.4 where I introduce a model of human reasoning (Pennycook et al., 2015) that tries to synthesize both accounts, and that will serve us to clarify how

Heilmann's framework should be modified to reflect the state-of-the-art view on how the two types of processing work and interact.

The ambiguities around how exactly the two types of processing work and interact in Heilmann's framework were the main limitations that I found in it if we were to apply the framework in contexts where such interactions matter. In particular, we saw that it was not clear whether decisions are taken primarily by Type 1 processes, sometimes going against preferences previously (before the need for a decision to be taken) formed by Type 2 processes, or whether both types of processes are triggered whenever a decision must be made. As shall be clear by the end of the section, such ambiguities are not something original in Heilmann's framework but a reflection of a dichotomy in the field.

Let's start considering the *parallel-competitive* account (I.e., Sloman, 1996; Smith and DeCoster, 2000). Leaving aside the differences in terminology between particular proposals, we could say that the mind idea behind this account is that both types of processes operate simultaneously from the very beginning, but they go on independently so that each of them proposes a judgment. Moreover, it is important to note that only one type of process will impose its say. While the final decision might be context-dependent, it is widely accepted that in a great number of cases, Type 1 responses prevail thanks to their faster pace. In Sloman's words (2002, p. 391, *italics* added):

> Both systems seem to try, at least most of the time to generate a response. The rule-based system (*Type 2 processes*) can suppress the response of the associative system (*Type 1 processes*) in the sense that it can overrule it. However, the associative system always has its opinion hear and, because of its speed and efficiency, often precedes and neutralizes the rule-based response.

It is easy to realize that this account resonates with the first of the possible readings of Heilmann's dual-process language that we introduced above. However, some researchers have pointed out some limitations of this account. An important question is to figure out how Type 2 processes could overrule Type 1's outputs. Here, it is important to note that parallel-competitive accounts reserve a *monitoring* feature for Type 2 processes. Thus, in the cases in which the two types of processes reach conflicting responses, Type 2 processes have the potential to detect the conflict and *conduct* further analysis. There is of course the possibility that no conflict is detected due to a malfunction of the Type 2 processes or a lack of time if the Type 1 outcome is quickly enacted. Both the gap between the responses and the lack of clearly distinguishing between the firstly triggered Type 2 process and its monitoring function are the two main criticisms that parallel-competitive accounts face (Evans and Stanovich, 2013, p. 237; Pennycook et al., 2015, p. 37).

A more relevant problem with this account according to Evans and Stanovich is that "Type 2 processing requires extremely limited and precious working

memory resources [...] these must be selectively allocated to the most important task at hand" (ibid.). In other words, it seems unlikely that Type 2 processes would be triggered every time a decision or judgment needs to be taken given the vast number of cognitive resources that these types of processes require. Bringing back the examples from the previous section, according to the critics of the parallel-competitive accounts, it would not be realistic to presuppose that every time someone is searching for a snack in the supermarket, she not only has automatic Type 1 responses for every potential snack but also need to engage Type 2 processes to figure out her evaluation of them. Similarly, it would seem far from reality to assume that a Twitter user evaluates *every* tweet through Type 2 processes. Were this true and given the hectic behavior in that social platform, the high demand for cognitive resources would tire out the user after a very short time. The apparent implausibility of this account led some researchers to put forward an alternative view, the *default-interventionist* account.

Default-interventionist accounts take a different route in setting up people's cognitive architecture. According to this approach, there is no need for a join trigger of Type 1 and Type 2 processes. Instead, most behavior would be governed by Type 1 processes and only under specific circumstances Type 2 processes would jump in to take control of the situation. Let me quote Evans and Stanovich's words in length to have a detailed grasp of this account (ibid., *italics* added):

> In general, we believe that intuitive answers are often prompted rapidly and with little effort when people are confronted with novel problems. Where they lack relevant experience, however, these answers may be inappropriate and fail to meet the goals set. Thus, a key concept in this kind of dual-process theory is that of intervention with reflective (Type 2) reasoning on the default (Type 1) intuition. Often, humans act as cognitive misers (an old theme in cognitive and social psychology) by engaging in attribute substitution—the substitution of an easy-to-evaluate characteristic for a harder one, even I the easier one is less accurate (Kahneman and Frederick, 2002). However, when the decision matters, being a cognitive miser may lead us astray [...] Default interventionism allows that most of our behavior is controlled by Type 1 processes running in the background. Thus, most behavior will accord with defaults, and intervention will occur only when difficulty, novelty, and motivation combine to command the resources of working memory (*Type 2 processes*).

This account clearly dispels the *cognitive resources* problem of the parallel-competitive account by reducing the need for Type 2 processes. Only in exceptional cases that involve some difficulty, novelty, or particular motivation, extra resources are required to make a decision. For the rest of the cases, which are

the majority, *highly energy-efficient* Type 1 processes are in control. But as was pointed out in the quote this does not mean that efficiency equals rightness in the decision. People might behave as cognitive misers or they might face a hostile environment where other agents can "discern the simple cues that are triggering Type 1 processing [and] start to arrange the cues for their own advantage (e.g., advertisements or the deliberate design of supermarket floor space to maximize revenue)" (ibid., p. 229). Another advantage of this account compared to the parallel-competitive one is that since the two types of processes do not work simultaneously, default-interventionist accounts also "wave away" the problem with the gap in time response between types of processes.

Applying this approach to the snack example, we would have that in *normal circumstances* where the recurring decision about what snack to grab does not entail any difficulty and the eater is not *exceptionally* motivated for thinking carefully about what to eat, Type 1 processes would be in control of the situation. Thus, whatever particular craving has been encoded into Type 1 processes will be the typical response. On the other hand, when those factors do not align and, for example, the preferred unhealthy option is not available or the person has gone for lunch with her nutritionist friend, then Type 2 processes would have the chance to intervene in the decision-making process and make her reevaluate the situation. As we saw above, this was exactly the second of the possible readings of Heilmann's dual-process language that we consider in the previous section.

However, a criticism that the default-interventionist accounts have not been able to escape from is the one that problematizes the fact that "Type 2 processes are themselves responsible for the instantiation of Type 2 processing" (Pennycook et al., 2015, p. 36). Similar to what we saw above about the conflict monitoring feature in parallel-competitive accounts, neither the default-interventionist accounts clearly explain "what leads someone to engage deliberate and effortful reasoning instead of more intuitive and automatic cognitive processes" (ibid., p. 35). That is, beyond simple mentions of the difficulty, the novelty of the decision, and the motivation of the decision-maker more is needed about the factors that trigger Type 2 processes. In the next section, I will introduce Pennycook, Fugelsang, and Koehler's model of analytic engagement, which tries to overcome this problem by originally incorporating elements of both parallel-competitive and default-interventionist accounts in a three-stage unifying model that focuses on the bottom-up factors that lead people to engage Type 2 processing.

To sum up, in this subsection we have analyzed the two main types of cognitive architectures proposed in the dual-process literature: the parallel-competitive account and the default-interventionist one. While the former assumes that both Type 1 and Type 2 processes are triggered for every decision and reserves a conflict monitoring function for Type 2 processes, the latter

proposes that Type 1 processes are in charge most of the time and Type 2 processes only intervene under particular circumstances. Additionally, we have also pointed out the limitations faced by each of the accounts. Whereas the parallel-competitive account struggles in explaining (1) how the Type 2 processes could have a say given the fast working of the Type 1 processes and (2) how could be sustainable that high energy demanding Type 2 processes are at work all the time; both the parallel-competitive and the default-interventionist accounts seem to assume that Type 2 processes are caused by themselves. We have also seen that the two possible readings of Heilmann's dual-process language fit each of the accounts. In the next section, I will introduce an account that tries to unify both approaches in a three-stage model of analytical engagement (Pennycook et al., 2015), and that will serve as the basis with which to modify Heilmann's dual-process language in such a way that avoid both the ambiguities of the original language and the drawbacks of the two accounts.

2.3 How people think: Pennycook, Fugelsang, and Koehler's (2015) model of human reasoning

In this first part of the thesis, we set the goal of devising a framework that specifies the cognitive mechanisms behind the introduction of nudges. I started by introducing Heilmann's framework (Heilmann, 2014), which aims at doing exactly that by stating four cognitive assumptions behind nudges and nudge-like intervention using a dual-process language. Next, I pointed out what I see as the main drawbacks of the framework were we to apply it in relatively more-cognitively-complex contexts: the ambiguities regarding the working and interactions between the automatic and deliberative systems. To be able to clarify such ambiguities, I looked at a more up-to-date dual-process account (Evans and Stanovich, 2013). From there, I concluded that since dual-system theories seem to be out of fashion, our framework should stop referring to systems and instead use the *Type 1/Type 2 terminology*. More importantly, we could see that the two possible readings of Heilmann's dual-process account reflect a long-standing debate in the field, that between parallel-competitive and default-interventionist accounts of cognitive architecture.

In this section, the goal is to introduce a model of cognitive architecture (Pennycook et al., 2015) that tries to settle the debate, and therefore also neutralize the ambiguities with Heilmann's framework, by incorporating both accounts in a single model. I will do so by firstly, in section 2.3.1, looking at how Pennycook, Fugelsang, and Koehler place their model within the parallel-competitive/default-interventionism debate. Next, section 2.3.2 will carefully detail Pennycook and colleagues' three-stage model, covering each stage in sections 2.3.2.1, 2.3.2.2, and 2.3.2.3 respectively. Finally, section 2.3.3 concludes. By the end of this section, we will finally be in a position to, in section 2.4, devise a new dual-process language with which to spell out the four cognitive assumptions behind nudges identified by Heilmann (2014).

2.3.1 The model and its place in the parallel-competitive / default-interventionism debate

Were Type 1 and Type 2 processes all firing up when our Twitter user encountered (bluntly false) news criticizing a disliked politician? Or maybe she just decided to share the tweet after *fluent* Type 1 processing of the information, even though that would violate her previously formed Type 2 preferences? In other words, was the biased behavior (sharing misinformation) caused by the gap between Type 1 and Type 2 processes, or due to a lack of Type 2 processing altogether? As we saw in the first section of this part *when* Type 2 processes (deliberative system at that point of the argumentation) are triggered and *how* they relate to Type 1 processes (automatic system) were not entirely clear in

Heilmann's framework. Such ambiguity could hinder the understanding of how to implement nudges to counteract some biased behavior.

According to Pennycook, Fugelsang, and Koehler (2015), a clear understanding of what triggers Type 2 processes is also something lacking in the two main accounts of cognitive architecture, the parallel-competitive, and the default-interventionist. In their own words (ibid., p. 36): "[N]either of the major groups of dual-process theories adequately explains important aspects of cognitive architecture because both assume that Type 2 processing is effectively caused by *itself*. This is a problem of particular importance because the utility and explanatory value of dual-process theories are thought to depend, at least partially, on our understanding of the *sources* of analytic reasoning (Evans, 2009; Stanovich, 2009; Thompson, 2009)". This unsatisfactory situation led Pennycook and colleagues to formulate a model that breaks down the decision-making process into three stages in order to incorporate the strongest features of each account. Let's see what exactly the model consists of and then, in the final section, evaluate how it could impact Heilmann's model.

Before moving on, it is important to note here that, in their model, Pennycook and colleagues only consider *bottom-up* sources of Type 2 engagement, contrary to *top-down* sources like, for example, direct instructions of thinking analytically (Daniel and Klaczynski, 2006; Evans, Newstead, Allen, and Pollard, 1994), or individual differences in thinking dispositions (Stanovich and West, 2000). This is very relevant for the purpose of this thesis since nudge interventions aim at mainly playing with Type 1 processes while safeguarding some capacity for Type 2 processes to step in the decision-making. That is, nudges are not instructions to reconsider the problem at hand, nor do they focus on training people's thinking dispositions.

We saw above that parallel-competitive accounts of cognitive architecture have faced significant criticisms in the past, being the most relevant of them the fact that it seems unlikely that the brain could afford the triggering of Type 2 processes all the time. However, Pennycook and colleagues (2015) present compelling evidence showing that this could be actually the case in some circumstances. They seem to conclude that after considering a well-researched problem in the decision-making literature, the base-rate problem. In this kind of problem, there are two types of information, one based on the base-rate probability of something being the case, the other entails some stereotypes pointing at a sometimes-alternative response. Research about the base-rate problem has found that people tend to go for stereotypical responses because they can process them more fluently. But, importantly for the parallel-competitive accounts, Pennycook and colleagues also mention some studies concluding that even when people go for the stereotypical response, they can be somehow aware that there is a conflict between the base-rate and the stereotype (De Neys, Cromheeke, and Osman, 2011; De Neys and Franssens, 2009; De Neys, Vartanian and Goel, 2008). This is shown by the increase in response time in problems where the

base-rate and the stereotype responses point at different conclusions compared to when there are aligned, even if in both cases the person would go for the stereotypical response. Were the critics of the parallel-competitive accounts right, then we should not observe any kind of increase in the response time: whenever there is a chance for Type 1 processing of information, as is the case here following the stereotype, there is no need to engage any kind of more consuming processing. Otherwise, the brain would be wasting valued energy in triggering Type 2 processes. From the previous findings, Pennycook and colleagues conclude that their model of analytical engagement should include a conflict monitoring stage as a source of Type 2 processing.

However, monitoring conflicts is not the only kind of Type 2 processing. As default-interventionist accounts prominently stress, cognitive decoupling, and overriding Type 1 responses are very important instances of Type 2 processing. According to those accounts, we should not assume that Type 2 processing enters the picture from the very beginning. They do so only in cases in which they need to override a Type 1 response. This type of analytical engagement has been also found in the base-rate problem. In particular, it has been observed that when people go for the base-rate response in incongruent problems, they take more time to respond than in congruent problems (where both base-rate and stereotypical responses point in the same direction). That extra time is thought to be spent in decoupling the initial Type 1 responses and then overriding the one that would be intuitively acted upon (De Neys and Franssens, 2009; De Neys, Vartanian and Goel, 2008). From this then, Pennycook and colleagues take for their model cognitive decoupling as a source of Type 2 processing to be differentiated from the conflict monitoring one.

2.3.2 The actual model

Let's move now to the description of Pennycook, Fugelsang, and Koehler's model to see how they integrated both sources of Type 2 processing into a single model (Figure 2.1 below shows what the model looks like). As was mentioned above, the way in which they did so, and what constitutes the main contribution of the model, is by breaking up the decision process into the following three stages:

2.3.2.1 Stage 1

As we can see in the top part of Figure 2.1, in the first stage of the decision process we have the problem cueing different intuitive, Type 1 responses (IR_1, IR_2, IR_n). There are four important characteristics of this stage that we can infer from Pennycook et al.'s description. Firstly, each of the Type 1 responses is prompted by different features of the problem. Secondly, those Type 1 processes operate in *parallel*. Thirdly, it may be the case that such responses point at competing directions. And fourthly, those responses may diverge on the speed and fluency in which they are processed.

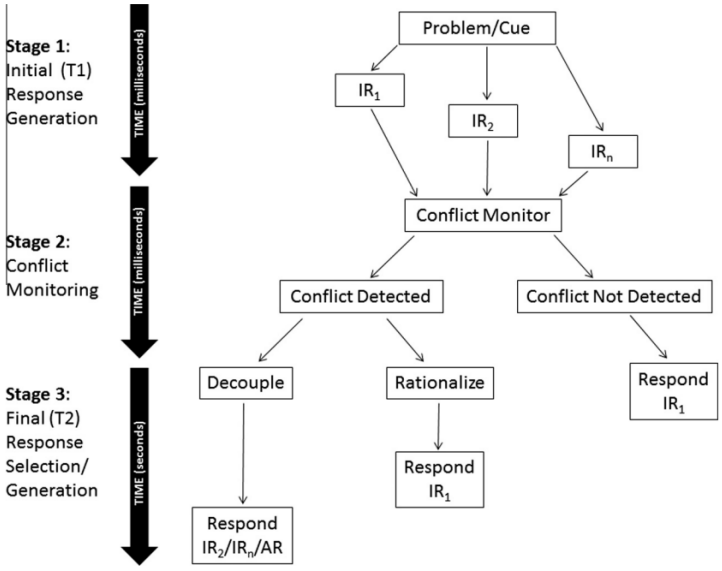

Fig. 1. Three-stage dual-process model of analytic engagement. T1 = Type 1 "intuitive" processing. T2 = Type 2 "analytic" processing. IR = initial response. IR's are numbered to reflect alternative speeds of generation. IR_1 is the most salient and fluent possible response. IR_n refers to the possibility of multiple, potentially competing, initial responses. AR = alternative response. IR_n refers to the possibility of an alternative response that is grounded in an initial response.

Figure 2.1: (Pennycook et al., 2015, p. 39)

Bringing back the (purely speculative) misinformation example, we would have that just as our Twitter user glances at the tweet covering false information about a disliked politician, some intuitive responses are cued. For example, we could establish that IR1 entails acceptance and willingness to retweet the information since it perfectly aligns with her political desires or expectations. IR_2 would be cued slightly after but it would also support retweeting the information as this intuitive response follows the fluent processing of the information due to the fact that it is not the first time that she encounters such a news. We could go on and also assume that a third intuitive response, IR_3, has a chance of making it through. In this case, IR_3 would suggest the user discredit the information given her increasing distrust of any extremist political view.

According to the characteristics mentioned above, we could postulate that different features of the tweet give rise to the intuitive response. Thus, IR_1 is cued by the statement made in the tweet; IR2 after quickly recognizing the image that accompanies the news; and IR_3 by the fact that the tweet popped up in the timeline some days after she watched a YouTube video about how to detect misinformation. Moreover, we can also assume that each IR is processed independently of the others, and that, as was specified above, they point at different conclusions and are processed at different speeds and fluency.

From this first stage of the decision-making process, we can already infer that the model put forward by Pennycook and colleagues at least partly diverges from a classic parallel-competitive account. While Pennycook et al.'s model does propose parallel processing, this regards only Type 1 processes. According to the model, there is no sight of Type 2 processing in the first stage. As we shall see, they will make it into the decision-making only in the second and third stages. In Pennycook and colleagues' own words (ibid., p. 66, *italics* in the original): "The three-stage model is consistent with default-interventionist models and may even be considered a default-interventionist model itself because Type 2 processing does not occur until *after* Type 1 processes output a response. The primary difference between the three-stage model and traditional default-interventionist models (e.g., Evans, 2007, 2010a, 2010b) is that the former is interested in the *causes* of analytic intervention whereas the latter is typically focused on determining the common defaults that undermine reasoning (e.g., prior beliefs) and the problem factors that require intervention to enter into reasoning (e.g., logical validity)". However, this is not to say that "it is [im]possible for a factor traditionally associated with analytic processing such as base-rate probabilities or logical validity to be the source of a Type 1 output (see Handley and Trippas, 2015) – and, in fact, for some individuals, factors such as logic may be more intuitive than factors such as belief (that is, logic cues IR1 and belief cues IR2)" (ibid.). In our misinformation example, that would be the case if, for example, our Twitter user had trained her skills for detecting misinformation up to an intuitive level.

2.3.2.2 Stage 2

The first source of Type 2 processing that we find in Pennycook and colleagues' model is conflict monitoring. As can be seen in Figure 2.1, the second stage of the model is solely constituted by the monitoring of possible conflicts between the intuitive responses cued in the first stage. There are only two possible options, either some conflict is detected or is not. If there is not, this could be either because there is in fact no conflict between the intuitive responses or due to a malfunction of the conflict monitoring feature. Whichever is the case, no Type 2 processing is carried out and the quickest IR would make it to the third stage "where it is accepted with cursory analytic (Type 2) analysis" (ibid., p. 39). It is important to note that the emergence of biases according to the default-interventionist accounts would be explained by the lack of engagement of Type 2 processing in this stage: Type 1 processing is the only one in charge of the decision. On the other hand, if a conflict is detected, then more Type 2 processing will be performed in the final stage.

Before moving on to the third stage, a clarification is due. As Evans and Stanovich (2013, p. 229) point out, it has been a recurring mistake in the literature to consider that "Type 1 processes (intuitive, heuristic) are responsible for all bad thinking and that Type 2 processes (reflective, analytic) necessarily lead to correct responses". There are indeed occasions in which Type 1 processes would yield an optimal or good-enough outcome. It all depends on whether the environment in which the decision is made is benign or not. According to Evans and Stanovich, what renders an environment benign is the presence of "useful cues that, via practice, have been well practiced by Type 1 mechanisms" (ibid.), and the lack of other agents that would sort out the cues for their own benefit.

There is no mystery when considering what this second stage looks like for the misinformation example. If the three intuitive responses are cued but no conflict is detected, then we must assume that a failure with the conflict detection has occurred, and no Type 2 processing has been triggered. This could be the case if, for example, IR_3 is prompted much slower than the first two IR, between which there are indeed no conflicts. Note that we would have the same outcome also in the case where only IR1 and IR2 are cued and therefore no conflict can be detected. However, for the misinformation case, it would not be as straightforward as for the base-rate problem to conclude that the response is biased since the normative response does not necessarily come from the laws of probability. But more on that below.

2.3.2.3 Stage 3

Lastly, in the third stage of the decision-making process, there are three possible paths leading to the final response. The first path has already been mentioned, the one that follows the lack of detected conflict in the second stage, and which ends with the bringing about of IR1. But there are two extra paths, each of them entailing a different form of Type 2 processing. On one hand, we have

what is known in the literature as *rationalization* and which consists of "the reasoner [focusing] on justifying or elaborating the first initial response (IR_1) without seriously considering the Type 1 output that was cued by the stimulus, but that did not come to mind as quickly and fluently (IR_2) as the first initial response (IR_1). This leads to a response in line with what would typically be considered bias (i.e., one's strongest intuition, which will often be personally relevant), but that has been bolstered by analytic reasoning (and "effortful' belief-based response; see Handley and Trippas, 2015)" (Pennycook et al., 2015, p. 40). On the other hand, individuals can engage in a form of Type 2 processing called *cognitive decoupling* (Stanovich, 2004; 2009) in which time is spent more carefully analyzing the possible responses and overriding the initial one, IR_1. Pennycook and colleagues identify three possible outcomes from the decoupling process: "(1) IR_1 is suppressed in lieu of IR_2 which, upon reflection, emerges as a stronger alternative, (2) IR1 is suppressed in lieu of some other initial response (IR_n), and (3) an alternative response (AR) is generated that represents a novel amalgamation of initial responses" (ibid.).

So, what would be the possible behavioral outcomes of our Twitter user? As we said above, if no conflict is detected (either because only IR1 and IR2 are cued or because the conflict detection feature fails), then the user would share/believe the misinformation contained in the tweet. However, if a conflict is detected, two options open up. The user can engage in rationalizing IR_1 (and maybe also IR_2 in this case) at the expense of a fair consideration of IR_3. That is, the user would disregard her vague intuition for detecting misinformation because she is focused on verifying the credibility of the (false) information (maybe due to the force of the combination of IR1 and IR2, or her recalling of memories that put the politician in a bad light). Alternatively, she could use her extra time decoupling the intuitive responses, and after careful consideration where she weighs all the options, an action is performed. As was pointed out above, it cannot be guaranteed that after the decoupling process, she will decide to discredit the information and not forward the tweet. For that to be the case, we would need to assume that she is a perfect reasoner and has the time and resources to go over her memories and relevant information in order to reveal the falsity of the news. But, as we shall see in the second part of the thesis, it might be perfectly rational for her to, for example, weigh more heavily information that reinforces her political identity (she might get more out of acceptance from a group than out of the pursuing of the *truth*). It could be also the case that after *logically* combining the new information with her prior beliefs, she decides to retweet the news. What this implies is that the difference between decoupling and rationalization may not be as clear as the model seems to suggest and that they lie on a continuum.

2.3.3 Concluding remarks

With this, we have arrived at the end of the section. Here, I have introduced Pennycook and colleagues' model of analytical engagement where they break

down the decision-making process into three stages to differentiate between two different sources of Type 2 processing: conflict monitoring and cognitive decoupling. According to Pennycook, Fugelsang, and Koehler, each of these sources relate to the two main accounts of dual-process cognitive architecture: the parallel-competitive and the default interventionist accounts respectively. The integration of the two accounts into a single model, even if it is more aligned with the default-interventionist approach, is a valuable contribution to the field since it sheds new light on both the ways in which Type 1, and Type 2 interact and the bottom-up sources that lead to Type 2 engagement. Moreover, the model also allows (1) to carefully specify the different ways in which the decision-making process can *go wrong* (failing to engage Type 2 processes and failing to override the initial wrong intuitive response after having detected a conflict between different intuitions), leading up to biased behavior; and (2) to help to devise interventions that counteract such mistakes. All these reasons also make Pennycook and colleagues' model useful for the purpose of this thesis. Not only does it avoid the *system terminology* and stick to the Type 1/Type 2 one as was suggested by Evans and Stanovich (2013), but it makes explicit the cognitive assumptions that were ambiguous in Heilmann's model: it discards the initial parallel working of Type 1 and Type 2 processes, which was one of the possible readings of Heilmann's framework and specifies under which circumstances Type 2 processing is triggered. In the final section of this first part of the thesis, I will spell out the four cognitive assumptions behind nudges and nudge-like interventions identified by Heilmann with a new dual-process language derived from Pennycook and colleagues' model.

2.4 A new proposal for understanding Nudges

In the last two sections, I have both contextualized the understanding of the dual-process theories that can be inferred from Heilmann's (2014) framework and introduced a dual-process model (Pennycook et al., 2015) that seems to avoid the problems found in the previous one. The intended goal of this last section is to arrive at a framework that specifies the same four assumptions as in Heilmann's but with a new language based on the contextualization made by following Evans and Stanovich and Pennycook and colleagues' model. Thus, the road ahead is as follows: while 2.4.1 introduces the new language, which switches the *system terminology* by *Type 1/Type 2 terminology*; the next four subsections, 2.4.2.1, 2.4.2.2, 2.4.2.3, and 2.4.2.4, focus on reformulating the assumptions behind nudges: the initial state of mind, the changes intended by the intervention, the nudge position, and the choice position, respectively. Section 2.4.3 briefly explores how some nudge-like interventions would look when formulated in the new language. Lastly, section 2.4.4 concludes.

2.4.1 New language

A crucial part of Heilmann's (2014) argument is to differentiate between the assumptions behind nudges and nudge-like interventions and the dual-process language used to spell them out as a way of safeguarding the former from the criticisms against the dual-process theories. In any case, we also saw that Heilmann keeps the language as minimal as possible –maybe with the intention of not tying the framework to any particular dual-process theory. Thus, the language merely refers to the outcomes supported by the *automatic* and the *reflective systems*, and to the conjunction of both as the decision-maker (e.g., **AqRp** meaning that at the given time, the decision-maker's automatic system promotes q while the reflective system does p). However, after Evans and Stanovich's reevaluation of the evolution of dual-process and dual-system theories which concludes that the Type 1/Type 2 terminology is the most accurate, a revision of Heilmann's minimal language is also due:

- I will use **T1** for the Type 1 processing, **T2D** for the decoupling Type 2 processing, and **T2R** for the rationalization Type 2 processing, while not setting for any particular formula to refer to the decision-maker. Contrary to Heilmann's language, which writes **AR** for the decision-maker, our language cannot just simply translate as **T1T2** given the possibility of mental states without Type 2 processing.

- I follow Heilmann's formulation and keep **p** and **q** for referring to what we could take as the 'prudent proposition' and its negation, respectively. However, in the current framework, additional letters for alternative propositions will take an important role.

- Similar to Heilmann's language, to indicate which propositions are endorsed by each type of processing we can write, for example, **T1p**, **T1q**,

T2Rp, or **T2Dr**. However, the novelty in our case comes from the fact that more than one proposition can be triggered as Type 1 processes at the same time. Thus, for example, we would use **T1pqr** for a mental state in which a problem cues the intuitive responses **p**, **q**, and **r** (as we shall see below, the possibility of having contradictory intuitions will be a key element in the new framework since it is the way of expressing the situations that might call for T2 processes). Note as well that the propositions are written down in chronological order: **p** proceeds **q**, which is followed by **r**.

- In the current language, bold question marks will also denote situations in which there is uncertainty regarding the proposition promoted by the type of processing in question. Note again that the fact that more than one proposition could be endorsed by Type 1 processing opens the door for situations like the following, **T1p?r**, **T1qp?**, and **T1?r**, to name a few.

- Lastly, I will also mark with an arrow a change in the endorsed propositions. Thus, for example, **T1qr** $\rightarrow$ **T1pqr** means that a faster intuitive response **p** has been added to the initial **q** and **r**.

Before moving to the reevaluation of the cognitive assumptions behind nudges, it is worth remarking on the same caveat as Heilmann (ibid., p. 79): "This terminology does not play the role of a formal language: all it introduces are a few abbreviations and shorthand expressions that will make it much easier in the following to use the framework of dual process in describing Nudges". The only difference here is, of course, that the framework of dual process in the present case includes Evans and Stanovich's (2013) contextualization and Pennycook, Fugelsang, and Koehler's three-stage model.

2.4.2 The new framework

2.4.2.1 New initial state of mind: T1q?

The basic idea behind nudges is to prevent people from *making mistakes after following their guts* and assist people in choosing what they would do if they deliberated about it. Thus, we could establish that a 'mistake after following the guts' is the initial state of mind. Such a state was spelled out in Heilmann's framework as **AqRp**, meaning that the automatic system supports the 'no-prudent' proposition while the deliberative system does support the 'prudent' one. With other initial states of mind -when both systems are aligned- there is no need for nudges since they would work against people's deliberative decisions (Heilmann, ibid.).

However, the initial situation must be different if we analyze it using Pennycook and colleagues' three-stage model. As we saw, according to such a model, the initial stage is one in which a problem cues one or more intuitive (Type 1

processing) responses. Type 2 processing will only if at all, enter the picture later on. If this is correct, then it might not be very accurate to say that Type 2 processes support any kind of proposition in the initial stage. Of course, one could say that by, for example, **T1qrT2p** is only meant that the decision-maker's Type 2 processing would arrive at proposition **p** if it were triggered, or so has been seen in the past. Others could argue that **T2p** in the initial state of mind only implies that the decision-maker has expressed that preference when she was questioned about it, independently of previous behavior. But against these possibilities, I would argue, following Pennycook, Fugelsang, and Koehler, that (1) regardless of what has been observed in previous behavior and of stated preferences, there is no activation of Type 2 processing in the initial stage. Full stop. Previous behavior or stated preferences could only make an impact in the initial stage if they were internalized as Type 1 processes (e.g., when someone masters a skill to the point of automatization), and even so, their force is limited since the decision-maker might not pay attention to them and go for a quicker intuitive response. Additionally, I would also argue that (2) extrapolating past behavior or previously stated preferences may run the risk of oversimplifying the functioning of Type 2 processes. Under those views, Type 2 processing seems to be understood as a unitary process, but we pointed out above that there are, at least, two kinds of Type 2 processing, cognitive decoupling, and rationalization. Since those two ways of Type 2 processes may yield different outcomes, failing to specify which one was responsible for the past behavior, or the stated preferences seems like an important limitation if the whole point was the align Type 1 processes with what the decision-maker *really* wants.

The previous considerations lead me to propose **T1q?** as the initial state of mind. What such a state shows is that the only thing we could be sure about is that the problem at hand cues in the decision-maker the intuitive response **q** with such speed and strength that makes her go for that behavior without further consideration. The bold question mark merely suggests that there may well be other intuitive propositions being prompted by the problem but of which we cannot necessarily be sure. It could be the case that the problem cues other Type 1 responses pointing in different directions but what is important to note is that they have no real impact on the decision. This could be because they were not cued fast or strongly enough to create a conflict that could be detected via Type 2 processing. What this implies is that the exact initial state of mind depends on the case at hand, being possible to have, for example, states such as **T1qsp** or **T1qp**, but where ultimately the propositions **s** and **p**, and **p** respectively have no impact on the final decision, **q**.

2.4.2.2 Intervention: T1q? → T1pq?

Similar to Heilmann, we will also understand nudges as interventions that aim at modifying the choice architecture in order to change the decision-maker's initial state of mind. The difference with Heilmann's position is that here we will not talk about "chang[ing] the state of the automatic system such that it

now endorses the 'prudent proposition'" (ibid.), but about cueing an intuitive response that endorses such proposition, **p**, and that it does so in such a way that **p** is now the quickest/strongest response to be triggered. It is also important to remark that the intervention makes **p** the most salient response without completely overshading the previous intuitive responses, **q?**. In other words, the nudge just *adds* a new-and-quicker intuitive response or *moves to the forefront* an existing one, while retaining the original ones (almost needleless to say that the intervention might not succeed and fail to either add a new response or move an existing one to the forefront). Why this is important will be clear after we reevaluate the next assumption, the nudge position.

2.4.2.3 Nudge position: T1pq?T2

One of the most important features of nudges, and what differentiates them from manipulations, is that they still give room for the decision-maker to reconsider her position. Following Pennycook and colleagues' model, we could say that a nudge does not intend to trigger an intuitive response so quickly or strongly that bypasses any kind of Type 2 processing – by making it impossible for them to detect a conflict between different intuitive responses. That is why in the previous assumption I was particularly careful with keeping the original intuitive responses and not making the 'prudent proposition', **p**, the only one cued by the problem after the nudge. The availability of more than one Type 1 response is what makes possible the further engagement of Type 2 processing – from a bottom-up perspective. If there is no conflict to detect because there is only one intuitive response, then, naturally, such a response would be the one acted upon. But in that case, the intervention would be a manipulation and not a nudge.

An important difference concerning Heilmann's position is that I do not presuppose that, in the nudge position, we should establish which proposition is ultimately endorsed by Type 2 processes. That is why I write **T1pq?T2** as the nudge position, without specifying anything after **T2**. Note further that I do not even say which kind of Type 2 processing, either cognitive decoupling or rationalization, needs to be triggered by the intervention. What is relevant about the nudge position is to guarantee that a conflict is detected between the initial intuitive responses and that that will lead to further Type 2 processing, with which the decision-maker could still 'correct' for the nudge.

2.4.2.4 Choice position: T1pq?T2Rp leads to choice according to p

And here comes probably the most important step in my proposal. I shall argue that nudges are a kind of intervention that is based on the *rationalization* of the intuitive response put forward by the nudge. That is, more Type 2 processing is required following the detection of a conflict in the nudge position, but such Type 2 processing must be of the *rationalization* kind and not a *cognitive decoupling* of the intuitive responses. In my view, this is the only way of guaranteeing

that "[t]he result of an ideal and successful Nudge is that the decision-maker *voluntarily* chooses according to the prudent proposition, *without the costs of deliberation* that are usually associated with such a choice" (ibid., *emphasis* added). The fact that it is a *voluntary choice* comes from the active participation of Type 2 processes; the final choice is not the result of unchecked Type 1 responses, but one brought about after a conflict between intuitive responses has been solved. Moreover, the only possible way that I see for avoiding the cost of deliberation is making the prudent proposition, **p**, one that can be rationalized by the decision-maker. Otherwise, we would be losing part of the *essence* of a nudge since cognitive decoupling is another way of referring to deliberating, whose costs nudges try to avoid. In a nutshell, interventions that trigger Type 2 processes of the cognitive decoupling kind (i.e., the one following the *boost* approach) cannot be labeled as nudges because they incur too high cognitive costs and nudges aim at changing behavior more subtly.

I am aware that making nudges dependent on the rationalization of intuitive responses might be controversial. Some might challenge this view and deny that rationalization of intuitions and voluntary behavior are compatible notions. Those potential critics would see rationalization as a not-good-enough kind of Type 2 processing, one that renders whichever outcome out of it less than *fully* sincere and therefore making any intervention that plays with rationalization something similar to manipulations. Against this possible view, I would argue that rationalization does not necessarily need to imply *irrational* or *damaging* behavior and it may well be a *good* or *efficient* strategy with which to process new information. Take for example the case of someone who values above all the feeling of belonging to a community and who takes whatever her preferred political party says as a guide for behavior. Faced with a problem cueing different intuitive responses, rationalizing the one which the political party supports might make her better off – in terms of time and energy consumed – than engaging in a decoupling process that would arrive at the same conclusion (unless she is someone who enjoys engaging in such kind of thinking, of course).

2.4.3 New nudge-like interventions

Before wrapping up this section, let's see in what ways some of the nudge-like interventions differ from nudges if we apply the new framework. Manipulations have already been mentioned along the way in detailing the framework. We have seen that **manipulation** is a kind of intervention that focuses on introducing a new intuitive response or making more salient an existing one in such a way that overshades any other intuitive response – if there is any left. This leads to the impossibility for Type 2 processes to detect any conflict and, therefore, the proposition put forward by the manipulation is the one brought about. We saw above that manipulations and **undetected nudges** only differ from each other in the assumption about the initial state of mind: while the former disregard the initial state altogether, the latter at least assumed that the '*deliberative system*' would initially support the 'prudent proposition'. However, since in the new

framework, there is no mention of any Type 2 processing, the previous difference has somehow faded away. Of course, we could still assume that undetected nudges are introduced in contexts where previous behavior and/or stated preferences suggest that the decision-maker would go for the 'prudent proposition', but there is no way of specifying such a thing with the new language.

The case of **social benefit nudges** was an interesting one. According to Heilmann, in this kind of intervention, the social planner aims at changing the 'automatic system' so that it promotes an outcome that is socially beneficial, regardless of the decision-maker's stated preferences/previous behavior, which pointed in another direction. But the gist was the fact that, while targeting the 'automatic system', the intervention creates a *nudge position* that Heilmann labeled as 'conflicting'. In his own words, "[t]he dis-alignment, and the insistence of the social planner that the prudent proposition really describes the better choice creates a conflict for the decision-maker" (ibid., p. 85). That is, the intervention seems to play with both the automatic and the reflective systems. I would argue that this kind of intervention would not be so *strange* if we analyze it with the new framework. Here, we would have that the initial situation is one in which the problem cues an intuitive response, **T1q?**; the intervention changes it so that now there is a quicker/stronger new intuitive response, **T1sq?**; and this leads to the detection of a conflict by Type 2 processes, which in turn call for more Type 2 processing. So far so good. But, given the concern around the creation of such a conflict ("By insisting on the choice that the Nudge promotes because it is a socially beneficial choice, decision-makers can experience *deep conflict, which incurs considerable costs*" (ibid., *emphasis* added)), we could assume that the extra Type 2 processing is of the cognitive decoupling kind. This would render this type of intervention into something else than a nudge, according to the new framework.

Finally, **social advertising** was a kind of intervention that directly appeals to people's deliberative capacities to convince them to reconsider their choices. That is, these interventions take a top-down approach to behavioral change and as such, they would escape our framework, which is based on a bottom-up approach.

2.5 Concluding remarks

In this section, I have synthesized the results of the previous three in a framework whose goal is to improve the understanding of the cognitive mechanism behind nudges. Such a framework follows the path opened by Heilmann (2014) insofar as it maintains the same four cognitive assumptions, but it also introduces important modifications. The first one has to do with the development of a new dual-process language with which to spell out the assumptions. This language takes into account (1) the outcomes of the defense that Evans and Stanovich (2013) made of dual-process theories, switching Heilmann's 'system terminology' to a 'process terminology' one; and (2) the findings of Pennycook and colleagues' (2015) dual-process model. This model breaks down the decision process into three stages: in the first one, some intuitive (Type 1 processes) responses are cued by a problem, after which, in the second stage, conflicts between the responses are monitored via Type 2 processes, if any is detected, more Type 2 processes are inquired in the third stage. These additional Type 2 processes can take the form of either *cognitive decoupling* or *rationalization*, each of them probably arriving at different outcomes. Thus, we arrive at expressions such as **T1pq?T2Rp**, meaning that the problem firstly cues Type 1 intuitive responses **p**, **q**, and an unknown other denoted by **?**; the conflict between the responses is detected and Type 2 processing engages in rationalizing response **p**. As it is evident from the example, the new language is less elegant than the one used by Heilmann. This is the price that must be paid in exchange for the codification of a slightly more complex dual-process model. Whether or not it is worth paying such a price would naturally depend on the intentions with which the language is used. In the present case, I see it as a requirement for bringing together the diversity of theories explaining the causes for the believing/sharing of misinformation in the next part of the thesis.

The second sort of modification comes from the implications of using the new language in order to detail the four assumptions. Contrary to Heilmann's framework, the initial state of mind in the new framework does not mention the existence of any Type 2 processing. This is due to the fact that biases are thought to be originated from the instantiation of an intuitive response after no conflict between cued intuitive responses is detected. Thus, according to the new framework, the initial state of mind would be formulated as, for example, **T1q?**. The intervention assumption, **T1q?** $\rightarrow$ **T1pq?**, does not differ much from the one in Heilmann's framework, but it reinforces the point of leaving, in this case, the intuitive response **q** as one of the cued ones in order to be able to arrive at the nudge position assumption. This latter assumption, **T1pq?T2**, also significantly differs from the one formulated by Heilmann. It does so in two ways, firstly because in the new framework, this assumption clearly specifies in which way is guaranteed that the decision-maker could have a chance to resist the nudge, via the detection of a conflict between some intuitive responses. And secondly, the new assumption limits itself to state that some Type 2 processing will be triggered, without specifying which sort. Finally,

the new assumption regarding the choice position is also more detailed than in Heilmann's framework, since the former explicitly states that for an intervention to be considered a nudge, it must trigger the decision-maker the rationalization of the desired response, otherwise, it would be either manipulation or entail too-high cognitive costs.

In the third chapter of this thesis, I will test the usefulness of the newly developed framework by applying it to the context of the spread of misinformation in social media. As we saw in the introductory chapter, different interventions aiming at curtailing misinformation and labeled as nudges have been proposed in the literature. What turns them interesting for us is the fact that each of them emerges from different lines of research on the cognitive mechanisms responsible for the believing and/or sharing of misinformation. Thus, the goal of the next part will be to analyze whether any of such mechanisms grants the introduction of nudges, according to the new framework. To do so, I will first translate the say mechanisms into the language developed in section 2.4.2 to then check whether they meet the requirements established by the framework for the introduction of nudges.

Chapter 3

The potential of nudges to curtail online misinformation

In the previous chapter, we used a toy example that starts with a Twitter user encountering a piece of misinformation in her timeline containing unflattering news about a disliked politician. The example goes on and imagines both the possible cognitive and behavioral responses triggered after the user sees the tweet. In particular, in section 2.3.2.1 we hypothesized that, upon reading the tweet, the user might experience different *intuitive responses* depending on whether the (mis)information (1) aligns with her political desires/expectations, (2) is not the first time that she encounters it, or (3) comes from a (dis)trusted source. After such intuitive responses, if a conflict between them is detected, the user can engage in two different types of Type 2 process reasoning – decoupling or rationalization – which would ultimately determine the user's final decision as to whether or not to share/believe the information. This being said, it is crucial to remark that the goal of such an example was only to flesh out the dual-process framework and the cognitive assumptions behind nudges, but nothing more than that. Thus, even if they are reasonable assumptions (as should be clear by the end of the present part), with such an example, I did not intend to claim that the cognitive mechanisms mentioned and the interaction between them represent the state-of-the-art of psychology of misinformation; the use I made of them was purely illustrative of the abstract framework. It will be the goal of the current chapter to turn to the actual research on the cognitive aspects of misinformation to analyze whether the main theories that try to explain its spread meet the requirements to introduce nudges (as defined in the previous chapter) that could curtail the spread of misinformation.

In particular, we will focus on two hypotheses about the cognitive causes for the spread of online misinformation: the one that blames it on people engaging

in politically motivated reasoning (section 3.1) and that which postulates that people lack attention to accuracy when consuming news on social media (section 3.2). For both hypotheses, we will follow the same strategy: introduction of the hypotheses on their own terms (3.1.1 and 3.2.1 respectively), translation of the hypotheses into the dual-process language developed in the previous chapter (3.1.2 and 3.2.2), and finally the application of the definition of nudge to each hypothesis in turn (3.1.3 and 3.2.3). Section 3.3 briefly presents the concluding remarks.

3.1 (Politically) Motivated Reasoning

In the toy example used in the previous part, we seem to have assumed, at least implicitly, that sharing misinformation was the *cognitively* wrong thing to do. The share of misinformation was due either to following intuitive responses or to a rationalization of one of those. In any case, it was the result of a lack of *proper thinking*. The use of more energy-intense decoupling Type 2 processes would have allowed the user to differentiate between false and accurate information. As we shall see in the next section, such an assumption is key in the line of research followed by Pennycook and colleagues, which put the focus on people's lack of attention. But are people *really* making cognitive mistakes when they share misinformation online? Are biases that lead to the spread of misinformation online an indication that people are *irrational*? The theory that will be considered in this section, *politically motivated reasoning* (PMR henceforth), challenges such a conclusion. As we shall see in this section, PMR would argue that, from people's point of view, solely focusing on *accuracy* does not always make them better off. PMR's proponents claim that especially in those occasions in which processing new information based on their accuracy threats people's position within their affinity group, it might be more rational for individuals to credit the new information piece in such a way that their social identity is protected. In the following subsections, we will specify how exactly people's minds would operate according to PMR.

However, while, at least sometimes, it seems individually rational to engage in identity-protective cognition, that does not mean that it is also socially rational. For example, Kahan (2017), one of the main PMR theorists and the one that we will follow here, has coined the term 'the tragedy of the science communications commons' to describe those situations in which there is a scientifically established right position, but some groups have taken a different stance, making it an important part of their definition as a group. If in such situations people engage in PMR to protect their well-being as members of such groups, "the citizens of a pluralistic democratic society are less likely to converge on the best possible evidence on threats to their collective welfare" (ibid., p. 7).

Thus, in order to prevent the societal harm caused by people engaging in PMR, different kinds of strategies have been proposed. For example, Kahan has argued for the introduction of "interventions that remove the *expressive incentives* individuals face to form perceptions of risk and related facts on grounds unconnected to the truth of such beliefs" (Kahan, 2013, p.419). In other words, such policies would aim at disentangling the social meaning from the pieces of false information so that it is no longer rational to credit them. How the particular interventions would look is something that Kahan does not specify.

Since in this thesis we are interested in nudge interventions and the goal of this part is to study whether different theories about the cognitive mechanisms that make people fall for/share misinformation allow for the introduction of

nudges in social media to curtail their spread, the rest of this section will proceed as follows: 3.1.1 introduces PMR in its own terms, putting especial emphasis on (i) the description of the cognitive mechanisms behind PMR and its comparison with other information processing styles, and (ii) the prediction according to which the smarter a person is, the more able to engage in PMR she would be. Section 3.1.2 looks at how the PMR account would look if it were spelled out using the dual-process language depicted in 2.4.1. Finally, section 3.1.3 assesses whether nudge interventions as described in 2.4.2 are possible if people display PMR.

3.1.1 The approach in its own terms

Especially since the 2016 US Presidential Election, political polarization has become a popular concept with which researchers, politicians, journalists, and the general public try to theorize and understand the increasing political division among citizens. While part of that division could be explained by differences in the political and moral values held by members of society (for example, people with different understandings of 'freedom' and 'equality' disagree on the desired distribution of taxes; Moore, 2015), more puzzling is the one due to disagreements over questions of fact. This later sort of division occurs when people contend about empirical questions, that is, questions that, even if complex, should be clarified by looking at the *evidence*. One of the most popular examples of such disputes concerns whether humans are responsible for climate change, as scientists have been claiming for decades. Usually, people on the political right reject such a claim, and those on the left argue for it. Explaining how people may turn their back on clear evidence is the goal of Kahan's (2016a) conceptual model of *politically motivated reasoning* (PMR).

Simply stated, PMR is just one of the flavors that *motivated reasoning* can take. In Kahan's words, motivated reasoning "refers to the tendency of individuals to unconsciously conform assessment of factual information to some goal *collateral* to assessing its truth" (ibid., p.2; emphasis in the original). Thus, according to Kahan, the political variant of this tendency would pose that *identity protection* is the collateral and truth-independent goal; where identity protection just means "the formulation of beliefs that maintain a person's status in an affinity group united by shared values" (ibid., p. 3). That is, people who disregard the scientific evidence about human-caused global warming would be doing so because holding such a belief is what, supposedly, partly defines being right-wing.

But, as just mentioned, PMR is only one of the forms of motivated reasoning, and other collateral goals have also been researched in detail. For example, Russo and colleagues have studied people's aversion to complexity, by which they prefer coherence to the task of disentangling the complex truth whenever an important decision is due (Russo, Carlson, Meloy, and Yong, 2008), while Dunning (2003) theorizes people's goal of forging a positive self-conception. In

this thesis, given its limited space, I will solely focus on PMR since I take it to be a strong candidate for the explanation of the spread of misinformation online. Of course, this choice does not mean that other collateral goals do not play any role in such a phenomenon, but their exact compatibility with the dual-process framework devised in the previous part is left for future works.

Coming back to PMR, according to this theory, accuracy might not be the only aspect of information that people care about. There could be some occasions when whether or not to believe a specific piece of information, regardless of its veracity, has significant social consequences. In particular, such occasions often involve information about topics around which a community has formed its identity. There, an individual that feels part of the community, would have incentives to credit information that reinforces the community's position and discredit the ones that threaten it, independent of their veracity. In Kahan's own words (2016a, p.2):

> "Where positions on some policy-relevant fact have assumed widespread recognition as a badge of membership within identity-defining affinity groups, individuals can be expected to selectively credit all manner of information in patterns consistent with their respective groups' positions. The beliefs generated by this form of reasoning excite behavior that expresses individuals' group identities. Such behavior protects their connection to others with whom they share communal ties."

Moreover, such an information processing strategy could be read as *rational* if we consider that a person's life might be more heavily affected by how tightly the connection with her community is than by the acquisition of accurate information regarding a topic over which she has close to no influence. After all, being part of a community open the door to a safety net of material and immaterial resources that *just being right* cannot offer. Coming back to the example from the previous part of the thesis, we can think of our Twitter user as someone with sympathies towards the Spanish Popular Party. Such a party has as one of its badges of membership demonizing Pedro Sánchez, Spain's Prime Minister and leader of the Socialist Party. Since the feeling of belonging to the party, in this case, manifested in her interactions with other Twitter users, provides her with emotional support, and potential access to material resources, she would have incentives to process every piece of information regarding Pedro Sánchez in a negative light, even those one that covers his *right* decisions. Otherwise, if she attempts a truthful evaluation of the information, she risks a backlash from her own community. Being right does not always pay off.

Here, it is important to remark that PMR does not only has to do with misinformation but any kind of information could also be subjected to this type of processing. What is relevant about the information is its direction; that is, whether it strengthens or threatens the group's position. In other words, polarization and PMR are perfectly possible in a world where only true information

circulates; in such a case, people would just dismiss information detrimental to their interests. This observation does not mean to imply that misinformation is not important after all, it is, but it helps us to specify how people exactly interact with misinformation.

In particular, PMR allows Kahan to differentiate between two models of misinformation and favor the "motivated public" model over the "passive aggregator" one. While according to the latter "a largely credulous public is assumed to be maneuvered into states of misunderstanding and confusion by economically or ideologically interested groups, which transmit misinformation through the media" (2017, p.4), the former does not leave the people out of the hook and emphasizes "the stake individuals have in holding beliefs that protect their identities creates a profitable opportunity to supply them with information, including misinformation" (ibid.). In the next sections, we will study whether, assuming that people engage in PMR, it is possible to introduce nudges that would let them dodge the otherwise profitable misinformation. Before that, we still need to introduce PMR in a more structured way so that we can translate it into the dual-process framework developed above. For this, Kahan's politically motivated reasoning paradigm will come in handy.

PMR model

Following what we said above, we could define PMR as 'the tendency of individuals to unconsciously conform assessment of factual information to the protection of their identity as members of a community, regardless of its truth'. To fully grasp what this 'pursuing identity protection at the expense of truth' means, Kahan proposes a conceptual model of PMR formulated as a comparison with a Bayesian information processing model, representative of one 'truth-convergent' processing of information. That is, in order to understand PMR, Kahan introduces a model of what *unmotivated reasoning* would look like. Let's see in detail the two models, while also illustrating them through the Twitter user toy example.

To model situations in which someone must deal with new information, Kahan makes use of what is called a *barebones Bayesian model*. Such a model, which will serve as the starting point for the two models that we are interested in, the 'truth-convergent' and the PMR ones, consists of:

i a person with a *prior* – an initial estimation of the probability of a hypothesis –,

ii who encounters *new evidence*,

iii that makes her *revise* the initial factual beliefs.

To illustrate, think of our Twitter user. We have imagined a Spanish right-wing citizen scrolling down her Twitter feed. Since people with right-wing

sympathies in Spain tend to dislike Spain's Prime Minister, Pedro Sánchez, we can pose that our user's *prior* consists of a probability of 4 to 1 above the hypothesis that 'Pedro Sánchez is not apt for his job'. The *new evidence* comes in the form of a tweet posted by the news outlet OKdiario claiming that 'Pedro Sánchez would have made some comments about possible illegal founding of his political party'. After reading the news, the Twitter user updates her beliefs so that now, the odds of Pedro Sánchez not being apt for his job is 9 to 1.

As simple as this structure and the example seem, the key to the matter, and that which sets apart different ways of processing information, lies in how the *revision* is conducted; that is, how the prior beliefs are affected by the new evidence. The feature of Bayes' theorem that deals with this is the *likelihood ratio*, a characteristic attributed to the new information at which we can arrive through different ways. Once we know the likelihood ratio, if we want to know the new estimates, Bayes' theorem simply asks to multiply such a ratio by the odds that the prior beliefs grant to the hypothesis. In the example above, we said that, before encountering OKdiario's tweet, the user gave a 4 to 1 probability of 'Pedro Sánchez not being apt for his job as Prime Minister'. If the revised odds were 9 to 1, then we have that, according to the user, the information's likelihood ratio is 2,25.

If we accept this way of proceeding, then everything would depend on *how the likelihood ratio is derived*. For that, Bayes' theorem is of no help; as Kahan points out, "Bayes' theorem does not say *how* to figure out the likelihood ratio, only what to *do* with it: treat it as a factor by which one multiplies one's prior odds" (ibid, p. 4). In our example, we simply said that the likelihood ratio was 2,25 since it is the one that took the user's beliefs on the hypothesis from 4:1 to 9:1; however, we did not mention how the user arrived at that likelihood ratio – what cognitive mechanisms made her credit the information in such a way. As mentioned before, we will follow Kahan and focus on two possible ways of deriving the likelihood ratio, the truth-convergent Bayesian way and the PMR one. Let's see now what the two models say about it.

The first model that we will consider is the truth-convergent Bayesian information processing. As can be seen in Figure 3.1, in this case, we have that the likelihood ratio of the new evidence is derived following a 'truth-convergent criteria'. The ratio is then multiplied by the prior odds so that we get the posterior odds, as Bayes' theorem mandates. Importantly, Kahan does not extend much on what a truth-convergent criterion entails. The only thing he has to say is that "[the likelihood ratio] reflects how much more consistent the information is with the hypothesis than with some alternative" (ibid., p.3).

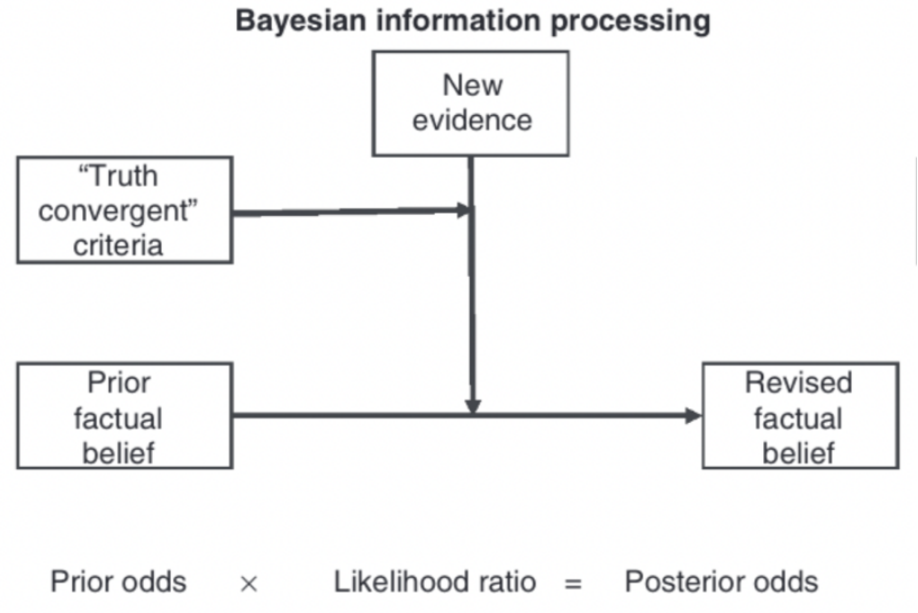

Figure 3.1: (Kahan, 2016, p.3)

Applying the model to our example, we would have that the 'truth seeker' user found the new evidence (tweet claiming that 'Pedro Sánchez would have made some comments about possible illegal founding of his political party') 2,25 times more consistent with the hypothesis that 'Pedro Sánchez is not apt for his job' than with the rival hypothesis that 'he is'. That is all.

As sparse as the truth-convergent Bayesian model is, it is sufficiently informative as a model with which to compare PMR. In this later model, according to Kahan, someone engaging in politically motivated reasoning would not follow a truth-convergent criterion to derive the likelihood ratio, but she would do so "from the impact crediting [the new evidence] will have on aligning her beliefs with those of others in an identity-defining group" (ibid., p.4). As can be seen in Figure 3.2 *political predispositions* but not 'truth-convergent criteria' intervene in the processing of assessing new evidence so that the *revised factual belief* is similar to those held by members of the desired group.

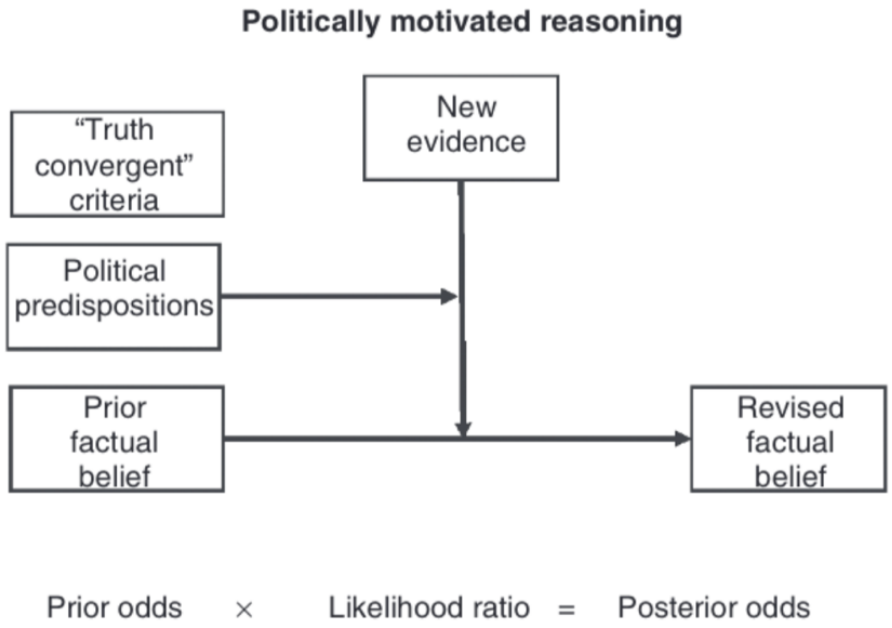

Figure 3.2: (Kahan, 2016, p.3)

For our Twitter user, what matters now about the new information is not how consistent is with the hypothesis but whether giving credit to it would reinforce her identity as a member of the group – the Spanish Popular Party supporters. Since the likelihood ratio was larger than 1, it indicates that believing the information contributes to the user's position within her group.

Before moving on to the next subsection, it is worth quickly differentiating between PMR and *confirmation bias*, two ways of information processing that given their similarity have given rise to some misunderstanding. Figure 3.3 shows Kahan's model for confirmation bias processing. Analogously to PMR, someone engaging in confirmation bias does not derive the likelihood ratio for the new evidence in a truth-convergent way but, unlike PMR, it neither does it following identity-protective motives. In this case, what matters is "[new information] consistency with one's existing beliefs" (ibid.). In other words, confirmation bias' distinctive feature is to derive the likelihood ratio from one's priors. For our Twitter user that would mean that whether or not to believe Pedro Sánchez's supposed illegal funding confession depends on how much this aligns with her previous beliefs. Since the likelihood ratio was established at 2,25, then we can assume that this is so because the new information is consistent with what she thought about Pedro Sánchez before encountering the tweet.

To sum up, in this subsection, we have introduced Kahan's conceptual model of politically motivated reasoning. PMR is a way of information processing

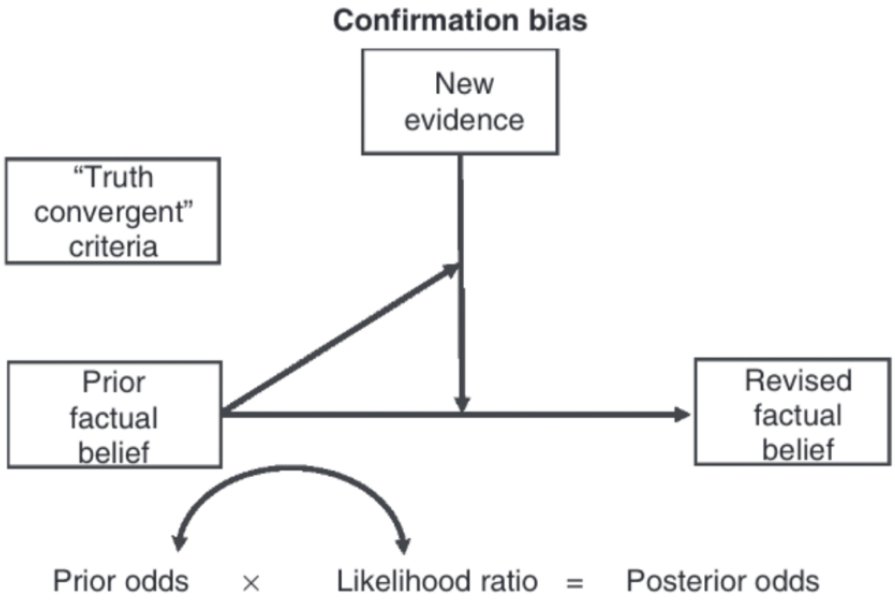

Figure 3.3: (Kahan, 2016, p.3)

according to which people engaging in it will weigh new information based on the congeniality of such information with their political predispositions. This way of information processing can be distinguished from, for example, confirmation bias or that one in which the goal is to assess the veracity of the information. Kahan's conceptual model helps us locate with precision where these three ways of information processing differ when modeled in Bayesian terms: how people derive the likelihood ratio of the new information.

As was also pointed out above, if people engage in PMR, it might open the door for interested parties to feed them misinformation. The possibility of introducing nudges for curtailing misinformation and that work on the assumption that people engage in PMR will be discussed after we translate Kahan's PMR model into the dual-process framework developed above but before that, let me delve into a feature of PMR – named, how it correlates with people's cognitive sophistication – that sets it apart from the hypothesis that emphasizes people's lack of attention as the cause of the spread of misinformation.

PMR and cognitive sophistication

Besides the identity-protecting goal when (dis)crediting new evidence, researchers have focused on another crucial element of PMR: the possible correlation between politically motivated reasoning and cognitive sophistication. In

a nutshell, the hypothesis claims that if people who engage in politically motivated reasoning do fit better within their affinity group, then the smarter – more cognitively sophisticated – a person is the better she can deploy her PMR abilities and therefore the more her identity is protected. In the following, I will present how Kahan and others have researched this empirical claim, which will in turn serve as a bridge for the translation into our dual-process framework in the next section.

In *Ideology, motivated reasoning, and cognitive reflection*, Kahan (2013) runs an empirical study to test three different hypotheses about the psychological mechanisms that make people polarize over questions of fact – that is, that makes some people disregard what science has to say about certain topics –, named 'bounded rationality position' (BRP), 'ideological asymmetry position' (IAP), and 'expressive utility position' (EUP). The first one, BRP, poses that the main reason for such polarization is that some people are driven by heuristics when processing information. In explaining the origin of this *heuristic-driven information processing hypothesis*, Kahan roughly grounds it on dual-process theories in the following way (ibid., p. 408): "Many scholars attribute controversy over societal risks to the disposition of members of the public to over-rely on the heuristic-driven, "System 1" (Stanovich and West, 2000; Kahneman, 2003) reasoning style. The centrality of visceral, emotion-guided modes of perception can cause laypeople to overestimate the incidence and harm associated with more sensational risks [...] Expert opinion does not reliably correct these distortions because members of the public too often lack the time or ability to engage in the more effortful, more dispassionate "System 2" style of reasoning suited to understanding the technical evidence that experts use to assess risks (Loewenstein, Weber, Hsee and Welch, 2001; Sunstein, 2003, 2006, 2007; Weber, 2006)". As should be evident at this point, this understanding of dual-process mechanisms does not perfectly align with the one depicted in the previous part of the thesis, but this should not stop us to get the main gist of this hypothesis: presented with certain information, the Type 1 intuitive responses cued are so *strong* that the person acts upon the fastest one without further considerations. According to Kahan, this heuristic-driven processing might also interact with PMR, but the latter takes a secondary role. In particular, PMR would only enter the picture as the reason that explains the ideological turn of polarization, thus "[m]any of the emotional associations that drive System 1 risk perceptions, it is posited, originate in (or are reinforced by) the sorts of affinity groups that share cultural or ideological commitments (Leiserowitz, 2005; Sunstein, 2007)" (Kahan, 2013, p. 409).

The *ideological asymmetry position*, also known as the neo-authoritarian personality thesis, grounds both the use of heuristics and PMR on right-wing personality traits, such as dogmatism, need for closure, or aversion to complexity. It is not that people with other ideologies do not express those traits, but they are "disproportionately associated with that ideology by virtue of the negative correlation between conservatism and the traits of open-mindedness, and

critical reflection that would otherwise check and counteract it (Jost, Hennes, Lavine, 2013; Nam, Jost and van Bavel, 2013)" (Kahan, 2013, p. 409).

Finally, the *expressive utility position* sees politically motivated reasoning as the main driver. When explaining PMR here, Kahan does not only use a similar exposition to the one detailed above – seeing PMR as a way of information processing that takes them to perceive facts congruently with what is thought by members of affinity groups –, but he also stresses how we should understand this type of information processing through the lens of (his version of) dual-process reasoning. Thus, Kahan claims that contrary to what should be expected according to BRP, people with more skilled System 2 would be better at aligning their beliefs with those of their peers. To fully grasp such a dynamic, let me cite Kahan in length (ibid.; *emphasis* in the original):

> If we imagine that socially adaptive pressures will favor reasoning styles that maximize this form of "expressive utility" (Gigerenzer, 2002), we might, on this account, expect the use of more effortful, System 2 forms of information processing to *magnify*, not mitigate, ideological differences. Individuals disposed to resort to heuristic-driven, System 1 cognitive processing should not have too much difficulty fitting in: Conformity to peer influences, receptivity to elite cues, and sensitivity to intuitions calibrated by the same will ordinarily guide them reliably to stances that cohere with and express their group commitments (Zaller, 1992; Gastil, Braman, Kahan and Slovic, 2011). But *if* individuals are adept at using more effortful, System 2 modes of information processing, then they ought to be even *better* at fitting their beliefs to their group identities. Their capacity to make sense of more complex forms of evidence (including quantitative data) will supply them with a special resource that they can use to fight off counterarguments or to identify what stance to take on technical issues more remote from ones that figure in the most familiar and accessible public discussions (Chen, Duckworth and Chaiken, 1999). More importantly still, it will make them more likely to *understand* the significance of competing claims, and related forms of evidence, for the status of their group, and thus be more likely to experience unconscious motivations to form identity-congruent assessments of them.

Here, again, Kahan uses a dual-process language different from the one that we developed above – the translation to it will be the task of the next section – but for now, it is sufficient to remark on the differences between EUP and BRP. The latter, we saw above, poses that some people's inability to tell true from false information – that in turn deepens polarization – is due to their overreliance on Type 1 reasoning coupled with insufficient, if any at all, Type 2 one. In other words, people fail to distinguish truth from falsity because they do not think enough. On the other hand, EUP hypothesizes that the more

sophisticated Type 2 reasoning skills a person has, the more successfully can such a person process information in a way that strengthens her identity as a member of an affinity group; that is, the smartest a person is, the higher the chances for believing misinformation/discrediting true information if that preserves her identity.

To test the three hypotheses, Kahan first measured individuals' cognitive reflection through what is known as Cognitive Reflection Test (CRT). Such a test encompasses three questions aimed at gauging people's "disposition to engage in the conscious and effortful form of information processing associated with System 2 as opposed to the heuristic-driven form associated with System 1" (Kahan, 2013, p. 410). After the CRT scores were collected, the actual test looked for people's willingness to accept evidence pointing out that people holding opposing views on heavily disputed topics were open-minded and reflective.

I will not go into the experiment's details since the current thesis' argument is sufficient to briefly report on the results. The experiment did not find evidence supporting BRP or IAP. Contrary to what is predicted by the former, the study found that the more System 2 processing was in place – according to the CRT scores – the bigger the impact of PMR. Against the latter, the experiment did not detect any relevant correlation between right-wing ideology and CRT nor perceived significant differences between right-wing and left-wing supporters when they reported about the open-mindedness of members of the opposing group. On the other hand, Kahan's study showed some support for EUP: one of the two experiments that provided subjects with information expected to polarize found that that was indeed the case as CRT scores increased.

As we shall see in section 3.2, these results are far from being uncontroversial, and the discussion around the existence of (politically) motivated reasoning and its relationship with the spread of misinformation is a hotly debated one. However, Kahan's study makes a very valuable contribution since it addresses the question from an angle not very developed in the field: An in-deep study of the "status of motivated reasoning within dual process reasoning theories" (Kahan, 2013, p. 418). Moreover, he does so in a way that goes against what has been amply assumed in the field, "that ideologically motivated cognition is a manifestation of unconscious, heuristic-driven reasoning process amenable to being overridden by dispositions that promote reflection and critical engagement with counter-attitudinal evidence (e.g., Lilienfeld, Ammirati, and Landfield, 2009; Sunstein, 2006; Westen, Blagov, Harenski, Kilts, and Hamann, 2006)" (ibid.). As we saw above, the results of his experiments point to the opposite conclusion: motivated reasoning positively correlates with higher levels of cognition. In the next section, we will take up Kahan's call for more research into "the relationship between ideological polarization and information processing" and translate his account of PMR into the dual-process language developed in the first part of the thesis. While this translation might be further developed to

generate empirical hypotheses, in this thesis it will serve us to assess in section 3.1.3 whether it is possible to introduce nudges against misinformation under the assumption that the spread of it is due to PMR.

3.1.2 Translation

The goal of this second part of the thesis is to apply the nudge framework devised in the previous part to two competing hypotheses about the cognitive mechanisms responsible for the believing/sharing of misinformation. Since none of these hypotheses are necessarily grounded on – nor expressed in – the required dual-process language, we will translate them into the language developed in section 2.4.1. Just to remember the reader, such a language consisted of a modification of Heilmann's dual-process language so that it accommodates Evans and Stanovich's view on dual-process theories and Pennycook and colleagues' three-stage dual-process model. Thus, the new language's two main novelties lay in the use of *processes* instead of *systems* when referring to the two types of reasoning, and in the further distinction of two kinds of Type 2 processes, *cognitive decoupling*, and *rationalization*. For example, in this new language, we could find expressions like **T1pqT2Dp**, which refers to the situation in which a problem cues the Type 1 intuitive responses **p** and **q**, and decoupling Type 2 processing promotes **p**. The question is, naturally, how could we express PMR in such a language? Before directly tackling the translation, let me recap the two main characteristics of PMR that we saw above and that therefore need to appear in the language.

In the previous section, we have introduced Kahan's account of politically motivated reasoning in its own terms. Firstly, as a conceptual model that defined PMR as 'the tendency of individuals to unconsciously conform assessment of factual information to the protection of their identity as members of a community, regardless of its truth', while comparing it to a truth-convergent Bayesian model and a confirmation bias one. And secondly, emphasizing the relationship between PMR and cognitive sophistication, by which the higher the level of reasoning – more use of System 2 reasoning as measured by CRTs –, the more motivated reasoning is shown. That is, we could see PMR as a cognitive phenomenon according to which the more *cognitively sophisticated* a person is, the more capable of an *unconscious* assessment of information is too. Let's see now how this apparent contradiction can be expressed in our three-stage dual-process language.

As we saw above, the kinds of situations where people show politically motivated reasoning are those in which they have to evaluate new information regarding a topic about which they have some prior beliefs. If we think of this situation in terms of the three-stage model, we could say then that in the first stage, we have a piece of information that cues some Type 1 intuitive responses. From Kahan's description, what those intuitive responses would entail cannot be clear. However, what we can be sure about, given that Kahan discarded

the bounded rationality position and its heuristic-driven information processing, is that those more politically motivated individuals do not have intuitive responses overwhelming enough that would make them act upon them without further consideration. That is, since according to Kahan, PMR is not due to heuristics – Type 1 processing – then we must presuppose at least two contradicting intuitive responses, **T1pq?**, with the question mark just leaving the door open for other intuitive responses.

Bringing back the Twitter user example, we could assume that quickly seeing the tweet about Pedro Sánchez's alleged comments cues in our user at least two intuitive Type 1 responses. The first one, **p**, promotes accepting the information since it puts Pedro Sánchez in a bad light, and believing that he is incompetent is one of Popular Party supporters' defining features. On the other hand, intuitive response **q** promotes not believing the information since it comes from a source that the user takes as a misinformation distributor. As mentioned above, there could be more intuitive responses, for example, one that would call for believing the information since it would not be the first time that she sees it, which makes it easier and faster to process. But since for our purposes, we only need two contradicting intuitive responses, **p**, and **q** are enough.

In the second stage, given the presence of opposing intuitive responses, **p** and **q**, and the fact that the fastest of them was not directly implemented, then we must conclude that a Type 2 monitoring process has been carried out and a conflict between **p** and **q** has been detected. This Type 2 processing calls for more – and different – Type 2 processing in the third, and last, stage of the reasoning process. In our example, the second stage would consist of just the Type 2 monitoring process detecting the conflict between the two Type 1 intuitive responses – whether to believe/share the information – and demanding more Type 2 processing in the next stage so that a decision can be reached.

Finally, in the third stage, after the detection of a conflict between intuitive responses, we saw that two different kinds of Type 2 processing can evaluate the situation and promote a response, *cognitive decoupling*, and *rationalization*. Given that we discarded Type 1 processes as the source for politically motivated reasoning, this must be caused by either cognitive decoupling or rationalization (or both). Let's start with the latter possibility. We claimed above that PMR can lead people to credit false information if that reinforces their identity as members of an affinity group. Since disregarding the truth could be seen as a sort of cognitive mistake – a bias – then it might be tempting to consider PMR an instance of rationalization. In other words, PMR might not be the kind of bias caused by intuitive Type 1 processes, but it must not be the outcome of a conscious decoupling process where the intuitive responses are weighted according to their accuracy. Thus, PMR should be seen as a middle way between heuristic bias and fully engaged cognitive capacities. It is important to note that this reading of PMR assumes that the quickest intuitive response is the one that processes the information in line with what is expected from a member of the

affinity group. Otherwise, another response would be rationalized, one that disregards the implications for the individual's identity. This assumption is thus a limiting factor for understanding PMR as derived from a rationalization Type 2 process since it confines PMR to only those cases in which identity-protection processing is engrained enough to be the first intuition that comes to mind. In any case, the possibility of PMR as an instance of rationalization is thus supported by the *unconscious* component that Khan attributes to PMR (recalling Kahan's definition of motivated reasoning, "[it] refers to the tendency of individuals to *unconsciously* conform assessment of factual information to some goal collateral to assessing its truth" (Kahan, 2016a, p.2; *emphasis* added), one of the two features that we were looking to translate into our dual-process framework.

However, rationalization might not be the only Type 2 processing that caused PMR. *Cognitive decoupling* could be argued to be a strong candidate as well. As we saw in section 2.3.2.3, this kind of Type 2 processing consists of an analysis of the intuitive responses cued in the first stage, which in turn allows the individual to override the quickest of them. Cognitive decoupling would therefore demand a conscious and longer exam of the different possibilities than what was required by the rationalization Type 2 processing. Understanding politically motivated reasoning as caused by cognitive decoupling would require then that we take Kahan's experiment, which seems to conclude that conscious engagement is actually what lead people to show PMR, as implying that PMR needs careful consideration of the consequences that each intuitive response has for reinforcing the individual's identity as a member of an affinity group. In other words, it would be hard to interpret PMR as rationalization if the former, in Kahan's words, claims that "*if* individuals *are* adept at using more effortful, System 2 modes of information processing, then they ought to be even *better* at fitting their beliefs to their group identities. Their capacity to make sense of more complex forms of evidence (including quantitative data) will supply them with a special resource that they can use to fight off counterarguments or to identify what stance to take on technical issues more remote from ones that figure in the most familiar and accessible public discussions (Chen, Duckworth and Chaiken, 1999). More importantly still, it will make them more likely to *understand* the significance of competing claims, and related forms of evidence, for the status of their group, and thus be more likely to experience unconscious motivations to form identity-congruent assessments of them." (Kahan, 2013, p. 409). Finally, an advantage of understanding PMR as an instance of cognitive decoupling instead of rationalization is that the former does not require the PMR *response* to be the fastest of the intuitions in the first stage of the reasoning process, one of the main limitations of the latter.

In conclusion, it is possible to understand politically motivated as being derived from rationalizing Type 2 processes (**T1pqT2Rp**) as well as from cognitive decoupling ones (**T1pqT2Dp**), although none of them seems to capture all the nuances of PMR. Thus, while the latter cannot really explain how to

reconcile the analysis of the intuitive responses with the assumption that PMR should be an *unconscious* tendency to conform facts to the goal of identity protection; the former would have trouble explaining how might be the case that people who understand better what is at stake when processing information that might reinforce their identity as members of a group, and who therefore need to analyze all the intuitive responses, show higher levels of PMR. Which of the two paths for PMR is the most common is of course an empirical question that we cannot answer in this essay. Indeed, to tackle such a question, further defining work is needed, for example, a clearer explanation of what exactly rationalizing Type 2 processing consists of, and how much different is from Type 1 processing. However, this lack of clarity should not stop us from our goal of assessing whether is possible to introduce nudges assuming that politically motivated reasoning is the cause for the spread of misinformation online, which is our task in the following section.

3.1.3 Nudges

In section 2.4.2, we developed a framework that spelled out four assumptions about the cognitive mechanisms behind nudges. Such a framework was based on the one developed by Heilmann (2014) but changed its minimal dual-process approach to a more complex, three-stage one (Pennycook, Fugelsang, and Koehler, 2015), and expressed in a language that follows Evans and Stanovich's (2013) takes on dual-process theories. In the two previous sections, we first introduced *politically motivated reasoning*, a theory that emphasizes the role of identity protection as the reason for which people process information without attending to its veracity, to then trying to explain it using the language and the three-stage dual-process model just mentioned. It is the aim of this section to analyze whether it is possible to introduce nudges to curtail misinformation, assuming that people engage in PMR. To do so, we will just spell out each of the four assumptions in turn.

3.1.3.1 Initial state of mind: T1pq?T2Rp or T1pq?T2Dp

We finished the previous section, 3.1.2, explaining the two possible paths by which someone could show PMR, either rationalizing the first intuitive response, which would make her credit the information that aligns her with her affinity group or decoupling the different intuitive responses so that she would choose the same one but now after carefully considering what it at stake with her decision. If we establish that **p** refers to the proposition led to after conforming the assessment of factual information to the protection of her identity as a member of her affinity community, and **q** refers to the proposition that follows the processing of the information according to its veracity, then we have that the initial state of mind is either **T1pq?T2Rp** or **T1pq?T2Dp**, depending on the kind of Type 2 processing taken. Needless to say that in order to arrive at such PMR initial state, it is necessary to assume that the information originally cued, at least, the Type 1 intuitive responses **p** and **q** in a way in which none of

them was strong enough to make the person directly go for such a proposition; that is, **T1pq?**.

3.1.3.2 Intervention: T1qp?

Since we established in section 2.4.2.4 that nudges play with the rationalization of the intended *prudent* intuitive response, then a nudge that looks for people to believe/share true information needs to bring to the forefront the proposition that would favor truthfully crediting the information, **q**, making it the quickest of the intuitive responses. This intervention, of course, should not eliminate any of the other original intuitive responses, **p** and the *possible* **?**. Thus, **T1qp?**.

3.1.3.3 Nudge position: T1qp?T2

The nudge position assumption just aims to guarantee that the conflict between the propositions **p** and **q** is detected by the monitoring Type 2 processing. Detecting the conflict implies that more Type 2 processing will be requested in the third stage of the reasoning process, where the individual would have a chance to still wave away the proposition to which she has been nudged, **q**.

3.1.3.4 Choice position: T1qp?T2Rq

Finally, we arrived at the choice position assumption in which the individual, if the nudge has worked as intended by the policy-maker, would rationalize the proposition **q**, processing the (mis)information in a truthful way, **T1qp?T2Rq**. But at this point, it is crucial to note that the success of the nudge depends in a great way on the particular PMR style of the individual. That is, if we assume that the individual is one who tends to show PMR after engaging in a cognitive decoupling Type 2 processing, then making the *prudent* proposition, **q**, quicker should not change much for her, she will still credit the (mis)information in a way that aligns with what is expected in her affinity group, **p**. On the other hand, if PMR is due to the rationalization of **p**, then making the proposition **q** the quicker intuitive response might make the individual change her mind and rationalize **q** instead. Note however that the latter mechanism is just a possibility since it is not granted that the nudge would work in such a way. It may be very possible that bringing **q** to the forefront would make the individual engage in cognitive decoupling Type 2 processing after which she would choose to credit **p**. In other words, a nudge targeted to someone engaging in PMR can only be successful under two conditions: (1) the PMR displayed is not one that follows cognitive decoupling Type 2 processing, and (2) in case that PMR is due to rationalizing Type 2 processing, nudging her towards the prudent proposition should not make her engage now in cognitive decoupling Type 2 processing – regardless of the final embraced proposition.

Here, it is important to remark that nudging someone that engages in PMR could be seen as violating the spirit of what a nudge should be in the first place, an intervention that should help individuals choose *what they really want*. If

someone considers that crediting (mis)information in a way that aligns her beliefs with those present in an affinity group is what matters most to her – a choice made after careful deliberation, or cognitive decoupling Type 2 process – then nudging her away from such a choice would make her worse off. Thus, even if the libertarian component, the nudge position, is still present, the paternalist side of the intervention might turn too big to still consider it a nudge. This observation does not mean to imply that nudges under PMR are altogether impossible but that careful work needs to be done to properly understand people's motivations when processing information.

3.2 Inattention-based account

As was pointed out above, politically motivated reasoning has not gone undisputed. Even if it still serves as a paradigm that continues to inspire new research, some of their original findings and claims have proved hard to replicate (Tappin, Pennycook, and Rand, 2020a; 2020b). In the following, I will consider two lines of experiments that get to the conclusion that cognitive sophistication does not magnify politically biased processing – PMR's main hypothesis – but quite the opposite. These criticisms will serve as the bridge to the theory that explains the spread of misinformation online by pointing at people's lack of attention to accuracy, in section 3.2.1. Such a theory, following the same strategy used to study PMR, will be first translated into the dual-process language developed in the first part of the thesis (section 3.2.2), to then assess whether, and if so how, is possible to implement nudges to curtail misinformation under the assumption that the theory is right (section 3.2.3).

First criticism

In section 3.1.1, we saw that Kahan (2017) explained politically motivated reasoning by comparing it with two other types of information processing, barebone Bayesian processing, and confirmation bias. All of them consisted of determining the odds of a hypothesis after reading about some new information by multiplying the person's prior odds by the likelihood ratio of the new information. However, the three processing styles diverge in the way of determining the likelihood ratio; thus, while the barebones Bayesian follows truth-convergent criteria, the other two do not. Under PMR the likelihood ratio is derived through identity-protective motives (the new information is credited depending on its alignment with what is believed by an affinity group), and confirmation-biased processing does so from the priors (that is, based on the similarity between the new information and what was already believed). From this depiction of the three processing styles, Kahan alerts us of the possibility that PMR and confirmation bias can lead to the same final interpretation of new information and, therefore, of the importance of properly designing experiments that distinguish between the two mechanisms so that there are no false attributions (ibid., p.6-11). The high difficulty of designing such experiments is precisely the main argument against PMR made by Tappin, Pennycook, and Rand (2020a).

Indeed, Tappin and colleagues (ibid., p.2) claim that multiple recent studies have failed in keeping PMR and cognitive bias apart and wrongly suggested that cognitive sophistication magnifies politically motivated reasoning (i.e., Kahan, 2013; Kahan, Peters, Dawson, and Slovic, 2017; Nurse and Grant, 2019). Such studies, in Tappin et al words, "randomly assign people to receive one of two pieces of information, holding constant the substantive detail of the information across treatments while varying its implication for their political identities between treatments. The outcome variable is typically people's self-reported evaluations or interpretations of the new information" (Tappin et al., 2020a,

p.3). The conclusion reached in every study was that people's evaluations of the information vary in each treatment and correlate with people's political identities. In other words, how much weight people grant to the information depends on its impact on the person's group identity. From these results, and those about people's cognitive sophistication, the researchers conclude that *cognitive sophistication magnifies politically motivated reasoning.* In Tappin, Pennycook, and Rand's view, such an inference is not granted since political group identity also correlates with variables other than PMR that might cause the same behavior, for example, *prior beliefs* (which in turn may be just "proxies for people's unobserved political information environment – comprising, for example, their exposure to media, discussions with friends, coworkers, and so on" (ibid.)). They illustrated their explanation with the following *oversimplified* diagram:

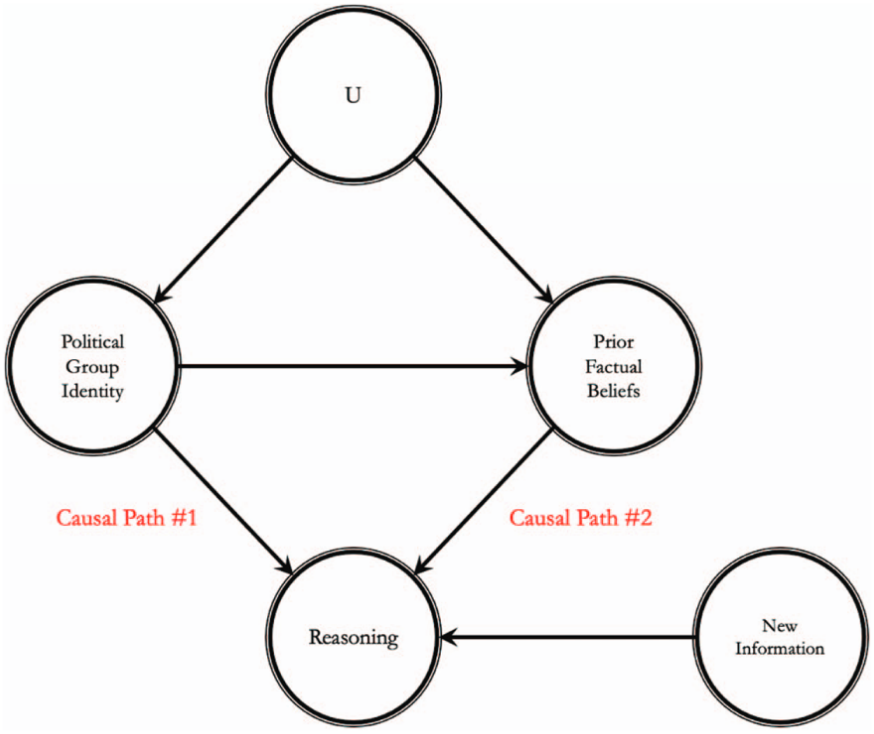

Figure 3.4: (Tappin et al., 2020a, p.3)

In such a diagram, we could interpret reasoning as the likelihood ratio, which is derived from the influence of (1) the content of the new information together with (2) the person's political group identity and/or (3) her prior factual beliefs. What the previous studies get wrong, according to Tappin and colleagues, is to assume that their results are solely caused by the effect of political group

identity alone (*causal path 1* in the diagram above). Such an inference is not necessarily the case at least for three reasons: (i) It is possible that political group identity *indirectly* affects reasoning by causing prior factual beliefs which in turn directly influence the reasoning (*causal path 2* in the diagram), (ii) prior factual beliefs may affect reasoning without they being caused by political group identity, or (iii) there might be other common variables that cause both political group identity and prior factual beliefs (U in the diagram), for example, "one's political information environment, including exposure to media, discussions with friends, family, and coworkers, and the resulting path-dependent and self-reinforcing perceptions about which sources of information are trustworthy and should thus be listened to (vs. ignored)" (ibid., p.4).

Thus, studies that fail to recognize the correlation between political group identity and prior factual beliefs can only show ambiguous evidence for PMR since they would not be specifying the causal path taken. According to Tappin, Pennycook, and Rand, studies of this type are the ones that have been used to provide evidence of the magnification of PMR (ibid.). On the other hand, studies that account for the correlation and asses whether political group identity affects reasoning independently of prior factual beliefs (causal path 1) would be the only ones that can provide strong evidence of PMR. This is exactly what Tappin and colleagues attempt in the empirical part of their article (ibid., p.6-17). There, the authors run two studies replicating that of Kahan (2013) and that analyze "whether cognitive sophistication magnifies a direct effect of political group identity on reasoning (Causal Path 1) [... and] whether cognitive sophistication magnifies a direct effect of prior factual beliefs on reasoning; that is, holding constant people's political group identity (Causal Path 2)" (ibid., p.5), respectively. That is, they tested whether people show PMR after "statistically controlling for people's relevant prior factual beliefs" (ibid., p.13).

The results from Tappin and colleagues' studies are clear, they "found little evidence to suggest that cognitive sophistication magnifies a direct effect of political group identity on reasoning [but] fairly consistent evidence to suggest that cognitive sophistication magnifies a direct effect of prior factual beliefs on reasoning" (ibid.). These results call into question the evidence provided by those studies that suggest that cognitive sophistication causes higher levels of PMR. According to Tappin et al., the evidence from such studies can only show that cognitive sophistication magnifies the influence of prior factual beliefs on reasoning. But, as was pointed out above, prior factual beliefs are not necessarily caused by people's political group identity; it might be the case that both of them are caused by a third variable – in the case that there is a correlation between them – or that prior factual beliefs are a proxy for one's political information environment. Thus, and to reiterate, evidence for PMR can only come from experiments that clearly control for the distinction between political group identity and prior factual beliefs. So far, at least in Tappin and colleagues' view, such evidence is lacking.

Second criticism

However, the usual conflation of people's political group identity and their prior factual beliefs is not the only design element of PMR studies that the research group formed by Tappin, Pennycook, and Rand have criticized. In the article *Bayesian or biased? Analytic thinking and political belief updating* (Tappin et al., 2020b), the authors point out two other features of studies concluding that cognitive sophistication magnifies PMR that might weaken such a hypothesis: (1) such studies arrive at that conclusion without specifying how politically *unbiased* reasoning should look like, and (2) they focus on people's interpretation of new information instead of on how such information affects their beliefs.

Tappin and colleagues' goal is to test whether cognitive sophistication does indeed magnify politically biased reasoning if the experiment's design does not have any of the three previous elements. Thus, they run two studies "in which people receive noisy but informative information about the truth or falsity of factual political questions (Hill, 2017). In this design, [they] measure people's prior beliefs about the questions, and define (Study 1) or measure (Study 2) their perception about the informativeness of the information. Based on these data, [they] calculate the posterior beliefs that are expected according to Bayes' rule. [They] then compare individuals observed posterior beliefs to this Bayesian benchmark; evaluating the direction and extent to which their posterior beliefs diverge from the benchmark as a function of the political favorability of the new information (i.e., whether it is favorable or unfavorable for the stated political affiliation)" (ibid., p.2). This design has some advantages over those that focus on people's evaluation of new information alone. I will not list all of them, but it is worth mentioning that the new design allows the authors to calculate a *barebone Bayesian* benchmark that accounts for people's prior beliefs. And, while Tappin and colleagues alert us that their way of establishing the Bayesian benchmark is only one of the possibilities, having one is very important since it offers them the chance to compare it with people's (un)biased actual reasoning.

Unsurprisingly at this point, the results of the two experiments carried out by Tappin, Pennycook, and Rand found no evidence for the hypothesis that cognitive sophistication magnifies political bias of posterior beliefs. Indeed, they did find evidence for the alternative hypothesis that cognitive sophistication is correlated with posterior beliefs that are closer to those predicted by the Bayesian benchmark. In Tappin and colleagues' words: "[...] while we observed fairly consistent evidence to suggest that higher CRT scorers updated more (less) on politically favorable (unfavorable) signals than lower CRT scorers, the Bayesian benchmark implies that these patterns offer little evidence of magnified political bias. This is because subjects who scored lower on the CRT tended to report posterior beliefs that exceeded the benchmark on politically *un*favorable information but fell short of the benchmark on politically favorable information; a

seeming anti-political bias. The patterns of posterior beliefs among higher CRT scorers, by contrast, mostly resulted in a correction of this tendency observed among lower CRT scorers; thus, resulting in more normative, rather than biased, posterior beliefs" (ibid., p.10). All in all, these results took the authors to conclude that more empirical research is needed to understand why "the most cognitively sophisticated opposing partisans often disagree most strongly over various factual political questions" once they have discarded the possibility that this is due to a correlation of analytical thinking and PMR.

These two lines of experiments led the same research group to investigate a different hypothesis, this time not about society's polarization, but about the spread of misinformation online, which in turn "leads to inaccurate beliefs and can exacerbate partisan disagreement over even basic facts" (Pennycook et al., 2021, p.590): people share such content because they do not pay attention to accuracy when scrolling through their news feeds. As we shall see below, such a hypothesis follows the line explored in Kahan (2013) under the name of Bounded Rationality Position, which explains that the sharing of misinformation is due to the use of heuristics – a lack of proper reasoning. It also resonates with what is known as the *deficit model*, which blames cognitive limitations or ignorance "for the belief in the implausible or the irrational" (Levy and Ross, 2021) – opposed to PMR models for which such beliefs are rational as long as they advance people's interests.

3.2.1 The approach

In *Shifting attention to accuracy can reduce misinformation online*, Pennycook and colleagues (2021) run a series of experiments to test three different theories that try to explain why people share misinformation online: (i) people might be confused about what is truth, (ii) people have preferences other than accuracy, for example, for partisanship, and (iii) people do not pay attention to accuracy when they are deciding what to share online. Starting with the confusion account, this theory simply suggests that "people share misinformation because they mistakenly believe that it is accurate" (ibid., p.590). While there might be several explaining why they have those mistaken beliefs, a prominent one is politically motivated reasoning. As we saw extensively above, people engaging in PMR involuntarily form wrong beliefs as a way of protecting their group identity. To analyze whether solely confusion about the veracity of the information can explain the share of misinformation, Pennycook and colleagues run an experiment in which people were presented with several news pieces, half of them being true and the other half false while also evenly divided in their political leaning. Some of the participants were asked to assess the veracity of the information while the rest were asked whether they would share the news online. The results of the experiment showed that while in the accuracy condition, people were more predominantly able to tell apart true from false headlines, in the sharing condition whether the news aligned with their political identity was a better predictor of sharing intentions. For example, while 51.1%

of the Republicans in the study would share the headline "Over 500 'Migrant Caravaners' Arrested With Suicide Vests", only 15.7% rated it as accurate. This means that, in this case, the confusion account could only explain 30.72% of the shared headlines (ibid., p591).

The preference-based account of misinformation sharing would explain the discordance between accuracy judgments and sharing intentions by claiming that value some factors much more than they do accuracy, for example, partisanship. In other words, since accuracy might not be the only – nor main – driver of sharing intentions, people are willing to knowingly share misinformation if doing so promotes other more important values. However, this hypothesis seems at odd with people's stated preferences at the end of the studies, where they claimed that sharing only accurate content is 'extremely important' (ibid.). To incorporate such preferences in a theory that seeks to explain the sharing of misinformation, Pennycook and colleagues introduced the inattention-based account, "in which (i) people do care more about accuracy than other content dimensions, but accuracy nonetheless often has little effect on sharing, because (ii) the social media context focuses their attention on other factors such as the desire to attract and please followers/friends or to signal one's group membership" (ibid.).

In order to empirically tell apart the two theories, Pennycook and colleagues run a survey experiment where participants were asked about their sharing intentions. The participants were divided into two groups: while those in the control condition had simply to say how likely they were to share the headlines, those in the treatment condition had to do the same *after* rating the accuracy of a non-partisan headline – that is, "with the concept of accuracy more likely to be salient in their minds" (ibid.). Unsurprisingly, the results of the experiment showed that "participants in the treatment group were significantly less likely to consider sharing false headlines compared to those in the control group [...] but equally likely to consider sharing true headlines" (ibid.).

In the last survey experiment, Pennycook and colleagues looked for quantifying the relative contribution of each account. To do so, they carried out again an experiment like the one just explained but adding a 'full attention' treatment in which participants had to rate the accuracy of the headline before showing their sharing intentions. With such a treatment, the authors were able to quantify the percentage of participants that mistakenly thought to be sharing truthful information, 33.1%. On the other hand, 51.2% of the sharing intentions for false headlines were explained by the inattention-based account and only 15.8% of sharing by the preference-based account. From such results, the authors concluded that "inattention does not merely operate on the margin, but instead has a central role in the sharing of misinformation in the experimental paradigm" (ibid., p.592).

Lastly, Pennycook and colleagues tested whether the hypothesis of their inattention-based account travels outside the lab and carried out a digital field

experiment on Twitter. There, they sent private messages to users that had previously shared links to famous misinformation websites asking them to rate the accuracy of a non-political headline. Later on, the authors compared the veracity of the posts shared after the intervention with those posted before and found that "the single accuracy message made users more discerning in their subsequent sharing decisions" (ibid., p.593). While the arguments made in this thesis do not rest on the efficacy of specific interventions, the one just introduced will be very illustrative when we analyze how nudges would look under the assumption that the inattention-based account is correct.

In this section, we have followed Pennycook and colleagues (2021) to introduce the inattention-based account in its own terms. Such an account aims to explain the sharing of misinformation online by adducing that people do not pay attention to accuracy when they have to decide what to share in their timelines – greatly due to the design of social media platforms. As we saw as well, this account is opposed to the one considered in detail in section 3.1, politically motivated reasoning, which has in turn been extensively criticized by the same group of researchers that proposed the former account. In the next section, we will translate the inattention-based account into the three-stage dual-process language developed in section 2.4, to then analyze whether nudges are methodologically possible in section 3.2.3.

3.2.2 Translation

In the previous section, we introduced the inattention-based account in its own (practical) terms. As is evident, in doing so, we did not make use of any dual-process framework. However, this fact should not be seen as implying that such an account has not been thought of in those terms – nothing further from the truth. Indeed, at the end of section 3.2, we already suggested that the inattention-based account is in some ways related to theories like for example, Kahan's bounded rationality position – which, as shown in section 3.1.1, is explicitly formulated in dual-process language. There, we saw that the bounded rationality position takes the *heuristic-driven information processing* as the driver in the dynamic with motivated reasoning when trying to explain people's polarization over questions of fact. And, in simple terms, Kahan defined such information processing as one in which people over-rely on "the heuristic-driven, 'System 1' reasoning style" (Kahan, 2013, p. 408), at the expense of more effortful 'System 2' style of reasoning with which would be possible to avoid biased processing. In the same vein, we could understand the inattention-based account as hypothesizing that the design of social media platforms encourages people to engage in heuristic-driven Type 1 reasoning instead of an 'attentive to accuracy' and analytic reasoning that would make them behave according to their stated preferences. But, as should be evident at this point, such a simple formulation is not satisfactory for our interests in this thesis; a translation into the previously developed three-stage dual-process language will follow.

The main claim of the inattention-based account is that the decision of whether or not to share (mis)information is made without attending to its accuracy. To translate this account into our language, the key will naturally be to get a clearer understanding of what 'not attending to accuracy' means. Thus, in the first stage of the reasoning process, we would have that reading a headline cue in the user some intuitive Type 1 responses. It would be tempting to conclude that, given that we assumed that the person does not pay attention to accuracy, accuracy is not a feature of any of the intuitive responses. However, this would be a mistake since the only requirement that we need is that 'attention to accuracy' is not acted upon; that is, accuracy might be one of the features of one of the intuitive responses, but such a response would be shone upon by another – quicker – intuitive response. In other words, we have, at least, two candidates for the first reasoning stage, **T1p?** if accuracy is not represented by any intuitive response, and **T1pq?** in the case that accuracy is embodied by the intuitive response **q**.

The second reasoning stage is the one in which Type 2 processes monitor the existence of conflict between intuitive responses. In this case, if we had **T1p?** in the first stage, then no conflict can be detected, and response **p** can make it through without opposition. On the other hand, if **T1pq?** is the case but we assumed that accuracy plays no role in the final decision on whether to share misinformation, then we seem to have to conclude that the response **p** comes to mind so much quicker than response **q** that no conflict is detected between them by the pertinent Type 2 process. Whatever path is taken, none of them would call for more Type 2 processing in the third reasoning stage, and response **p** is enacted. However, one could claim that rationalization and inattention to accuracy are not contradictory processes. Under this view, it would be possible for Type 2 processes to detect the conflict in **T1pq?** so that more Type 2 processing is required in the third stage – as long as this extra Type 2 processing engages in the rationalization of **p**. As has been already pointed out in the thesis, the difference between rationalizing Type 2 processes and quick Type 1 ones might be extremely context-dependent. Therefore, we cannot completely rule out the possibility that the inattention-based account takes the form **T1pq?T2Rp**.

Thus, the only option that seems safe to discard is **T1pq?T2Dq**. In such a state, the conflict detected in the second stage enquires more Type 2 reasoning processes, but this time of the cognitive decoupling kind that would support the *accurate* response, **q**. Assuming that cognitive decoupling Type 2 processing would arrive at such a conclusion looks like the logical step given (1) the results of experiments mentioned above that show that people are competent in telling apart true from false headlines (as long as they have a minimum of relevant information about the topic at hand), and (2) people's stated preference for sharing only truthful information.

3.2.3 Nudges

Similarly to what we did in section 3.1.3 for the case of politically motivated reasoning, now it is time for analyzing whether it is possible to introduce nudges under the assumption stating that people share online misinformation because they do not pay attention to the accuracy of the content seen, possibly due to how social media platforms are designed. To do so, let's see what the four assumptions of the framework developed in section 2.4.2 would look like.

3.2.3.1 Initial state of mind: T1p?, T1pq?, or T1pq?T2Rp

As we have seen in the previous section, 3.2.2, there are three possibilities when it comes to understanding sharing online misinformation due to a lack of attention to accuracy in our dual-process language, **T1p?**, **T1pq?**, and **T1pq?T2p**. In them, while the proposition **p** refers to the intuition cued by the piece of misinformation favoring its sharing (the reason for which is not important to us at this point, but we could think for example of *familiarity bias* in the case that the user has encountered the information before), proposition **q** represents the intuition calling for paying attention to accuracy before taking any action with the piece of misinformation. Thus, in the cases **T1p?** and **T1pq?**, we have that the intuition **p** is acted upon because of the speed at which is cued, regardless of any other intuition triggered – including that about accuracy, **q**. Alternatively, in the case in which proposition **q** comes to mind fast enough so that more Type 2 process is demanded, **T1pq?T2Rp**, this extra processing could only be of the rationalizing type. Otherwise, if the user would engage in Type 2 processing of the decoupling kind, given the assumptions stating that the user is capable of discerning between true and false information and her preferences for sharing only truthful news, it would not be possible that the proposition enacted is **p**.

3.2.3.2 Intervention: T1qp?

Since we established in section 2.4.2.4 that nudges play with the rationalization of the intended *prudent* intuitive response, then a nudge that looks for people to believe/share true information needs to bring to the forefront the proposition that would favor truthfully crediting the information, **q**, making it the quickest of the intuitive responses. This intervention, of course, should not eliminate any of the other original intuitive responses, **p** and the *possible* **?**. Thus, **T1qp?**.

3.2.3.3 Nudge position: T1qp?T2

The nudge position assumption just aims to guarantee that the conflict between the propositions **p** and **q** is detected by the monitoring Type 2 processing. Detecting the conflict implies that more Type 2 processing will be requested in the third stage of the reasoning process, where the individual would have a chance to still wave away the proposition to which she has been nudged, **q**.

3.2.3.4 Choice position: T1qp?T2Rq

Before describing the choice position in more detail, note that the intervention, the nudge, and the choice position are exactly the same as in the case of politically motivated reasoning. This should not be surprising if we take into account that, fundamentally, a nudge consists of triggering the desired intuition, **q**, fast enough to be the quickest one but not so fast as to shine upon other intuitions and guaranteeing that, in the end, **q** is rationalized through Type 2 processes. What changes between different hypotheses about the reasons that make people share misinformation are, naturally, the initial state of mind and whether the desired intuition **q** is in each case one that could be rationalized in the choice position.

Here, similarly to the PMR hypothesis, we arrive at the choice position assumption in which the individual, if the nudge has worked as intended by the policy-maker, would rationalize the proposition **q**, processing the (mis)information in a truthful way, **T1qp?T2Rq**. What is up for discussion is whether q can be indeed rationalized. As I see it, there are two possibilities, each of them pulling in a different direction. On the one hand, we could understand **q** just as an intuition favoring the proposition at which the user would arrive if she would analyze its veracity. On the other hand, we could interpret **q** as an intuition that alerts the user about the importance of analyzing the accuracy of the information. In such a scenario, more than an intuition that can be rationalized, the intuition **q** might also entail a call for the user to engage in a decoupling process, **T1pq?T2Rp**, which of course would not be a nudge under our definition.

3.3 Concluding remarks

In this chapter, our goal has been to apply the framework developed in the previous chapter to the context of online misinformation. In particular, we have limited our analysis to two hypotheses that explain the sharing of misinformation: politically motivated reasoning and people's lack of attention to accuracy. For both hypotheses, we have followed the same strategy: first, an introduction of the hypothesis in their own terms, followed by a translation of the hypotheses into the dual-process language that we developed in the previous chapter to then the application of the definition of nudge.

From the results of the application of the framework to the two hypotheses, we cannot *conclusively* assert that it is possible *successfully* introduce nudges in any of the two scenarios. We have seen that while for the case of politically motivated reasoning, it would depend on the kind of PMR shown (either as a consequence of the rationalization of the response or an analytical decoupling of it), for the lack of attention to accuracy hypothesis, whether the intervention is a nudge would depend on whether the intuition triggered makes the person to rationalize sharing truthful information or on the other hand it leads to a decoupling Type 2 process. These results could be interpreted as discouraging the nudge endeavor altogether (after all, as Heilmann (2014, p.92) concluded, the cognitive requirements for the introduction of nudges are not easily met); however, as I see them, they should be better seen as a call for further research about the different understanding of the hypotheses considered here as well as the application of the developed framework to other theories explaining the sharing of misinformation.

Chapter 4

Conclusions

In this thesis, I had the ultimate goal of assessing whether the cognitive research about the spread of online misinformation grants the introduction of nudges in order to curtail it. In order to be able to answer the research question, I divided the task into two steps:

1. Developing a conceptual framework that specifies in a detailed way the cognitive requisites for the introduction of nudges.
2. Apply the framework to assess whether nudges are possible under two competing theories about the cognitive causes of the spread of misinformation.

I devoted Chapter 2 to the construction of the conceptual framework. Such a framework had as its starting point the one developed by Heilmann (2014). In his work, Heilmann defined nudges by clarifying the cognitive assumptions behind them (as originally understood by Thaler and Sunstein (2008)) in the following way:

1. Initial state of mind: **AqRp**
2. Intervention: **Aq** $\rightarrow$ **Ap**
3. Nudge position: **ApRp**
4. Choice position: **ApRp** leads to choice according to **p**

Importantly, Heilmann distinguishes between the assumptions themselves and the dual-system language in which they are expressed. Thus, once I introduced Heilmann's framework in sections 2.1.1, 2.1.2, and 2.1.3, I went on and presented what I understood as its main limitations: the simplified dual-system language may lead to ambiguities regarding the understanding of the working and relationship between the two systems. Given the possibility that in some of the contexts in which nudges could be implemented the research about the cognitive causes for the behavior that is meant to be altered has used dual-process

theories that are more complex than the one employed in Heilmann's framework, I decided to switch the dual-system language in the framework while keeping the assumptions so that it could be more fruitfully applied in those contexts. This is, of course, not to say that the following way of amending is the only nor best way of doing it: its validity and appropriateness would depend both on the overall validity of the dual-process paradigm (which, as mentioned, has been called into question) and the particular context of application.

To be able to modify Heilmann's framework so that the assumptions are preserved, I first contextualized Heilmann's take on the dual-system approach following Evans and Stanovich (2013) in section 2.2. There, we not only saw that the literature seems to have moved away from talking of systems and instead uses *processes* but also that the ambiguities found in Heilmann's framework can be traced back to two different accounts of cognitive architecture: parallel-competitive and default-interventionist. In section 2.3, I then introduced a model of human reasoning that aims at reconciling both accounts (Pennycook, Fugelsang, and Koehler, 2015). Such a dual-process model breaks down reasoning processing into three stages that clearly specify the ways in which Type 2 processes are triggered and how they relate to Type 1 processes to produce a final response. It is important to remind here that the decision of supplementing Heilmann's framework with this particular three-stage model is just one of the possibilities, other models are of course available, both within and outside of the dual-process approach. Likewise, it is also worth mentioning that the present attempt is only a first exploration, and more work is required to fully develop the integration of Pennycook and colleagues' model into Heilmann's framework. Nonethless, with the considerations of the last two sections, I was in disposition in section 2.5 of modifying Heilmann's framework such that the four assumptions are now like this:

1. Initial state of mind: **T1q?**
2. Intervention: **T1q? → T1pq?**
3. Nudge position: **T1pq?T2**
4. Choice position: **T1pq?T2Rp** leads to choice according to **p**

This definition of nudge contains an important novelty since it makes them depend on the *rationalization* of the desired behavior. The reason for this move is that rationalizing the desired response seems to be the only way of achieving it while, at the same time, giving the person the opportunity to resist the nudge and not spend the time and effort to think thoroughly about the decision at hand.

Chapter 3 consists of the application of the previous framework in the context of online misinformation. However, as mentioned, given the limited space available and the ever-growing number of hypotheses explaining the cognitive

causes for the sharing of online misinformation, I restricted the study to two of them: the one that blames it on people engaging in *politically motivated reasoning*, and that that postulates people's lack of attention to accuracy when they are using social media.

In section 3.1, I focused on politically motivated reasoning. There, I started by presenting the hypothesis in its own terms (Kahan, 2013, 2016a, 2016b), defining it as 'the tendency of individuals to unconsciously conform assessment of factual information to the protection of their identity as members of a community, regardless of its truth'. To make things clearer, we also followed Kahan (2017) and compared PMR with a Bayesian mode of information processing and with confirmation bias, focusing on how they diverge in the way in which they establish the *likelihood ratio* of the new information. Moreover, we also saw that another defining characteristic of PMR is that those that have shown higher levels of PMR in certain experiments are those who also scored higher in tests measuring their analytical skills. The goal of section 3.1.2 was then to explore how PMR can be expressed in the dual-process language developed in the previous chapter in order to apply the definition of nudge in section 3.1.3. Thus, we have that the assumptions behind such a nudge would be:

1. Initial state of mind: **T1pq?T2Rp** or **T1pq?T2Dp**
2. Intervention: **T1qp?**
3. Nudge position: **T1qp?T2**
4. Choice position: **T1qp?T2Rq**

From this description of what a nudge under PMR looks like, we conclude that its success would depend on the kind of PMR shown by the user. If she engages in the type of PMR that makes her rationalize the response that aligns with that of her community, then nudging her towards another response would only be effective if the nudge does not trigger her into a decoupling Type 2 process, which would render the intervention into something other than a nudge according to our definition. On the other hand, if the user shows PMR after decoupling her intuitive responses, nudging her towards another response would not seemingly work.

Finally, section 3.2 is devoted to the study of the lack of attention hypothesis. This section starts by briefly introducing two criticisms of PMR stating that the empirical evidence in its favor might not be as robust as previously thought (Tappin, Pennycook, and Rand, 2020a, 2020b). Then, I followed the same strategy as in the previous study case, first presenting the inattention account in its own terms to then translate it into the dual-process language before applying the definition of nudge to it. Thus, in a nutshell, according to the proponent of this hypothesis, Pennycook and colleagues (2021), the reason for people to share misinformation is that the *architecture* of social media distracts them

from paying attention to accuracy – even though experiments have shown that people have the ability to tell true from fake news, and they claim to have strong preferences for sharing only truthful information. Or, in the dual-process language, people share misinformation because they either act upon very fast intuitions (Type 1 processes) or rationalize the quickest of them. This would imply that the assumptions behind a nudge would be:

1. Initial state of mind: **T1p?**, **T1pq?**, or **T1pq?T2Rp**
2. Intervention: **T1qp?**
3. Nudge position: **T1qp?T2**
4. Choice position: **T1qp?T2Rq**

However, as in the case of PMR, the introduction of nudges might not be straightforward. While they would be successful if the intuition promoted by the nudge is just the one that promotes sharing truthful information, it might not be the case if the intervention consisting of 'making accuracy more salient' leads people to engage in decoupling processes.

With this, we arrive at the end of the thesis. While technically speaking we have seen that it might be *coherently* possible to introduce nudges to curtail online misinformation, it is important to remark that this could be extraordinarily hard to achieve. This is so not only because, as Heilmann (2014, p. 92-93) concludes, nudges' assumptions are difficult to meet – generally speaking – but also because some ambiguities remain when we try to interpret the hypotheses about the cognitive causes of the spread of misinformation through the lenses of the developed dual-process framework. Taking this into account, I cannot conclude but reinforcing the idea that the present thesis is only a first and tentative step and that more theoretical and practical work is needed in order to arrive at a more robust answer to the research question. Such future work would need to explore both parts of the thesis, the general framework and its application to the case of online misinformation. For example, one of the potential avenues of research could try to provide a characterization of the cognitive assumptions behind nudges using a different dual-process model to the one employed in the thesis, or even one that does not follow the dual-process paradigm. Another possible future path might explore the application of the (new) framework to cases in which the spread of online misinformation could be explained by making use of theories other than politically motivated reasoning and lack of attention. As can be seen in the recently published handbooks of political epistemology, such potential theories abound (Edenberg and Hannon, 2021; Hannon and de Ridder, 2021). That being said, I hope that the present thesis can contribute to framing future research in this recent interdisciplinary issue of interventions fighting online misinformation.

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. Journal of economic perspectives, 31(2), 211-236.

Berinsky, A. J. (2017). Rumors and health care reform: Experiments in political misinformation. British journal of political science, 47(2), 241-262.

Bovens, L. (2009). The ethics of nudge. Preference change: Approaches from philosophy, economics and psychology, 207-219. Brown, É. (2021). Regulating the spread of online misinformation. In The Routledge handbook of political epistemology (pp. 214-225). Routledge.

Chen, S., Duckworth, K., & Chaiken, S. (1999). Motivated heuristic and systematic processing. Psychological Inquiry, 10(1), 44-49.

Daniel, D. B., & Klaczynski, P. A. (2006). Developmental and individual differences in conditional reasoning: Effects of logic instructions and alternative antecedents. Child Development, 77(2), 339-354.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. PloS one, 6(1), e15954.

De Neys, W., & Franssens, S. (2009). Belief inhibition during thinking: Not always winning but at least taking part. Cognition, 113(1), 45-61.

De Neys, W., Vartanian, O., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. Psychological Science, 19(5), 483-489.

Dunning, D. (2003). The relation of self to social perception. In M. R. Leary J. P. Tangney (Eds.), Handbook of self and identity (pp. 421–441). The Guilford Press.

Edenberg, E., & Hannon, M. (Eds.). (2021). Political epistemology. Oxford University Press.

Evans, J. S. B. (2007). On the resolution of conflict in dual process theories of reasoning. Thinking Reasoning, 13(4), 321-339.

Evans, J. S. B. (2009). How many dual-process theories do we need? One, two, or many?. In J. S. B. Evans K. Frankish (Eds.), In two minds: Dual processes and beyond (pp. 31-54) Oxford, England: Oxford University Press.

Evans, J. S. B. (2010a). Intuition and reasoning: A dual-process perspective. Psychological Inquiry, 21(4), 313-326.

Evans, J. S. B. (2010b). Thinking twice: Two minds in one brain. Oxford University Press.

Evans, J. S. B., Newstead, S. E., Allen, J. L., & Pollard, P. (1994). Debiasing by instruction: The case of belief bias. European Journal of Cognitive Psychology, 6(3), 263-285.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. Perspectives on psychological science, 8(3), 223-241.

Gastil, J., Braman, D., Kahan, D., & Slovic, P. (2011). The cultural orientation of mass political opinion. PS: Political Science Politics, 44(4), 711-714.

Gigerenzer, G. (2002). Adaptive thinking: Rationality in the real world. Oxford University Press on Demand.

Gigerenzer, G. (2010). Personal Reflections on Theory and Psychology. Theory & Psychology, 20(6), 733–743. https://doi.org/10.1177/0959354310378184

Greifeneder, R., Jaffe, M., Newman, E., & Schwarz, N. (2021). The psychology of fake news: Accepting, sharing, and correcting misinformation (p. 252).

Grüne-Yanoff, T. (2012). Old wine in new casks: libertarian paternalism still violates liberal principles. Social Choice and Welfare, 38(4), 635-645.

Grüne-Yanoff, T. (2016). Why behavioural policy needs mechanistic evidence. Economics & Philosophy, 32(3), 463-483.

Grüne-Yanoff, T., & Hertwig, R. (2016). Nudge versus boost: How coherent are policy and theory?. Minds and Machines, 26(1), 149-183.

Guala, F., & Mittone, L. (2015). A political justification of nudging. Review of philosophy and psychology, 6(3), 385-395.

Handley, S. J., & Trippas, D. (2015). Dual processes and the interplay between knowledge and structure: A new parallel processing model. In Psychology of learning and motivation (Vol. 62, pp. 33-58). Academic Press.

Hannon, M., & de Ridder, J. (Eds.). (2021). The Routledge handbook of political epistemology. Routledge.

Hausman, D. M., & Welch, B. (2010). Debate: To nudge or not to nudge. Journal of Political Philosophy, 18(1), 123-136.

Heilmann, C. (2014). Success conditions for nudges: a methodological critique of libertarian paternalism. European Journal for Philosophy of Science, 4, 75-94.

Hill, S. J. (2017). Learning together slowly: Bayesian learning about political facts. The Journal of Politics, 79(4), 1403-1418.

Horne, B. D., Gruppi, M., & Adali, S. (2019). Trustworthy misinformation mitigation with soft information nudging. In 2019 First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA) (pp. 245-254). IEEE.

Hundley, A. C. (2017). Fake news and the first amendment: How false political speech kills the marketplace of ideas. Tul. L. Rev., 92, 497.

Jost, J. T., Hennes, E. P., & Lavine, H. (2013). " Hot" political cognition: Its self-, group-, and system-serving purposes.

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. Judgment and Decision making, 8(4), 407-424.

Kahan, D. M. (2016a). The politically motivated reasoning paradigm, Part 1: what politically motivated reasoning is and how to measure it Emerg. Trends Soc. Behav. Sci.

Kahan, D. M. (2016b). The politically motivated reasoning paradigm, Part 2: Unanswered questions. Emerging trends in the social and behavioral sciences: An interdisciplinary, searchable, and linkable resource, 1-15.

Kahan, D. M. (2017). Misconceptions, misinformation, and the logic of identity-protective cognition. Cultural Cognition Project Working Paper Series No. 164, Yale Law School, Public Law Research Paper No. 605, Yale Law & Economics Research Paper No. 575.

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. Behavioural public policy, 1(1), 54-86.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. American economic review, 93(5), 1449-1475.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. Heuristics and biases: The psychology of intuitive judgment, 49(49-81), 74.

Kahneman, D., Tversky, A. (1979). Prospect Theory: an Analysis of Decisions under Risk Econometrica 47.

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. Perspectives on psychological science, 4(6), 533-550.

Kruglanski, A. W., Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. Psychological review, 118(1), 97.

Levy, N., & Ross, R. M. (2021). The cognitive science of fake news. In The Routledge handbook of political epistemology (pp. 181-191). Routledge.

Lilienfeld, S. O., Ammirati, R., & Landfield, K. (2009). Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare?. Perspectives on psychological science, 4(4), 390-398.

Lindeman, M., & Aarnio, K. (2007). Superstitious, magical, and paranormal beliefs: An integrative model. Journal of research in Personality, 41(4), 731-744.

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. Psychological bulletin, 127(2), 267.

Nam, H. H., Jost, J. T., & Van Bavel, J. J. (2013). "Not for all the tea in China!" Political ideology and the avoidance of dissonance-arousing situations. PloS one, 8(4), e59837.

Nieminen, S., & Rapeli, L. (2019). Fighting misperceptions and doubting journalists' objectivity: A review of fact-checking literature. Political Studies Review, 17(3), 296-309.

Nurse, M. S., & Grant, W. J. (2020). I'll see it when I believe it: Motivated numeracy in perceptions of climate change risk. Environmental Communication, 14(2), 184-201.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bull-

shit. Judgment and Decision making, 10(6), 549-563.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. Nature, 592(7855), 590–595.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). What makes us think? A three-stage dual-process model of analytic engagement. Cognitive psychology, 80, 34-72.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. Cognition, 188, 39-50.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. Trends in cognitive sciences, 25(5), 388-402.

Pennycook, G., & Rand, D. G. (2022). Nudging social media toward accuracy. The Annals of the American Academy of Political and Social Science, 700(1), 152-164.

Ross, R. M., Rand, D. G., & Pennycook, G. (2021). Beyond "fake news": Analytic thinking and the detection of false and hyperpartisan news headlines. Judgment and Decision making, 16(2), 484-504.

Russo, J. E., Carlson, K. A., Meloy, M. G., & Yong, K. (2008). The goal of consistency as a cause of information distortion. Journal of Experimental Psychology: General, 137(3), 456.

Sahlin, N. E., Wallin, A., & Persson, J. (2010). Decision science: from Ramsey to dual process theories. Synthese, 172, 129-143.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. Psychological bulletin, 119(1), 3.

Sloman, S. A. (2002). Two systems of reasoning.

Smith, E. R., & DeCoster, J. (2000). Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. Personality and social psychology review, 4(2), 108-131.

Stanovich, K. E. (1999). Who is rational?: Studies of individual differences in reasoning. Psychology Press.

Stanovich, K. E. (2004). Balance in psychological research: The dual process perspective. Behavioral and Brain Sciences, 27(3), 357-358.

Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory?.

Stanovich, K. (2011). Rationality and the reflective mind. Oxford University Press.

Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. Behavioral and brain sciences, 23(5), 701-717.

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. Journal of economic literature, 38(2), 332-382.

Sunstein, C. R. (2003). Terrorism and probability neglect. Journal of Risk and Uncertainty, 26, 121-136.

Sunstein, C. R. (2006). The availability heuristic, intuitive cost-benefit analysis, and climate change. Climatic change, 77(1-2), 195-210.

Sunstein, C. R. (2007). On the divergent American reactions to terrorism and climate change. Colum. L. Rev., 107, 503.

Sunstein, C. R., & Vermeule, A. (2008). Conspiracy theories.

Tappin, B. M., Pennycook, G., & Rand, D. G. (2020a). Thinking clearly about causal inferences of politically motivated reasoning: Why paradigmatic study designs often undermine causal inference. Current Opinion in Behavioral Sciences, 34, 81-87.

Tappin, B. M., Pennycook, G., & Rand, D. G. (2020b). Bayesian or biased? Analytic thinking and political belief updating. Cognition, 204, 104375.

Thaler, R. H., & Sunstein, C. R. (2009). Nudge. Penguin. Thompson, V. A. (2009). Dual-process theories: A metacognitive perspective. Oxford University Press.

Thornhill, C., Meeus, Q., Peperkamp, J., & Berendt, B. (2019). A digital nudge to counter confirmation bias. Frontiers in big data, 2, 11.

Waldman, A. E. (2017). The marketplace of fake news. U. Pa. J. Const. L., 20, 845.

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. Political Communication, 37(3), 350-375.

Weber, E. U. (2006). Experience-based and description-based perceptions of long-term risk: Why global warming does not scare us (yet). Climatic change, 77(1-2), 103-120.

Westen, D., Blagov, P. S., Harenski, K., Kilts, C., & Hamann, S. (2006). Neural bases of motivated reasoning: An fMRI study of emotional constraints on partisan political judgment in the 2004 US presidential election. Journal of cognitive neuroscience, 18(11), 1947-1958.

Zaller, J. (1992). The nature and origins of mass opinion. Cambridge University Press.