Topoi
https://doi.org/10.1007/s11245-024-10083-2

Regulating Misinformation: Political Irrationality as a Feasibility Constraint

Bartlomiej Chomanski[1]

Accepted: 26 July 2024

Abstract
This paper argues that the well-established fact of political irrationality imposes substantial constraints on how governments may combat the threat of political misinformation. Though attempts at regulating misinformation are becoming increasingly popular, both among policymakers and theorists, I intend to show that, for a wide range of anti-misinformation interventions (collectively termed "debunking" and "source labeling"), these attempts ought to be abandoned. My argument relies primarily on the fact that most people process politically-relevant information in biased and motivated ways. Since debunking or factual correction of politically relevant misinformation (as well as source labeling) themselves consist of providing politically-relevant information, they are also very likely to be processed in irrational ways. This makes it extremely difficult to effectively correct people's political beliefs and political information processing. Since governments should not pursue policies likely to be futile, they should refrain from mandating such interventions. My conclusion is of relevance to considerable literature in digital ethics of misinformation. It shows that many celebrated works in the field ignore political irrationality and fail to consider its implications.

Keywords Misinformation ethics · Social media ethics · Free speech · Political irrationality · Political ignorance

1 Introduction

In this paper, I argue that governments should not pass laws aiming to curb, limit, or eliminate the spread of questionable[1] information on social media, to the extent that these laws rely on debunking interventions such as fact-checking, and source reliability labels (SRL). Governments should also refrain from coercively intervening in political discourse with the aim of limiting ideological sorting.[2] I focus specifically on information that is politically relevant, in the broad sense of being treated by many of its consumers as having relevance to some political matter(s). Political matters include both matters of politics (elections, politicians' character etc.) and matters of policy (immigration policy, pandemic policy etc.). Unless otherwise specified, all mentions of "information" in what follows refer to politically relevant information[3]. I follow standard practice among misinformation researchers (Roozenbeek et al. 2023) and define *debunking* as providing corrective information concurrently with or subsequently to the alleged misinformation, and *source reliability labeling* as providing assessments (in the form of a prominent label) of the reliability of an information source alongside the information (e.g. a news story) the source presents. In practice, debunking encompasses a number of distinct interventions. For example, in addition to displaying a corrective message (e.g. "disputed by third-party fact-checkers") alongside the false or misleading posts, Meta, the owner of Facebook, Instagram, and WhatsApp, would generally also reduce the spread of such content (Meta 2024). X (formerly Twitter), in turn, relies on user-generated Community Notes displayed alongside disputed information, seeking to correct it or provide more context,

[1] I will use the general term "misinformation" to refer to such information. Misinformation, on my understanding, could be false, partially true, or true but misleading. I make no distinction between intentional and non-intentional misinformation.

[2] For recent work along similar lines, see e.g. Messina (2023); Gibbons (2023).

[3] For a definition of "political beliefs" along similar lines, see, e.g., Hannon and de Ridder (2021, 156).

✉ Bartlomiej Chomanski
b.chomanski@gmail.com

[1] Department of Philosophy, Adam Mickiewicz University, Poznan, Poland

Published online: 14 August 2024

but the company takes no further steps regarding the message's spread (X Corporation, 2024). YouTube's approach to fact-checks appears similar, except the corrections come from externally validated publishers, rather than being user-generated (Google 2024).

In this paper, I am concerned with debunking in the sense of providing corrective information, without imposing further sanctions on users; I leave aside the question of content removals and user bans.

In making my argument, I do not presuppose any particular normative theory (for example, I do not assume any substantive account of the value of free speech). I rely on appeals to empirical evidence and, when necessary, commonsense moral intuitions. Nevertheless, the argument has broader philosophical and practical import. Its conclusions run counter to celebrated works in digital ethics, and stand athwart a number of vigorously pursued policy proposals.

1.1 Motivating the case and Presenting the Arguments

I will start with an unofficial sequel to a story once told by Pigliucci (2018): In the original, Massimo, a distinguished philosopher, finds himself discussing some controversial matters with a relative of his, a conspiracy theorist by the name of Ostinato. The discussion covers a broad variety of issues, from the 9/11 attacks to links between vaccines and autism. Despite having truth, reason, and logic on his side, Massimo fails to get Ostinato to budge on any of the topics. Ostinato

> denied relevant expertise …, while at the same time vigorously—and apparently oblivious to the patent contradiction—invoking someone else's doubtful expertise …. He continually side-tracked the conversation, bringing up irrelevant or unconnected points … and insisting we should look "beyond logic," whatever that means. The usual fun. I was getting more and more frustrated, the wine was running out, and neither I nor Ostinato had learned anything or even hinted at changing our mind (2018, 7).

For the sequel, let us introduce a third protagonist - Prepotente, the mayor of the town where Massimo and Ostinato live.

Prepotente recognizes that Ostinato's views are not just mistaken; they might be downright dangerous (what if Ostinato acts on his crazy misinformed beliefs? What if he spreads them? What if he *votes*?), and it would be good for Ostinato to change them in line with the evidence. This much seems reasonable. But Prepotente then decides to intervene in the conversation in the most unusual way: he threatens *Massimo* with significant financial penalties unless he continues the conversations with Ostinato, on a regular basis, so that the latter's mind can finally be changed. If Massimo refuses, Prepotente will confiscate a substantial amount of money from him. He can even, when all else fails, throw Massimo in a cage.

To make matters starker, imagine one last thing: imagine it turns out that when it comes to issues such as the pseudoscience of vaccines, Ostinato is literally immune to evidence. No matter what Massimo says, no matter how many times he hears it, Ostinato's prior confidence in his belief will equal his posterior. For him, when it comes to matters of this sort, $p(A|B) = p(A)$.

Now, what should we make of Prepotente's actions in this story? It seems that Prepotente is acting unjustly. By using coercion to get Massimo to engage in what is essentially a fool's errand, Prepotente appears to violate Massimo's rights (a lawyer might say that Massimo's forced conversations with Ostinato amount to compelled speech), *for no benefit whatsoever*. This is unjustified. Even if Massimo has good reasons to generally follow Prepotente's other orders (including orders about what Massimo should say), this case is an exception.[4]

When governments insist on mandating misinformation interventions, such as fact-checks and SRLs (especially online), they mandate that social media companies (SMCs) and other content providers engage, like Massimo, in certain types of compelled speech. But if people are like Ostinato in the above story, the governments would be coercing SMCs for no benefit whatsoever. Since it was unjust for Prepotente to do this, it seems unjust for governments to do as well.

As it turns out, when it comes to political information, most of us seem to harbor the habits of thought that approximate Ostinato's. Thus, governments may not mandate (certain types of) misinformation interventions.

More explicitly:

(1) (Most) people are (epistemically) politically irrational (henceforth simply "politically irrational") – that is they

[4] One could worry that what drives the intuition condemning Prepotente's act is the mere presence of coercion rather than the futility of the coerced actions. But that would be too quick. For starters, many reasonable people would be willing to accept the permissibility of compelling speech at least in some cases (e.g. many reasonable people support truth-in-advertising laws; a priori, we should not want to condemn such laws merely because they amount to compelled speech); indeed, it's not unreasonable to think that in some instances, the government's use of coercive power to get *one party to police the speech of another* is legitimate. Employers are sometimes legally required to restrict expressive activity in the workplace, lest they be found liable for creating a hostile work environment. While some bemoan such laws, reasonable people can disagree on how well-justified they are and I do not want to prejudge the issue.

do not form their political beliefs in truth-conducive ways.

(2) If (most) people are politically irrational, then debunking questionable information in their information diets[5] will very likely not lead them to becoming better informed and make better-informed political choices.

(3) Therefore, debunking questionable information in people's information diets will very likely not lead them to becoming better informed[6] and make better-informed choices. (from 1 to 2)

(4) If governments are justified in passing a law, they need to have a good reason to think the law has a high probability of meeting its goals.

(5) Therefore, governments are not justified in passing laws aimed at curbing/removing questionable information from the citizens' information diets. (from 3 to 4)

I rely on empirical findings in political psychology to motivate premises (1) and (2). I rely on commonsense moral intuitions, and an appeal to public reason liberalism, to motivate premise (4).

If sound, the argument gives us a strong reason to oppose government interventions aimed at combating questionable information online. Suitably modified, it can also yield the conclusion that governments should not attempt to mitigate ideological *sorting* online (ideological sorting, or sorting for short, occurs when people tend to interact online primarily with those that share their political views). Assuming that a reduction in sorting - perhaps achieved by a reduction, or even banning, of microtargeted political ads, sometimes thought to be one of the culprits behind sorting - would mean expanded access to what psychologists call "counter-attitudinal" information (information that in some way goes against one's views), the second version of the argument is this:

(6) (Most) people are politically irrational.

(7) If most people are politically irrational, then expanding their access to a broader range of information in their online information diets will very likely not lead them to becoming better informed and making better-informed choices.

(8) Therefore, expanding people's access to a broader range of information in their online information diets will very likely not lead them to becoming better informed and making better-informed choices. (from 6 to 7)

(9) If governments are justified in passing a law, they need to have a good reason to think the law has a high probability of meeting its goals.

(10) Therefore, governments are not justified in passing laws aimed at adding information to the citizens' online information diets, such as laws preventing sorting. (from 8 to 9)

The assumption of political irrationality yields the prediction that actual attempts to debunk misinformation will fail to change people's minds. There is a growing body of empirical evidence that confirms this prediction. Debunking fake stories, especially in online environments, fails either to correct misinformed beliefs or to change evaluative attitudes previously based on misinformed beliefs (or both). It fails because people who process information irrationally are not epistemically benefited by improvements to the quality, quantity or diversity of information they receive. Moreover, if misinformation contributes to the adoption of more extreme attitudes, fact-checking, in virtue of its failure to correct belief in misinformation, will fail to prevent increased extremism. For the same reason, we can expect that exposure to alternative viewpoints will not lead people to becoming less extreme, less biased or better informed.

The argument has some implications for the philosophical literature on (how to deal with) social media misinformation and sorting. As we shall see, celebrated works in these areas – belonging to what I'll call the "digital ethics of misinformation" – tend to assume the negation of premise 1 and 6 in the above arguments, in order to make recommendations about what to do with the less-than-ideal conditions of political discourse on social media. These recommendations frequently include encouraging governments to institute policies restricting ideological sorting and curbing misinformation. If I am right, these demands are misguided.

Moreover, if I am right, then we should be pessimistic. Governments across the world – even those putatively committed to robust protections of freedom of speech – have either already employed a number of restrictive policies aiming to combat misinformation, or are planning to do more (see e.g., The European Commission's *Code of Practice on Disinformation* (2022)). These policies will *not* help us get better at deliberating online and will *not* improve our knowledge of complex political matters, nor will they help us better appreciate opposite viewpoints. So, real-world governments are likely already exceeding their legitimate

[5] For the purposes of this paper, "information diet" refers to the totality of politically relevant information a person consumes. The *quality* of a person's information diet refers to the proportion of non-misleading information in her information diet (the higher the proportion, the higher the quality). Misleadingness is context-dependent. For example, knowing the crime rate among new immigrants, or the fatality rate of some disease, when taken in isolation, could be misleading. When put in context (crime rates of other groups, fatality rates of other diseases), they could cease to be misleading.

[6] A person becomes better-informed when the proportion of her non-misleading beliefs increases. Choices are better-informed when a person choosing makes her choice at least on the partial basis of the new, better information.

functions by attempting to fight misinformation, and are also likely to continue doing so.

2 Defending the Premises

2.1 The Normative Premise

Premise 4 (9) can be justified by noting, first, that most government policies are coercively enforced. Second, in virtue of being harmful, coercion is presumptively unjustified. That is to say, whoever wishes to employ coercion must provide a good reason for doing so. Moreover, for coercion to be justified, it must at least have reasonable chances of achieving its goals. Huemer (2012) expresses this thought as follows:

> there is a kind of moral presumption against coercive interventions. Laws are commands backed up by threats of coercive imposition of harm on those who disobey them. Harmful coercion against an individual generally requires some clear justification. One is not justified in coercively harming a person on the grounds that the person has violated a command *that one merely guesses has some social benefit. If it is not reasonably clear that the expected benefits of a policy significantly outweigh the expected costs*, *then one cannot justly use force to impose that policy* on the rest of society (2012, 12, emphasis added).

The point seems intuitive (recall Prepotente from the previous section). Consider the following two cases:

RESCUE 1: A is drowning at sea. In order to save A, B decides to use a nearby boat, which belongs to C, and sail to A's rescue. B commandeers the boat, threatening to use violence against C if they were to refuse.

RESCUE 2: A is drowning at sea. In order to save A, B decides to use a reliquary, which belongs to C, and perform a magic ritual to calm the waters. B commandeers the reliquary, threatening to use violence against C if they were to refuse.

B's actions in RESCUE 1 seem justified. B's actions in RESCUE 2 don't seem justified. The only difference between them is that, in contrast to RESCUE 1, in RESCUE 2 it's extremely unlikely that the action B coerces C to perform will be effective in helping achieve the ethically justified goal of saving A. This can be analogized to what governments do. While government coercion may sometimes be justified, it may not be justified in cases where there is little to no chance of achieving the goal (however laudable).

There is in this regard a generally recognized asymmetry between coercive and non-coercive acts. Geoffrey Brennan and Loren Lomasky (2006), for example, say that:

> the onus of justification weighs much more heavily on coercive than on consensual activity. *Unless there is some overriding reason to coerce others*, *there is an overriding reason not to coerce.* Different theories of political authority will prompt different judgments concerning those circumstances in which coercion is justified, but *any remotely plausible theory will acknowledge that the justificatory bar is set considerably higher for force than for voluntary concurrence* (237, emphasis added).

Bare intuition and appeal to consensus are not the only ways to support this premise. In a wide-ranging exploration of the principles of public justification, Kevin Vallier (2021) argues as follows:

> we can only publicly justify coercion if members of the public have some way to convince one another that the policy in question will have certain effects and that the benefits of the policy will exceed the costs associated with lost opportunities for choice. I think this is arguably an implicit part of what Rawls called the "guidelines of inquiry," and I will call this part *policy epistemology* … Policy epistemology is critical for determining which public policies can be publicly justified, especially in the case of coercive regulations. Part of showing that persons have reason to submit to coercion is demonstrating that the coercion in question is an improvement according to each person's reflective perspective. [footnote omitted, emphasis in original]. (155–158)

On Vallier's account, a coercive act that cannot be demonstrated to have a reasonable chance of achieving its promised benefit is not publicly justified. If we're all political Ostinatos, then forcing us to consume (or forcing hapless Massimos to feed us) debunking information will not secure the benefit of expanding our political knowledge and improving our political information-processing. So, it will not be publicly justified.

All the foregoing favors the normative premise.

Still, one could object: perhaps the analogies are inapt. Coercive policies aimed at corporations do not carry the same normative significance as coercive policies aimed at individuals. When corporations are found in violation of the law, they typically have to pay a fine, sell off their assets, break up into smaller units, etc.; when individuals are, they could end up in prison, or worse. This is a significant moral difference not accounted for in my argument - after all, coercive policies I discuss will be aimed at corporations, not individuals.

However, recent practice of governments when it comes to policing internet speech belies this time-honored distinction. Real-world democratic governments appear to be willing to impose criminal sanctions (inclusive of imprisonment) on corporate personnel of various levels, from owners and CEOs to regular workers, for violations of government demands regarding online speech.

For one example, the executives of Rumble, an online video sharing site, risk criminal penalties in the UK for the company's failure to comply with the British governments' demands concerning what speech is "monetized" on the platform (Sellman 2023). For another, Brazil-based employees of X (formerly Twitter) were threatened with arrest for the company's failure to comply with the Brazilian government's demands for censorship (Chakraborty 2024). Thus, rank-and-file workers also seem to face threats of prosecution from governments dissatisfied with how their employers handle demands about online speech. Consequently, at least when it comes to the enforcement of various speech laws, analogizing them to coercive acts aimed at individuals seems apt.[7]

It could also be said that, though perhaps the laws are not likely to be particularly effective, they express a valuable signal: a government's (or even society's) commitment to a rational, fact-based public discourse. While this could be true, there are reasons to think that such symbolic virtues do not justify coercion (especially when combined with the law's practical ineffectiveness): first, while, to some people, forcing companies to correct political misperceptions in public discussion communicates the governments' commitment to improved public discourse, it's not unreasonable to view laws mandating fact-checks as expressions of epistemic paternalism, signaling, rather, a distrust in citizens' capacity to participate in democratic self-governance without oversight. The message sent by governments with laws like these seems ambiguous, which undermines their symbolic force.[8]

One can also wonder whether passing laws intended to achieve some goal, without caring whether they in fact do so (or while knowing, or having an obligation to know, that they probably won't), is an especially effective way of communicating one's commitment to that goal. Indeed, such cavalier attitude towards the law's outcomes undermines the message of commitment to the values the law is supposed to promote.[9]

Finally, there are non-coercive ways of signaling such a commitment. The government's expressive powers are vast and it seems possible for it to engage in a number of non-coercive actions that express endorsement of the value of reasoned debate, without threatening anyone with sanctions for failure to say what governments want them to say (I return to this point in the penultimate section).

2.2 The Empirical Premise

Premise 1 (6) is true in light of a vast literature on political psychology. That literature appears to converge on the finding that, when it comes to politics, we are all Ostinatos (nearly enough) or "partisan hacks" (Freiman 2021, p. 22 and *passim*). We process politically-relevant information irrationally. We are partisan and ignorant. We do not listen to reason, and frequently do not know (nor care to know) what policies our favored candidates endorse. We arrive at beliefs about politics in deeply biased and motivated ways. In a word, we fail to adhere to basic epistemic norms when it comes to forming political beliefs.

This is not to say, however, that such beliefs are formed haphazardly or randomly. Just because people are epistemically irrational about politics, it does not follow that they are irrational simpliciter.

In many theorists' view (Caplan 2007; Huemer (2016; Brennan (2016; Freiman (2021; Hannon and de Ridder (2021), Somin (2023) non-adherence to the norms of epistemic rationality might still be *instrumentally* rational, might still help us achieve our other goals - at least when it comes to politics. For when it comes to politics, Bryan Caplan

[7] Ultimately, it seems that getting corporations to obey laws would have to depend on getting individuals to obey commands from the state. If this is granted, then the moral strictures on coercion would apply to laws targeting corporations to roughly the same degree as they do to laws targeting individual behavior, since the chief mechanism for enforcement – coercion directed at individuals - is at work in both cases.

[8] For the same reason, it's not clear whether such legislation could succeed in setting the standards for public debate and fostering an environment where truth and accuracy are prized, regardless of short-term effectiveness. While it's possible that some reasonable people will see the demands for fact-checking as doing just that, other reasonable people may perceive such demands as fostering an environment of stifled discussion and elite distrust of common opinion. Since whether the standards are successfully set would depend on what the general public perceives the standards to be, it is unclear whether legally-required debunks would achieve this aim. Plus, there also seem to exist non-coercive ways of setting such standards.

[9] For an analogy, imagine I teach literature and want to instill love of Shakespeare in my students. To do so, I make them memorize long passages from *Hamlet*, and grade them exclusively on how well they do it. When told that this method is unlikely to make my students appreciate Shakespeare - in fact, it could lead them to resent his work instead - I reply that what really matters is that my method of teaching expresses my commitment to spreading the love of the Bard, and that I intend to make no changes to my teaching style. Such a reply seems to show that I am deeply unserious about my alleged commitment. It would also be odd to claim that my doing so fosters an environment and sets appropriate standards for the appreciation of fine literature.

says, “[b]eliefs that are irrational from the standpoint of truth-seeking are rational from the standpoint of individual utility maximization” (2007, 179). This is because the costs of being factually wrong about political matters are usually very low or externalized (when I vote on the basis of mistaken beliefs about some issue, my vote has a vanishingly small chance of changing the outcome, and even if I get my way, the costs of the policy my elected representatives pursue are borne by the entire society - see Joshi (2024) for more on this); on the other hand, the psychological and social benefits of holding beliefs endorsed by members of our political tribe (regardless of their truth) are significant. Being biased, unscientific, and tribalistic in forming political beliefs benefits us by way of signaling group-membership, solidifying our self-conception as decent human beings, and securing social approval - and does little harm. As Ilya Somin explains,

> when there are few or no negative consequences to error, it is rational to make little or no effort to control one’s biases. Thus, citizens routinely overvalue evidence supporting their preexisting [political] views while downplaying or ignoring anything that cuts the other way. These tendencies toward biased evaluations of information are significant and widespread among voters on both sides of the political spectrum. Many of the most attentive citizens tend to be highly biased “political fans.” *They follow politics closely for much the same reasons as sports fans follow their favorite teams: not to get at the truth*, *but to enjoy the camaraderie of their fellow fans*, *the process of cheering on their preferred political “team,” and – in many cases – detesting opposing “teams” (opposing parties and their supporters). There is nothing necessarily wrong with being a political or sports fan. But fan behavior is often at odds with truth-seeking. People who follow politics primarily to enhance their fan experience cannot objectively evaluate political information.* [2023, 289, emphasis added, references omitted]

In other words, in the realm of political belief, (most) people have strong incentives to forgo epistemic norms for belief-forming, and weak incentives to avoid untruths. Not all political beliefs fall into this category (e.g. the belief that there are 25 EU Member States would likely be easily corrected, regardless of one’s opinion of the institution), and some non-political beliefs do (even empirical beliefs - concerning, say, the efficacy of a vaccine, the lethality of a disease, or the main causes of crime - could become, under the right sorts of conditions, vehicles for partisan signaling). But insofar as there is a class of beliefs such that their falsehood does not harm us in achieving our goals, and holding them brings us non-epistemic benefits, we are incentivized not to be especially attentive to epistemic norms when forming them. It is rational to be irrational about such beliefs.[10]

In what follows, I choose to focus primarily on political beliefs as chief examples of this phenomenon, for three reasons: first, in so doing, I simply follow the practice of the theorists cited above who endorse some version of rational irrationality; second, political beliefs are often taken to be paradigmatic examples of rational irrationality; third, the correction of specifically political misinformation (or politically-relevant scientific misinformation) is frequently the main target of fact-checking and other forms of debunking.

The scientific evidence of widespread failures of epistemic rationality in forming political beliefs is substantial. Consider, for starters, this passage from Milton Lodge and Charles Taber (2007), summarizing their own work on political rationality:

> These studies show that [when it comes to political issues] people find it very difficult to escape the pull of their spontaneously evoked feelings. First, people simply feel that the information they agree with is stronger than the information with which they disagree. Second, when thinking about the evidence on a policy issue, people actively denigrate the information with which they disagree while accepting supportive information with little scrutiny. Third, people seek out confirmatory information and avoid evidence that might challenge their priors. Fourth, all of these biases conspire to drive attitudes further in the direction of priors the more they think and reason about the issues. Finally, all of these biases are particularly pronounced for citizens with more knowledge and stronger political attitudes, the very folks on whom democratic theory relies most heavily (35).

There is, of course, more. In a famous series of studies, Cohen (2003) gave ideologically sorted participants descriptions of two sorts of welfare policy, a “stringent” and a “generous” one, and asked them to pick the one they preferred. In one condition, participants were given no further information about the policies. In the other, they were told that their party opposed the ideologically congruent policy (i.e. that the Democrats *opposed* the generous policy or that the Republicans *opposed* the stringent one). Cohen found that

[10] I stress that I do not take this to impugn the general public’s intellectual capacities, nor to justify any form of paternalistic interventions ‘for the good of the ignorant masses’. Rather, it’s a simple recognition that people respond to incentives, and sometimes they’re highly incentivized not to take epistemic norms seriously.

> [e]ven under conditions of effortful [cognitive] processing, attitudes toward a social policy depended almost exclusively upon the stated position of one's political party. …. Nevertheless, participants denied having been influenced by their political group, although they believed that other individuals, especially their ideological adversaries, would be so influenced (808).

In yet another study, Dan Kahan and colleagues (2017) discover that when processing politically valenced information, the cognitive performance of even sophisticated participants significantly deteriorates when the information is discordant with the participant's values. As they summarize it,

> when policy-relevant facts become identified as symbols of membership in and loyalty to affinity groups that figure in important ways in individuals' lives, they will be motivated to engage empirical evidence and other information in a manner that *more reliably connects their beliefs to the positions that predominate in their particular groups than to the positions that are best supported by the evidence* (74, emphasis added).

This finding has been replicated in a number of other studies, such as Gampa et al. (2019), Calvillo et al. (2020), and Aspernäs et al. (2023).

Su (2022) finds similar results:

> people were more reluctant to update their beliefs for politically significant issues and … more likely to update their beliefs when they received information that aligned with their preexisting ideologies. Also, … providing subjects with ambiguous information caused them to be further divided based on their political ideology, although subjects may also discredit the ambiguous information altogether when it challenges their beliefs to a great extent (8).

These are all examples from individual studies, of course, but they seem representative of the literature as a whole. Consider these summaries:

> in studying the political psychology and behavior of citizens, every facet of the rational choice model appears to be violated to some degree. People prefer policies and engage in behavior such as voting that do not further their self-interest. Their preferences are often unstable, inconsistent, and affected by how alternatives are framed. They do not always respond to new information by updating their beliefs and modifying their preferences in accord with their goals. They do not gather enough information to make the optimal choice (Chong 2013, p. 96).
>
> Given the evidence cited so far, the question of whether political thought could ever amount to a normatively satisfying rational-choice process is probably not even a subject for debate (Taber and Young 2013, p. 546).
>
> The overwhelming consensus in political psychology, based on a huge and diverse range of studies, is that most citizens process political information in deeply biased, partisan, motivated ways rather than in dispassionate, rational ways. … Even [those] who lack strong ideologies… don't care enough about politics to form opinions, but if they started to care, they'd form opinions in biased ways (Brennan 2016, p. 32).
>
> the studies I've reviewed show that political partisans are generally unmoved by evidence. When confronted with information that threatens our side, we're masters at ignoring, downgrading, and discrediting it to maintain our partisan allegiance. Thus, even if the evidence points to our side being wrong, we'll continue to believe that we're right (Freiman 2021, p. 34).

Systematic biases and (rationally) irrational attitudes characterize how typical members of the electorate think about politics. Specifically, acquiring new information does not lead to a straightforward belief-updating one would expect if voters were rational. People respond to new information in ways that diverge from how they rationally *should* respond. In general, it seems more important for most to remain firm believers in an ideology than to have an accurate picture of political matters in light of the available evidence.

The foregoing suggests that interventions aiming at debunking misinformation or providing SRLs are unlikely to succeed. This is because information that some purported facts are not the way originally presented, or that a news story is missing context, or is put forward by a disreputable source, will be reasoned about in the same way that the target story was reasoned about: with the aim to maintain one's political identity, not to acquire a more accurate picture of how the world is and should be.[11] It also suggests that removing barriers to exposure to, and discussion with, people holding different political beliefs, will not significantly reduce biases or increase political knowledge, since people will reason about this new information in biased and motivated ways.

[11] Presumably, this should apply to cases in which misinformation is entirely prevented from entering into people's information diets, for example via preventive bans on certain content or content providers. Once again, improving the quality of information diets seems unlikely to improve how that information is processed. People are still likely to process true, non-misleading information in ways that are biased and motivated, even if their access to false political information is limited. Interventions addressing ignorance will not improve rationality.

3 Empirical data on Correcting Misinformation[12]

3.1 Debunking

If the above is correct, then we should expect actual debunking interventions, including interventions in online environments, to mostly fail at improving people's knowledge and reasoning processes. There is ample empirical confirmation of this prediction.

A number of studies have now found little effect of interventions aimed at countering *politically relevant* misinformation (and some have found them to achieve the opposite effect of increasing confidence in misinformation).

Let us consider some examples.

Prasad et al. (2009) find that people retained their belief in the link between Saddam Hussein and the 9/11 terrorist attack on New York City, despite being confronted with information to the contrary.

Nyhan and Reifler (2010) discover that "ideological subgroups failed to update their beliefs when presented with corrective information that runs counter to their predispositions." (304).

Ecker and Ang (2019) replicate these results, finding, however, that conservatives are more likely to stick to their guns in the face of counterattitudinal evidence.

Lewandowsky et al. (2012) summarize the (then current) literature on misinformation correction in a similarly pessimistic vein, documenting the so-called "continued influence effect," whereby the original misinformation continues to exert influence on how people reason:

> Research … has consistently found that retractions *rarely*, *if ever*, *have the intended effect* of eliminating reliance on misinformation, even when people believe, understand, and later remember the retraction …. In fact, a retraction will at most halve the number of references to misinformation, even when people acknowledge and demonstrably remember the retraction …; in some studies, a retraction did not reduce reliance on misinformation at all … (114, references omitted, emphasis added)[13]

Though I think it is fair to say that the preponderance of evidence is on the side of the ineffectiveness of debunking, some research suggests small positive effects.

For instance, a meta-analysis by Chan and colleagues (2017) discovers that debunking fake news stories fails to prevent the persistence of misinformation, though it seems to mitigate it somewhat (specifically, groups that are presented with misinformation but no debunking are worse informed than controls, groups presented with misinformation + debunking are better informed than the misinformation only group, but groups presented with misinformation + debunking are still less informed than controls), while recent work by Porter and Wood (2022) finds a robust effect of fact-checking on belief accuracy. So, how can this apparent discrepancy be reconciled?

Some have proposed that the difference can largely be explained by the kind of misinformation that gets corrected in different experimental paradigms. Information concerning singular, one-off events (e.g. the racial identity of the criminal in a news story) is easier to correct regardless of prior (in this case, racial) attitudes, than the more general information more tightly tied up with a person's self-conception as a member of the political tribe (e.g. that members of one racial group are on average no more likely to commit crimes than members of some other group). As Ecker and Ang put it,

> a retraction of an attitude-congruent one-off event might be easily accommodated with a person's worldview because it may not require attitude change. … By contrast, accepting a retraction of attitude-congruent misinformation that is more general is more likely to require attitude change: Misinformation that is general is more likely to be relevant to the associated attitude…. For instance, having received and accepted misinformation that Aboriginal people are generally more likely to commit robberies than Caucasians, a racially prejudiced person might be unwilling to accept a retraction of that misinformation because this

[12] My presentation of the empirical research could strike some as one-sided - aren't I straying too close to simply cherrypicking studies that happen to fit my preexisting normative conclusions? In trying to prevent such charges, I sought, where possible, to rely on meta-analyses and literature reviews; where I could find no such work, I sought to rely on multiple papers supporting my empirical claims, rather than a single reference. If this still seems unsatisfying, it's worth pointing out that such practice is par for the course in the literature on the topic. Even cursory discussions of the empirical methodology, or acknowledgments of research disputing their empirical claims are rarely included. Indeed, high-quality, insightful, well-received philosophical articles on the ethics of fake news and misinformation often rely on single studies or even trade paperbacks, in justifying some of their important empirical premises (Pham and Castro 2020; Fritts and Cabrera 2022; Fraser 2023). If one accepts that it is more secure to rely on meta-analyses and literature reviews than on individual papers, then the foundation for my empirical premises seems stronger than (some of) those authors'. If not, then I am at least in good company.

[13] One could say that this finding demonstrates at least some effectiveness of the corrections; but note that it requires acknowledging and explicitly remembering the correction to reduce reliance on the corrected misinformation by a half. It's doubtful whether we can rely on people - especially on partisans - to consciously keep counterattitudinal corrections firmly in mind outside of experimental settings.

> would necessitate a certain amount of attitude change (2019, 4)[14].

Secondly, while in some cases people may respond to corrections by rationally updating their beliefs, political rationality would also seem to require that they typically modify their evaluative attitudes towards questions about politics and policy accordingly (e.g., support for a candidate should decrease at least somewhat if it was based on faulty – but later corrected – information about her character or record). Generally, this isn't so, as demonstrated by empirical data explicitly examining the effect of debunking on changes in political attitudes - see e.g. Swire et al. (2017), Nyhan and Zeitzoff (2018), Nyhan et al. (2020), Swire-Thompson et al. (2020), Sides (2021), and Klofstad and Uscinski (2023). All these studies find no effect of improved relevant political knowledge on evaluative attitudes.

This is in line with Dennis Chong's (2013) summary of the literature on political belief-updating: "[i]n general, new facts do not change opinions as much as the perceived implications of those facts, which are themselves subject to partisan biases" (113).

Thirdly, it appears that studies with more ecological validity, either closely mimicking actual online environments or directly examining data therefrom tend not to find any positive effect of debunking. For example, in a study of 54 million Facebook users interested either in scientific or conspiracy content, Zollo and colleagues (2017) find that

> [debunking] information online is ignored [by people interested in conspiracy theories]. Indeed, our results suggest that debunking information remains confined within the scientific echo chamber and that very few users of the conspiracy echo chamber interact with debunking posts. Moreover, the interaction seems to lead to an increasing interest in conspiracy-like content. (9)

Lastly, meta-analyses of empirical work on debunking appear largely to confirm the ineffectiveness of the intervention. Walter and colleagues' (2020) paper is a case in point. Having examined a range of individual studies, the authors find that

> the effects of fact-checking on beliefs are quite weak and gradually become *negligible the more the study design resembles a real-world scenario* of exposure to fact-checking. For instance, though fact-checking can be used to strengthen preexisting convictions, its credentials as a method to correct misinformation (i.e., counterattitudinal fact-checking) are significantly limited. (367, emphasis added)

A similar finding has been reported by Chan and Albarracin (2023), whose meta-analysis of over 70 studies of debunking scientific misinformation, also discovered that "attempts to debunk science-relevant misinformation were, on average, not successful" (1514). Moreover, to the extent that small positive effects were observed, they concerned corrections of counterattitudinal misinformation. Just as the hypothesis of political irrationality would predict, people were more likely to believe corrections aligned with their worldview, and disinclined to believe counterattitudinal corrections.

One could worry that the research I cite at best shows little effectiveness of debunking *on committed partisans*; but, one could argue, political neutrals might still benefit from fact-checks, since they aren't as infected with bias.

There are two reasons to resist this worry: first, the objection might underestimate the proportion of partisans in the general population; after all, most of the studies demonstrating the ineffectiveness of fact-checks that I discuss here do not specifically select for partisanship but rather tend to rely on random sampling. So, if they do demonstrate the ineffectiveness of fact-checking/sorting prevention, they demonstrate the ineffectiveness of fact-checking on the beliefs of (what approaches) a random sample of citizens.

Second, empirical research appears to show that political misinformation is in fact overwhelmingly consumed and shared by committed partisans (Narayanan et al. 2018; Guess et al. 2018; Osmundsen et al., 2021). Consequently, the corrections of such misinformation will also overwhelmingly be accessed by political partisans.

Overall, despite the outliers, we should, it seems, retain substantial confidence that debunking interventions don't work.[15]

Though there's less empirical work on SRLs (see Roozenbeek et al. 2023), some of the most recent research appears to confirm the futility of these interventions, in line

[14] Indeed Ecker and Ang confirm this hypothesis.

[15] Can't we rather conclude that since the intervention works better in laboratory conditions than in the real world, we should seek to transform real-world conditions to approximate whatever happens in the lab? Then, the interventions would boast higher real-world effectiveness. This is easier said than done, however. Consider that people's relevant *incentives* in the laboratory are significantly shifted because the likely audience is changed: first, due to demand characteristics, participants might be more willing to respond in the way they believe researchers would like them to respond; second, in the relative privacy of the experimental conditions. participants incur no social opprobrium for disavowing their previously held views. These incentives are reversed in the real world. It's unclear whether and how they can be changed through policy - but if they can, if being factually correct about political matters can somehow become more beneficial, then people will have independent, self-regarding reasons to seek out correct information, making fact-checking mandates superfluous.

with what we should expect given political irrationality. As Aslett et al. (2022) put it,

> evidence from a preregistered randomized field experiment among a large representative sample of Americans reveals that the particular intervention studied here—providing dynamic, in-feed source reliability labels—does not measurably improve news diet quality or reduce misperceptions, on average, among the general population. Our estimates, based on both survey and behavioral data collected over an extended period, are precise and rule out even modest effect sizes by conventional standards. (7)

A good explanation of these findings (explicitly embraced by both Walter et al. and Chan & Albarracin) is politically motivated reasoning. People prefer to maintain their political identity at the expense of genuinely taking on-board information that contradicts their political views. This is why they do not change their minds even when told that the information they'd previously relied on was false.

There is good reason to think that debunking and SRL interventions will continue failing to achieve their goals.’ Failures of political rationality are not remedied by interventions targeting political ignorance. Consequently, governments should not mandate them.

3.2 Sorting

By the same token, governments should not pursue interventions that aim to reduce ideological sorting by increasing exposure to politically discordant points of view. Sorting occurs when different groups of politically opposed people tend only to interact with other partisans of the same viewpoints, tend to consume the same (often biased) news sources, and rarely try to engage the other side in a serious debate – or attempt to apprehend their opponents' perspectives charitably.[16]

Supposing that limiting[17] sorting would expand access to more diverse information, the sad fact of political irrationality once again throws a spanner in the works of better political discourse. As we shall see, much of the research demonstrates that exposure to alternative political viewpoints does little to persuade people to modulate their own positions. So it seems we should not really expect people to arrive at their political beliefs more rationally after being exposed to alternative perspectives.

This line of thought suggests that not just increased passive exposure to, but also more frequent active discussion with, people who hold different views would be unlikely to remedy the problems with political discourse on social media. And though the empirical literature's verdict on the effects of deliberation (on- and offline) appears less clear-cut than what the literature on political irrationality says, it is, still, hardly encouraging. In one philosopher's estimation,

> On its face, the empirical evidence seems to show us both that people are too [biased] to deliberate properly and that deliberation makes them more [biased] … [T]here is ample empirical evidence that *deliberation often stultifies or corrupts us*, *that it frequently exacerbates our biases and leads to greater conflict*. (Brennan 2016, p. 66, emphasis added)

Others are slightly less pessimistic. In their overview of the empirical literature on deliberation, political scientists C. Daniel Myers and Tali Mendelberg (2013) note that:

> Deliberation can help correct some of the pathologies of individual information processing, …, although it can lead to other information-based or socially based pathologies, such as group polarization or convergence. … However, while deliberation is supposed to result in more inclusive decision-making, and racially heterogeneous groups may provide information-processing benefits …, the process of deliberation is rarely free of the inequalities of social status, race, and gender. These problems can be addressed, but specific conditions must be in place to do so. … [O]ther forms of heterogeneity, such as preference heterogeneity, can have complicated effects on the quality and outcomes of deliberation. Process research can also identify biases that are not anticipated by normative scholars, such as [the] finding … that the influence of an

[16] This characterization bears a similarity to what Thi Nguyen has described as an "echo chamber" (2020), i.e. "an epistemic community which creates a significant disparity in trust between members and non-members. This disparity is created by excluding non-members through epistemic discrediting, while simultaneously amplifying members' epistemic credentials. Finally, echo chambers are such that general agreement with some core set of beliefs is a prerequisite for membership, where those core beliefs include beliefs that support that disparity in trust" (146, emphasis removed). Nguyen contrasts echo chambers with what he calls epistemic bubbles - epistemic communities where information is also filtered out, though not necessarily maliciously or even intentionally. Bubbles can be burst easily, whereas escape from echo chambers is often very difficult.

[17] How could governments do this? One proposal is to ban personalized advertising. Without personalized advertising, the argument goes, internet users would no longer receive specially tailored information that ensconces them in an information bubble. Indeed, the European Union has just recently banned Facebook from engaging in this practice (von Hoffman 2023).

> argument depends on whether the argument is introduced by someone who shares the majority's interests, not just on the informational value of the argument (721-2, references omitted).

This does not inspire confidence that getting people to engage with opposing viewpoints would result in a better political discussion on social media, especially in light of the fact that the above summaries also include in their scope highly structured and moderated deliberative exercises. In contrast, online political discussion is frequently much more free-wheeling and unstructured.

In any case, the more recent evidence on the effects of online political discussion on participants is, at best, ambiguous. Some studies (e.g. Bail et al. 2018; Suhay et al. 2018; Oswald and Bright 2022) find that exposure to political disagreement online increases partisan attitudes, including the dislike of opponents. Zhang and colleagues (2022) find that such exposure leads people to actively filter out (by blocking or unfriending) opposing viewpoints which may, paradoxically, lead to *more* ideological sorting. Torcal and Maldonado (2014) and Guidetti and colleagues (2016) find that political disagreement reduces interest in (and knowledge of) politics. It also contributes to feelings of psychological discomfort (Jeong et al. 2019).

Bago and colleagues (2022) find that deliberation has little to no effect on reducing belief in conspiracy theories. Wojcieszak and Price (2010), Robison and colleagues (2018) and Robison (2020) similarly find that exposure to and discussion of opposing views has no effect on bias reduction or attitude strength.

Collectively, the findings suggest that increased exposure to and discussion of alternative political viewpoints has negative to null effects on opinion change, partisan bias, political knowledge, and political interest. Kevin Simler and Robin Hanson (2018) argue that political irrationality explains these results:

> When our beliefs are anchored not to reasons and evidence, but to social factors we don't share with our conversation partners (like loyalty to different political groups), disagreement is all but inevitable, and our arguments fall on deaf ears. We may try to point out one another's hypocrisy, but that's not exactly a recipe for winning hearts and minds (298, footnote omitted).

Consequently, it is highly questionable whether governments can claim to have sufficient evidence that limiting ideological sorting will achieve their ends of improving political discourse, political knowledge, and political reasoning. So it is highly questionable whether they are justified in passing such mandates. In the parlance of public justification: some reasonable citizens would take the above cited studies to demonstrate (to a satisfactory degree) the ineffectiveness of anti-sorting interventions, grounding the opposition to mandating them. Such mandates would, therefore, not be publicly justified, in virtue of failing the requirements of policy epistemology.

In conclusion, governments should not engage in policing misinformation, whether by mandating debunking and SRLs, or by attempting to ban microtargeting in an effort to decrease ideological sorting.[18]

4 Digital Ethics of Misinformation

This conclusion runs counter to a number of proposals in the philosophical literature on social media ethics, having to do both with the spread of misinformation and increasing ideological sorting (see, for example, Cohen 2012; Tufekci 2014; Rini 2017; Singer 2017; Véliz 2020; Pasquale 2020; Castro and Pham 2020; Fritts and Cabrera 2022; Howdle 2023; most of whom take inspiration from the well-known work on polarization and echo chambers by Pariser (2011); see also Bozdag and van den Hoven (2015) for a survey of such arguments). These authors seem explicitly or implicitly committed to the denial of premise 1 & 6 in my arguments. That is, they seem to assume that people exposed to political information online will rationally update their beliefs and adjust their attitudes on the basis of this information. This – so the thinking goes – could become very dangerous when the information is false. In contrast, if people's beliefs were more accurate, if they had access to what the other side thinks, and if they engaged with those positions in good faith, the threat of misinformation would be diminished. Call this "the rationality assumption."

For instance, Véliz (2020) argues that

> Personalized ads fracture the public sphere into individual parallel realities. If each of us lives in a different reality because we are exposed to dramatically different content, what chance do we stand of *having healthy political debates*? When politicians have to design one ad for the whole of the population, they *tend to be more reasonable*, *to appeal to arguments that a majority of people are likely to support*. Personalized ads are more likely to be extreme … When we

[18] One could object that the studies at best offer evidence of short-term ineffectiveness of debunking. It's possible the intervention could offer longer-term benefits. While true, it seems to me that in light of the fact that, as discussed by Gardner (2012) and Tetlock (2017), any longer-term predictions are riddled with extreme uncertainty, it would be problematic to base a justification for coercion on necessarily speculative long-term predictions about the effects of, say, fact-check mandates.

> all see the same ads, *we can discuss them. Journalists, academics, and political opponents can fact-check and criticize them*. Researchers can try to measure their impact. All that *scrutiny puts pressure on political candidates to be consistent*. (83, emphasis added)

Véliz's point about the salutary effects of fact-checking is repeated by other scholars. Frank Pasquale, for example, claims that: "If Google and Facebook had clear and publicly acknowledged ideological agendas, *adult users could grasp them and inoculate themselves accordingly*, with skepticism toward self-serving content" (2020, 98, emphasis added).

Rini (2017) predicts in the same vein that "[a] story that has been flagged as disputed is, presumably, less likely to be trusted on the basis of testimony, and people who persist in sharing disputed stories may suffer reputational consequences" (57) and proposes, as a further development of counter-misinformation strategies, for SMCs to introduce reliability labels for individual users.

The European Commission itself is on board. In expanding upon the already mentioned Disinformation Code of Practice (as of now voluntary, but rumored to become mandatory soon), the august body says:

> signatories [to the Code] should commit to extend the cooperation with fact-checkers. Increasing the impact of fact-checking can be also achieved through a better incorporation and visibility of content produced by fact-checkers. Signatories should look into efficient labelling systems as well as the creation of a common repository of fact-checks, *which would facilitate its efficient use across platforms to prevent the resurgence of disinformation that has been debunked by fact-checkers*. (The European Commission 2021; np., emphasis added)

Véliz's claims about the harmful effects of "fragmentation" (or what I call sorting) are also embraced in the literature. Cohen (2012), for instance, says:

> networked citizen-consumers move within personalized "filter bubbles" that conform the information environment to their political and ideological commitments. This is conducive to identifying and targeting particular political constituencies, but not necessarily to fostering political dialogue among diverse constituencies in ways that might enable them to find common ground… *through robust and open debate*, *which liberal democracy requires to sustain and perfect itself* (1917, emphasis added).

Howdle (2023) argues, along the same lines, that, in the absence of political discussion with others,

> citizens are unable … to help each other develop their understandings and deliberative stances in the light of information they receive from one another. They cannot escape or supplement their limited pools of information. They are unable to update their beliefs and preferences in light of what they learn from their fellow citizens' responses to politicians' claims, proposals, and policies. (451)[19]

Singer (2017) also castigates fake news as a threat to democracy:

> fake news … is contrary to one of the fundamental premises on which democracy rests: that voters can make informed choices between contending candidates. (np.)

These authors appear to think that exposure to, or deliberation with, others not sharing our views (Cohen, Howdle), exposure to fact-checks (Véliz, Rini, the EC), and access to accurate information (Pasquale, Singer) will improve the way people think and talk about politics, because, presumably, people are (more or less) rational in how they process this information (what else could the explanation be?).

As we saw, the rationality assumption flies in the face of empirical evidence. Digital ethics of misinformation, as practiced by these scholars, needs to let go of the excessively optimistic view of what the interventions they support may accomplish (Fritts and Cabrera 2022 is one example of taking the ineffectiveness of political fact-checks seriously; the already mentioned Nguyen 2020 is an example of taking seriously the difficulties with preventing sorting[20]). Since the rationality assumption underlies the policy proposals put forward by Véliz, Pasquale, and others, it follows that

[19] The trouble is, of course, that citizens *don't* "update their beliefs and preferences" even when they have access to others' views.

[20] It seems to me that Nguyen's account bolsters the case I am making in this paper. On Nguyen's view, echo chambers are extremely difficult, if not impossible, to escape merely by increased exposure to counterevidence. Nguyen attributes such difficulties less to individual members' rational irrationality (the literature on which he does not engage with) and more to the structural features of such communities, but in either case, fact-checking and cross-partisan exposure are likely to be ineffective against echo chambers. Things may be different when it comes to breaking out of epistemic bubbles. Here, mere provision of neglected information is sufficient to "burst" them. However, given the sorts of psychological profiles of people antecedently interested in politically relevant information, it is unclear whether any politically partisan epistemic communities are better characterized as bubbles than echo chambers.

these proposals are not well-motivated, whether they target misinformation or ideological sorting.

A wrinkle on the above argument could be that I am overstating the consequences of political irrationality. The findings do not suggest, after all, that people will *never* reliably update their normative and descriptive beliefs about politics in light of new evidence; moreover, failures of rationality need not be particularly pronounced either, and are consistent with a serviceable manner of dealing with political information. After all, in general, all the well-studied cognitive biases notwithstanding, humans are pretty good at arriving at largely true beliefs. As Goldman (1999) puts it:

> None of the psychological literature on biases suggests that people are wholly incapable of forming veridical beliefs. Most of it contends that native cognitive heuristics just don't coincide with normatively appropriate procedures. … This hardly establishes—nor does it purport to establish—that they have zero capacity for accurate belief formation. Indeed, other segments of cognitive science confirm that people are extremely accurate in their beliefs (231).

However, there is a crucial difference between an average person's belief about most things and her beliefs about politics. As political psychologists persistently find out, the average person knows very little about politics (see, e.g., Somin 2013)[21]. This suggests that Goldman's reason to think that, in general, our ordinary thinking processes are by-and-large *reliable* at getting at the truth does not apply to the special case of politics. People are *extremely inaccurate* in their political beliefs. So we have reason to think that the biases operative in political thinking are more pronounced than those that operate in everyday thinking about non-political matters, and make our political belief-forming processes singularly unreliable. Interestingly, the already cited papers exploring deteriorating performance in politics-related cognitive tasks discover just the significant differences in the processing of political vs. non-political information predicted by uniquely political irrationality.

5 Regulation: Insufficient but Necessary?

Suppose one objects: while the policy of countering misinformation may be *insufficient* to attain the goal of improving public debate, such policy is nevertheless a *necessary* condition for improving public debate. Specifically, for public debate to be improved, it is *necessary* (though not sufficient) that the public *have access* to accurate information (this, I take it, is ultimately the point behind Howdle's complaints about microtargeting). Countering misinformation increases the probability of the public accessing correct information. This is sufficient for the intervention's feasibility. Hence, this paper doesn't show it should not be mandated.

I do not think the objection works: for starters, people *already have* access to accurate information – it's just a Google search (or a library visit) away; social media itself is awash with reliable data (it's not *just* misinformation, after all). Consequently, the necessary condition for, for example, rationally forming political opinions on the basis of the best evidence, is already met, absent any mandates.

It could be replied that prominently displayed fact-checks make access to accurate content easier, since they don't require people to actively search for information - scrolling through news feeds alone would suffice. This, in turn, could raise the probability of people accessing, consuming, and internalizing the corrections.

There is reason to doubt this claim, however: for starters, if the backfire effect is real, then consuming counterattitudinal fact-checks will *entrench* partisan views, rather than mitigating them; corrections could also result in *more* sharing of questionable content, as some researchers have found (Mosleh, et al., 2021). Moreover, since people generally find exposure to political disagreement unpleasant, and since encounters with counterattitudinal fact-checks are likely to be seen as encounters with political disagreement, people could be discouraged from using social media entirely - and seek out venues less committed to fact-checking (perhaps the comment sections of partisan news outlets).

It is also worth bearing in mind that governments appear to have at their disposal less coercive alternatives enabling them to raise the probability that misinformation is properly countered. They can simply use their own ability to command attention and project epistemic authority through official communications, and issue corrective messages across social media and other information channels, without mandating anything.

In addition to communication and messaging, it might be possible for governments to promote educational efforts aiming to foster media literacy and critical thinking (though see Gibbons 2023 for skepticism about the efficacy of such measures, in virtue of their failure to genuinely incentivize people to apply these skills to political discussion). Indeed, some have suggested that governments could incentivize being well-informed about politics by simply *paying people* for doing well on the annual national political knowledge test (Caplan, 2013) - this solution seems to address people's incentives, and could conceivably work better than education-based alternatives. In any case, there is a wide scope for government interventions that do not compel speech.

[21] And those in the minority who do know a lot show evidence of being a lot more biased anyway.

Lastly, learning new *accurate* information is not necessary for becoming better informed. It may sometimes make things worse. As a review of a wide range of evidence by Gerd Gigerenzer and Wolfgang Gaissmaier (2011) finds, in some cases having access to *more* information *impedes* judgment accuracy. As the authors put it "ignoring part of the information can lead to more accurate judgments than weighting and adding all information" (451). Thus, access to new, accurate data is not required for epistemic success, especially in the circumstances of "low predictability and small samples" (Gigerenzer & Gassmaier, 2011, 451). Some political information environments have precisely this feature. So, increased consumption of accurate information is not necessary to make us better informed or reach better decisions. [22]

6 Conclusion

Governments need to have good reason to think their policies will achieve their stated aims for those policies to be justified. Debunking misinformation (and providing SRLs) on social media is unlikely to improve people's political knowledge and political decision-making. So governments aren't justified in mandating that misinformation on social media be debunked. For the same reason, governments aren't justified in passing laws aiming to prevent the fragmentation of political discourse online.

Some issues remain: first, a number of interventions other than debunking and SRLs (e.g. cognitive nudges and "prebunking") have been extensively explored in the empirical literature. They seem to have moderately positive effects on political knowledge, rationality, and attitudes (see e.g. Pennycook et al. 2021). Would governments be justified in mandating *these*? Nothing I said so far suggests they would not. Further research is required - the thin normative presuppositions I am embracing in this paper only go so far.

Second, what should we make of SMCs' decision, of their own free will, to institute debunking and SRL? Again, nothing in this paper answers this question. Again, the thin normative presuppositions only take us so far.

Third, in principle, my conclusions are consistent with the permissibility of governments requiring the debunking of non-political information - though this seems to be of limited practical significance, since most debunking efforts are directed at politically relevant content. In most cases where the harms from holding false (or badly formed) beliefs are likely smaller than the benefits of irrationality, we should expect attempts at fact-checking to fail, even for uncontroversially non-political issues (e.g. those that animate devoted sports fans). However, as the costs of error increase, and/or are increasingly internalized (e.g. when it comes to making personal decisions about where to go to school or work, whom to date, or even what city to visit on vacation), this incentivizes people to try to get things right. So, relevant fact-checking would tend to work for them. Ultimately, it is a matter of the sorts of social, psychological, and other incentives people face, and it just so happens that, empirically, political beliefs tend to be ones where the incentives for indulging in irrationality frequently outweigh the incentives for factual correctness for most ordinary people.

Acknowledgements I am grateful to audiences in Leuven, Manchester, Poznan, and Prague, for comments on a previous version of this paper. Two reviewers for this journal made a number of insightful suggestions for which I am also extremely grateful. I wish to thank the Institute for Humane Studies, grant number: IHS017378 for funding that allowed me to present this paper at the ECPR General Conference in Prague. This research was funded by National Science Center, Poland, grant number: 2022/47/D/HS1/01110.

Declarations

Competing Interests The author declares no conflict of interest.

[22] For an intuitive example, consider a voter picking between two candidates. Suppose she has made - by all accounts - a reasonable decision, based on all information she had available, to vote for A over B. Suppose she then learns that 20 years ago, A used a derogatory term when talking about immigration. This upsets the voter to such an extent that she develops a strong bias against A, discounting all positive information about the candidate, and exaggerating all evidence against them. It seems intuitive that, epistemically, the voter is worse off after learning new information than before, and that her decision based on that plus all the information she had previously, will be worse. For another very schematic example, consider a conspiracy theorist, deeply skeptical about the truth of some generally accepted account of some historical event E. It's not unlikely that the theorist has substantial knowledge about minute details of E that most non-experts lack; but it's not obvious it would be right to say that she is therefore better informed than the non-experts who believe the official account. Alternatively, think of two people, A and B, with the same knowledge of E, except A, a believer in astrology, also knows that, at the time of E, Mercury was in retrograde. A knows more than B concerning E, but it's not at all obvious he is better informed about it.

References

Aslett K, Guess AM, Bonneau R, Nagler J, Tucker JA (2022) News credibility labels have limited average effects on news diet quality and fail to reduce misperceptions. Sci Adv 8(18):eabl3844

Aspernäs J, Erlandsson A, Nilsson A (2023) Motivated formal reasoning: ideological belief bias in syllogistic reasoning across diverse political issues. Think Reason 29(1):43–69

Bago B, Rand DG, Pennycook G (2022) Does deliberation decrease belief in conspiracies? J Exp Soc Psychol 103:104395

Bail CA, Argyle LP, Brown TW, Bumpus JP, Chen H, Hunzaker MF, Volfovsky A (2018) Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37), 9216–9221

Bozdag E, van den Hoven J (2015) Breaking the filter bubble: democracy and design. Ethics Inf Technol 17:249–265. https://doi.org/10.1007/s10676-015-9380-y

Brennan G, Lomasky L (2006) Against reviving republicanism. Politics, Philos Econ 5(2):221–252.

Brennan J (2016) Against democracy. Princeton University Press, Princeton, NJ

Calvillo DP, Swan AB, Rutchick AM (2020) Ideological belief bias with political syllogisms. Think Reason 26(2):291–310. https://doi.org/10.1080/13546783.2019.1688188

Caplan B (2007) The myth of the Rational Voter. Princeton University Press, Princeton, NJ

Caplan B (2103) A cheap, inoffensive way to make democracy work better. Econlib. https://www.econlib.org/archives/2013/10/a_cheap_inoffen.html

Castro C, Pham AK (2020) Is the attention economy noxious? Philosophers' Impr 20(17):1–13

Chakraborty P (2024), April 9 Elon Musk says employees in Brazil threatened with arrests. StratNews Global. https://stratnewsglobal.com/world-news/elon-musk-says-employees-in-brazil-threatened-with-arrests/

Chan MS, Jones CR, Hall Jamieson K, Albarracín D, Debunking (2017) A Meta-analysis of the psychological efficacy of messages countering misinformation. Psychol Sci 28(11):1531–1546. https://doi.org/10.1177/0956797617714579

Chan MPS, Albarracin D (2023) A meta-analysis of correction effects in science-relevant misinformation. Nat Hum Behav 7(9):1514–1525.

Chong D (2013) Degrees of rationality in politics. In: Huddy L, Sears D, Levy J (eds) The Oxford Handbook of Political psychology. Oxford University Press, Oxford

Cohen GL (2003) Party over policy: the dominating impact of group influence on political beliefs. J Personal Soc Psychol 85(5):808

Cohen JE (2012) What privacy is for. *Harvard Law Review*, 126, 1904

X Corp (2024) About Community Notes on X. *X Help*. https://help.x.com/en/using-x/community-notes

Ecker UK, Ang LC (2019) Political attitudes and the processing of misinformation corrections. Political Psychol 40(2):241–260

Fraser R (2023) How to talk back: hate speech, misinformation, and the limits of salience. Politics Philos Econ 22(3):315–335. https://doi.org/10.1177/1470594X231167593

Freiman C (2021) Who it's OK to ignore politics. Routledge, New York

Fritts M, Cabrera F (2022) Fake news and Epistemic Vice: combating a uniquely noxious market. J Am Philosophical Association 1–22. https://doi.org/10.1017/apa.2021.11

Gampa A, Wojcik SP, Motyl M, Nosek BA, Ditto PH (2019) (Ideo)logical reasoning: ideology impairs sound reasoning. Social Psychol Personality Sci 10(8):1075–1083. https://doi.org/10.1177/1948550619829059

Gardner D (2012) Future babble: how to stop worrying and love the unpredictable. Random House, London

Gibbons AF (2023) Bullshit in Politics pays. Episteme, pp 1–21. https://doi.org/10.1017/epi.2023.3

Gigerenzer G, Gaissmaier W (2011) Heuristic decision making. Ann Rev Psychol 62(1):451–482

Goldman A (1999) Knowledge in the Social World. Clarendon, Oxford

Google (2024) Find fact checks in YouTube search results. YouTube Help. https://support.google.com/youtube/answer/9229632?hl=en

Guess A, Nyhan B, Reifler J (2018) *Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign* Working paper

Guidetti M, Cavazza N, Graziani AR (2016) Perceived disagreement and heterogeneity in Social networks: distinct effects on political participation. J Soc Psychol 156(2):222–242. https://doi.org/10.1080/00224545.2015.1095707

Hannon M, de Ridder J (2021) The point of political belief. The Routledge Handbook of Political Epistemology. Routledge, pp 156–166

Howdle G (2023) Microtargeting, Dogwhistles, and deliberative democracy. Topoi 42(2):445–458

Huemer M (2012) In praise of passivity. Studia Humana 1(2):12–28

Huemer M (2016) Why people are irrational about politics. In: Anomaly J, Brennan G, Munger M, Sayre-McCord G (eds) [eds.]. Philosophy, politics, and economics: an anthology. Oxford University Press, Oxford, pp 456–467

Jeong M, Zo H, Lee CH, Ceran Y (2019) Feeling displeasure from online social media postings: a study using cognitive dissonance theory. Comput Hum Behav 97:231–240

Joshi H (2024) Socially Motivated Belief and Its Epistemic Discontents. *Philosophic Exchange*http://hdl.handle.net/20.500.12648/14805

Kahan DM, Peters E, Dawson EC, Slovic P (2017) Motivated numeracy and enlightened self-government. Behav Public Policy 1(1):54–86

Klofstad C, Uscinski J (2023) Expert opinions and negative externalities do not decrease support for anti-price gouging policies. Res Politics 10(3). https://doi.org/10.1177/20531680231194805

Lewandowsky S, Ecker UKH, Seifert CM, Schwarz N, Cook J (2012) Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological Science in the Public Interest*. 2012;13(3):106–131. https://doi.org/10.1177/1529100612451018

Lodge M, Taber CS (2007) *The Rationalizing Voter: Unconscious Thought in Political Information Processing*. Available at SSRN: https://ssrn.com/abstract=1077972 or https://doi.org/10.2139/ssrn.1077972

Messina JP (2023) Private censorship. Oxford University Press, Oxford

Meta (2024) About Fact Checking on Facebook, Instagram, and Threads. *Facebook.com*. https://www.facebook.com/business/help/2593586717571940?id=673052479947730

Myers CD, Mendelberg T (2013) Political deliberation. In: Huddy L, Sears D, Levy J (eds) The Oxford Handbook of Political psychology. Oxford University Press, Oxford

Narayanan V, Barash V, Kelly J, Kollanyi B, Neudert LM, Howard PN (2018) Polarization, partisanship and junk news consumption over social media in the US. arXiv Preprint arXiv:1803.01845.

Nguyen CT (2020) Echo chambers and Epistemic Bubbles. Episteme 17(2):141–161. https://doi.org/10.1017/epi.2018.32

Nyhan B, Reifler J (2010) When corrections fail: the persistence of political misperceptions. Polit Behav 32(2):303–330

Nyhan B, Zeitzoff T (2018) Fighting the past: perceptions of control, historical misperceptions, and corrective information in the israeli-palestinian conflict. Political Psychol 39(3):611–631

Nyhan B, Porter E, Reifler J, Wood TJ (2020) Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. Polit Behav 42(3):939–960

Osmundsen M, Bor A, Vahlstrup PB, Bechmann A, Petersen MB (2021) Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. Am Polit Sci Rev 115(3):999–1015

Oswald L, Bright J (2022) How do climate change skeptics engage with opposing views online? Evidence from a Major Climate Change Skeptic Forum on Reddit. Environ Communication 16(6):805–821

Pariser E (2011) The filter bubble: how the new personalized web is changing what we read and how we think. Penguin, London

Pasquale F (2020) New laws of Robotics. Harvard University Press, Cambridge, Mass

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

Pigliucci M (2018) *Nonsense on Stilts: How to tell Science from Bunk* Second edition. Chicago and London: Chicago University Press

Porter E, Wood TJ (2022) Political misinformation and factual corrections on the Facebook News feed: experimental evidence. J Politics 84(3):np

Prasad M, Perrin AJ, Bezila K, Hoffman SG, Kindleberger K, Manturuk K, Powers AS (2009) There must be a reason: Osama, Saddam, and inferred justification. Sociol Inq 79(2):142–162

Rini R (2017) Fake news and partisan epistemology. Kennedy Inst Ethics J 27(2):E–43

Robison J (2020) Does social disagreement attenuate partisan motivated reasoning? A test case concerning economic evaluations. Br J Polit Sci 50(4):1245–1261. https://doi.org/10.1017/S0007123418000315

Robison J, Leeper TJ, Druckman JN (2018) Do disagreeable political discussion networks undermine attitude strength? Political Psychol 39:479–494. https://doi.org/10.1111/pops.12374

Roozenbeek J, Culloty E, Suiter J (2023) Countering misinformation. Eur Psychol 28(3):189–205

Sellman M (2023), September 25 Rumble: platform hosting Russell Brand may be forced offline. *The Times of London*. https://www.thetimes.co.uk/article/russell-brand-rumble-uk-new-web-safety-laws-7ngkmk5v9

Sides J (2021) Do facts change public attitudes toward fiscal policy? In: Barker D, Suhay E (eds) The politics of Truth in Polarized America. Oxford University Press, New York, pp 305–329

Simler K, Hanson R (2018) The Elephant in the brain. Oxford University Press, New York

Singer P (2017) Free Speech and fake news. Project Syndicate https://www.project-syndicate.org/commentary/fake-news-criminal-libel-by-peter-singer-2017-01

Somin I (2013) Democracy and political ignorance. Stanford University Press, Stanford, CA

Somin I (2023) Top-down and bottom-up solutions to the problem of political ignorance. In: Samaržija H, Cassam Q (eds) The epistemology of democracy. Routledge, New York, p 287–315.

Su S (2022) Updating politicized beliefs: how motivated reasoning contributes to polarization. J Behav Experimental Econ 96:101799

Suhay E, Bello-Pardo E, Maurer B (2018) The Polarizing effects of online partisan criticism: evidence from two experiments. Int J Press/Politics 23(1):95–115. https://doi.org/10.1177/1940161217740697

Swire B, Berinsky AJ, Lewandowsky S, Ecker UK (2017) Processing political misinformation: comprehending the Trump phenomenon. Royal Soc Supplement 4(3):160802

Swire-Thompson B, Ecker UK, Lewandowsky S, Berinsky AJ (2020) They might be a liar but they're my liar: source evaluation and the prevalence of misinformation. Political Psychol 41(1):21–34

Taber C, Young E (2013) Political information Processing. In: Huddy L, Sears D, Levy J (eds) The Oxford Handbook of Political psychology. Oxford University Press, Oxford

Tetlock PE (2017) Expert political judgment: how good is it? How can we know? Princeton University Press, Princeton, NJ

The European Commission (2022) *The 2022 Code of Practice on Disinformation*. https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation

The European Commission (2021) Guidance to Strengthen the Code of Practice on Disinformation - Questions and Answers. https://ec.europa.eu/commission/presscorner/detail/en/QANDA_21_2586

Torcal M, Maldonado G (2014) Revisiting the dark side of political deliberation: the effects of media and political discussion on political interest. Pub Opin Q 78(3):679e706. https://doi.org/10.1093/poq/nfu035

Tufekci Z (2014) Engineering the public: big data, surveillance and computational politics. First Monday 19(7). https://doi.org/10.5210/fm.v19i7.4901

Vallier K (2021) Trust in a polarized age. Oxford University Press, Oxford, UK

Véliz C (2020) Privacy is power: why and how you should take back control of your data. Random House, London

von Hoffman C (2023) EU authorities ban Meta from using personal data for advertising. *MarTech.org*. https://martech.org/eu-authorities-ban-meta-from-using-personal-data-for-advertising/

Walter N, Cohen JR, Holbert L, Morag Y (2020) Fact-Checking: a Meta-analysis of what works and for whom. Political Communication 37(3):350–375. https://doi.org/10.1080/10584609.2019.1668894

Wojcieszak M, Price V (2010) Bridging the divide or intensifying the conflict? How disagreement affects strong predilections about sexual minorities. Political Psychol 31:315–339. https://doi.org/10.1111/j.1467-9221.2009.00753.x

Zhang X, Lin WY, Dutton WH (2022) The political consequences of Online disagreement: the filtering of communication networks in a polarized political context. Social Media + Soc 8(3):20563051221114391

Zollo F, Bessi A, Del Vicario M, Scala A, Caldarelli G, Shekhtman L, Quattrociocchi W (2017) Debunking in a world of tribes. PLoS ONE, 12(7), e0181821