Can arguments change minds?

Catarina Dutilh Novaes

Abstract: Can arguments change minds? Philosophers like to think that they can: by engaging in the (presumably rational) process of carefully considering reasons in favor or against a given position or view, we should update our beliefs accordingly. However, a wealth of empirical evidence seems to suggest that arguments are in fact not very efficient tools to change minds. What to make of these radically different assessments of the mind-changing potential of arguments? To address this issue, it seems that we need to look beyond the content and quality of arguments alone: we must also take into account the broader contexts in which they occur, in particular the propagation of messages across attention networks, and the choices that epistemic agents must make between alternative potential sources of content and information. These choices are very much influenced by perceptions of reliability and trustworthiness, which means that the source of the argument may be even more decisive than its content or quality when it comes to how persuasive it will be for a given person. In a nutshell: arguments may well be able to change minds, but only under conducive, favorable socio-epistemic conditions. In this paper, I deploy a *three-tiered model of epistemic exchange* that I've been developing over the past years (Dutilh Novaes, 2020b) to (hopefully) shed light on the mechanisms involved in these processes, and on the conditions under which arguments can change minds.

1. Introduction

Can arguments change minds? Philosophers like to think that they can: by engaging in the (presumably rational) process of carefully considering reasons in favor or against a given position or view, we should update our beliefs accordingly.[1] According to this optimistic view, famously defended by John Stuart Mill in particular, we not only *do* change our mind when exposed to (compelling) arguments (a descriptive claim), but we also *improve* our overall epistemic position by the careful considerations of reasons (an evaluative claim).

However, a wealth of empirical and anecdotal evidence seems to suggest that arguments are in fact not very efficient tools to change minds (Gordon-Smith, 2019) (McIntyre, 2021). For example, the well-documented phenomenon of polarization (Isenberg, 1986) (Sunstein, 2002) suggests that, when exposed to arguments supporting positions different from their prior views, people in fact often (though perhaps not always) become even more convinced of their prior views rather than being swayed by arguments (Olsson, 2013). Frequently, argumentative encounters look rather like games where participants want to score 'points' (Cohen, 1995) (Dutilh Novaes, 2021) rather than engage in painstaking consideration of different views for the sake of epistemic improvement.

[1] In this paper, I speak of 'changing minds' in a rather loose way, but the concept can also be treated more systematically. There are different formal frameworks that purport to give an account of what it means to change one's mind, such as Bayesian inference and various belief revision theories. For our purposes here, the differences between them are immaterial, as they all deal with how agents update their beliefs in view of incoming information.

What to make of these radically different assessments of the mind-changing potential of arguments? To address this issue, it seems that we need to look beyond the content and quality[2] of arguments alone: we must also take into account the broader contexts in which they occur, in particular the propagation of messages across attention networks, and the choices that epistemic agents must make between alternative potential sources of content and information. These choices are very much influenced by perceptions of reliability and trustworthiness, which means that the source of the argument may be even more decisive than its content or quality when it comes to how persuasive it will be for a given person. (In this respect, argumentation would be more akin to testimony than one might expect, as I argued elsewhere (Dutilh Novaes, 2020b).) In a nutshell: arguments may well be able to change minds, but only under conducive, favorable socio-epistemic conditions.

In this paper, I deploy a *three-tiered model of epistemic exchange* that I've been developing over the past years (Dutilh Novaes, 2020b) to (hopefully) shed light on the mechanisms involved in these processes, and on the conditions under which arguments can change minds. I start with the 'optimistic' view on the power of argumentation to change minds, in particular in John Stuart Mill's formulation, and its shortcomings as discussed in the literature (at least as an accurate *description* of the phenomena in question). I then offer a brief description of the three-tiered model and of its relevance for the issue at hand. In Part 4, I discuss two real-life examples of people who had epistemic breakthroughs which involved at least to some extent engagement with arguments, but only against the background of favorable socio-epistemic conditions. I part 5, I clarify a few pending issues. I then close with some concluding remarks.

2. The Millian conception of argumentation and its limitations

Mill is one of the main exponents of the view that interpersonal argumentative situations involving people who truly disagree with each other have the potential to change minds (primarily for the better, he thinks).[3] In *On Liberty* (1859) (Mill, 1999), he notes that, when our ideas are challenged by those who disagree with us, we are forced to evaluate critically our own beliefs.

[2] I understand the quality of an argument as pertaining to familiar criteria for argument quality such as validity and soundness. (Argument quality can also be defined probabilistically.)

[3] I have defended this view myself (Dutilh Novaes, 2020a) but with the important caveat that the beneficial epistemic effect of interpersonal argumentation will come about only against the background of specific circumstances that ensure good faith exchange of ideas (for example, within a community of mathematicians). See below for a discussion of circumstances where argumentative exchanges reliably lead to epistemic improvement.

> [Man] is capable of rectifying his mistakes, by discussion and experience. Not by experience alone. There must be discussion, to show how experience is to be interpreted. Wrong opinions and practices gradually yield to fact and argument; but facts and arguments, to produce any effect on the mind, must be brought before it. (Mill, 1999) (p. 41)

This process is often described as a *free exchange of ideas*, and according to Mill, it is beneficial *even* when we are right and our interlocutors are wrong. The expected result is that the remaining beliefs, those that have survived critical challenges, will be better justified than those held before such encounters. As Mill puts it, "both teachers and learners go to sleep at their post, as soon as there is no enemy in the field." (Mill, 1999) (p. 83) Dissenters thus force us to stay epistemically alert instead of becoming too comfortable with existing, entrenched beliefs—what Mill describes as 'dead dogma'.

But for this process to be successful, dissenters must be permitted to voice their opinions and criticism freely, and indeed Mill's forceful defense of free speech is one of his most celebrated positions. One of his main arguments for free speech is epistemic: he emphasizes the role played by the free exchange of ideas in facilitating the growth of knowledge in a society. The more dissenting views and arguments in favor or against each of them are exchanged, the more likely it is that the 'better' ones will prevail (Halliday & McCabe, 2019).

However, it is not sufficient that dissenters be given the opportunity to voice their opinions freely; it is also of crucial importance that receivers of these opinions and arguments be willing to engage in good faith and with an open mind.[4] Mill pays much attention to the structural conditions for the free exchange of ideas (in particular, that there should be no state-sanctioned censorship of any kind), but he does not seem to take sufficiently into account our well-documented tendencies to avoid engaging with dissenting views altogether, or to explain away contrary evidence so as to preserve prior beliefs (a point that will be further discussed shortly).

More recently, Alvin Goldman articulated a similar account of the social epistemology of argumentation (Goldman, 1994) (Goldman, 2004). The starting point for Goldman is the recognition of a situation of epistemic division of labor, where different members of an epistemic community know different things, and so can benefit from exchanging these epistemic resources with each other. Moreover, given our inescapable fallibility, these exchanges with other knowers may help expose our own mistaken beliefs (as also noted by Mill). A third feature of our socio-epistemic situation is that people sometimes have incentives to deceive and mislead, so a certain

[4] There is also the important issue (to be discussed shortly) of whether dissenting voices will attract attention at all, for example if they belong to marginalized groups.

amount of epistemic vigilance is needed. It is against these background conditions that argumentation becomes a valuable tool in the pursuit of truth and avoidance of error, according to Goldman.

> Norms of good argumentation are substantially dedicated to the promotion of truthful speech and the exposure of falsehood, whether intentional or unintentional. […] Norms of good argumentation are part of a practice to encourage the exchange of truths through sincere, non-negligent, and mutually corrective speech. (Goldman, 1994) (p. 30)

But does argumentation indeed *reliably* succeed in promoting truth and avoiding error in social epistemic contexts, as suggested by Mill and Goldman? Do we readily revise our beliefs when exposed to (good) arguments that contradict them? Do we really "gradually yield to fact and argument", as claimed by Mill? It seems that Mill and Goldman are overly optimistic regarding the power of arguments to change minds. In fact, argumentation appears to be a rather inefficient way to change minds in many real-life situations (Gordon-Smith, 2019).

The truth is that people typically avoid revising their views about firmly entrenched beliefs (a point famously made by Quine (Quine, 1951)). When confronted with arguments or evidence that contradict these beliefs, they tend either to ignore the evidence, explain it away (as we know from the literature on confirmation bias (Nickerson, 1998)), or to discredit the source of the argument as unreliable.[5] These tendencies are exemplified by so-called science deniers such as flat-earthers (McIntyre, 2021), but also in scientific practice where entrenched paradigms often resist a fair amount of counter-evidence before a 'scientific revolution' takes place (Kuhn & Hacking, 2012). In particular, arguments that threaten core beliefs, feelings of belonging, and identities (e.g., political beliefs) seem to trigger various forms of motivated reasoning whereby one ignores or rejects those arguments without engaging substantially with their content (Taber & Lodge, 2006) (Kahan, 2017). Engaging (or not) in argumentation is often a means to express and cement social identities rather than to come closer to the truth (Talisse, 2019) (Hannon, 2019).

Moreover, when choosing among a vast supply of options, there is a tendency to gravitate towards content and sources that confirm one's existing opinions, in so-called 'echo chambers' and 'epistemic bubbles' (Nguyen, 2020). Conversations with like-minded people may reinforce prior beliefs and even drive people to more extreme versions of those beliefs (Olsson, 2013). This means that the mere availability of dissenting opinions is not sufficient to ensure that knowers remain epistemically alert and consider all sides of a question. There is always the option of

[5] But see (Mercier, 2020) and (Coppock, 2022), who argue that epistemic agents do regularly, and competently, update their beliefs in view of new information, including on value-laden matters such as politics.

ignoring (i.e., not engaging with) these dissenters and the substance of their arguments, especially if they are perceived as untrustworthy (Dutilh Novaes, 2020b). This is the familiar phenomenon of polarization: instead of bringing parties closer together, argumentation and deliberation may have the opposite effect of drawing them further apart (Sunstein, 2002).

Another obstacle is the fact that the absence of government-sanctioned censure (as proposed by Mill) is no guarantee that all relevant voices will be truly *heard*. Dissenting views defended by marginalized social groups will tend to attract less attention than those with powerful proponents; the so-called free exchange of ideas is one were power differentials significantly affect the spread and uptake of views. This is the familiar problem of *inclusion* in democratic societies (Young, 2000), which has serious political as well as epistemic consequences. More often than not, it is not the force or quality of an argument alone that determines its uptake; the social position of its proponents is a decisive factor in how much it will spread and be viewed as persuasive.

To be sure, there *are* some contexts where the exchange of reasons in argumentative interactions does seem to lead reliably to people changing their minds and to epistemic improvement (Mercier, 2018) (Dutilh Novaes, 2020a) (Chapters 8 and 9).[6] The literature on group problem-solving has established that, for what are referred to as 'intellective problems', that is, those that have a unique answer within a given theoretical framework (e.g., a mathematical or logical problem), group discussion among peers has a clear beneficial, truth-conducive effect (Laughlin, 2011). Indeed, in specialized contexts such as in science or mathematics, argumentative 'friction' is a quintessential way to produce knowledge (Longino, 1990) (Lakatos, 1976). But this is less obviously the case for so-called 'judgmental problems', that is, those pertaining to values and judgments that do not have a straightforward 'right' answer (Laughlin, 2011). Importantly, in real-life situations, we are more often confronted with judgmental than with intellective problems, and for the former there is no conclusive evidence that argumentation reliably leads to better outcomes. In fact, many of them are instances of *deep disagreements* (Fogelin, 1985) that may not be amenable to being solved by means of reasoning and argumentation.

These observations suggest that we are not 'proper Millians' when it comes to argumentation and dissent. The epistemic alertness that Mill believed would be the natural, almost automatic consequence of being exposed to dissenting opinions and arguments often fails to come about.

[6] The concept of 'epistemic improvement' presupposes that there are suitable metrics that allow us to measure progress. One natural metric is simply what epistemologists call *accuracy*, which roughly corresponds to Goldman's 'pursuit of truth and avoidance of error' (veritism). But more fine-grained metrics may be considered, for example epistemic improvement in terms of *understanding* (Grimm et al., 2016).

The Millian account is thus descriptively inaccurate, or at the very least incomplete. One may retort that the Millian account is still *normatively* correct; but given that it appears to be highly idealized, it is arguably not suitable to offer *prescriptive* recommendations (in the sense of (Bell et al., 1988)) for concrete human agents.[7] Instead, we need a more realistic approach to the (social) epistemology of argumentation, one which takes into account not only the cognitive limitations of individual knowers but also the social complexities of these processes.

3. The three-tiered model of epistemic exchange

We've just seen that the free exchange of ideas is hindered by various factors such as structural power relations and cognitive and social tendencies, so much so that there is no guarantee that wrong opinions and practices will "gradually yield to fact and argument". To address some of the limitations of the Millian conception of argumentation, I've been developing a *three-tiered model of epistemic exchange*, which presents a more realistic account of epistemic exchange through argumentation by considering the costs, obstacles, and risks of engaging in argumentative exchanges (Dutilh Novaes, 2020b). While it is a model of social epistemic processes in general, the key idea is that argumentation truly consists in an *exchange*, where resources flow in both directions (from arguer to receiver but also from receiver to arguer), and thus is a specific kind of epistemic exchange.

This model was inspired by a theoretical framework known as *Social Exchange Theory* (SET) (Dutilh Novaes, 2020b). This is a framework developed by sociologists and social psychologists that seeks to explain human social behavior in terms of processes of exchange, involving costs and rewards, and against the background of social networks and power structures (Cook, 2013). It was originally developed in the late 1950s and early 1960s under the influence of research in economics (rational choice theory), psychology (behaviorism), and anthropological work by Malinowski, Mauss, and Lévi-Strauss. SET is an influential and empirically robust framework, which has been used to investigate a wide range of social phenomena (such as romantic relationships, business interactions, trust in public institutions, among many others). In particular, and relevant for our purposes, it has been extensively used to investigate interpersonal communication (Roloff, 2015). The SET models are neither purely descriptive—as they rely on certain idealized assumptions such as that agents seek to maximize rewards and minimize costs—nor purely normative, given that they incorporate experimental findings as well as extensive observational data. Moreover, SET combines a first-person perspective, which explains and predicts choices that individuals make between different potential exchange

[7] I don't think that the Millian story is fully convincing as a normative account either, but a thorough discussion of this point goes beyond the scope of this paper. See (Fantl, 2018) for a critique of the Millian idea that engaging with dissenters is always rational/desirable.

partners, with a third-person perspective, which focuses on structural features of these exchange networks.

The three-tiered model of epistemic exchange adapts insights and results from SET to exchanges that are specifically *epistemic*, that is, when epistemic resources such as knowledge, evidence, information etc. are involved (possibly alongside other kinds of resources).[8] The model allows for a meticulous account of the conditions under which successful epistemic exchange may occur or fail to occur. Crucially, there seem to be two preliminary stages that determine whether specific agents will be in a position to engage in fruitful epistemic exchange: the *networks* that determine which sources and which epistemic resources an agent is exposed to; and the *contrastive choices* that agents must make regarding which contents and sources to engage with (among those she is exposed to). Thus seen, the three stages for epistemic exchange are:

1. **Attention/exposure**. The first stage consists in establishing whether people are potential exchange partners of each other, given the relevant opportunity structures for epistemic engagement within a network. In simpler terms: who is in an agent's network of potential contacts? Who is in a position to attract the attention of others? It may be that potential lines of communication are cut, say in the case of structural censorship or epistemic bubbles. But it may also be that so many signals are being broadcast that many different sources are competing for the receiver's attention (Gershberg & Illing, 2022), in a so-called 'attention economy' (Franck, 2019).[9]
2. **Choosing whom to engage with**. The next level comprises the choices that agents make against the background of possibilities for exchange, as determined by the relevant opportunity structures. Typically, there will be a number of options for a given agent—for example, the various newspapers that I can read on any given day, among those that I have access to. Given limitations of time and attention, contrastive choices will have to be made. Among those sources that have caught my initial attention, who do I view as worthy of consideration as an exchange partner? At this point, considerations of *trustworthiness* (Hawley, 2019) and *expertise* (Goldman, 2018) come into play, as well as the perceived value of the content being offered by different potential exchange partners. In particular, trusting someone will often entail *not* trusting someone else, especially when their respective messages conflict (Dutilh Novaes, 2020b).
3. **Engagement with content**. It is only at a third stage that engagement with *content* properly speaking should occur; this is when the actual epistemic exchange takes place. At this point, the receiver will reflectively (and perhaps critically) engage with the

[8] See (Dutilh Novaes, 2020b) for further details on how the three-tiered model emerges from SET.

[9] See (Dutilh Novaes & de Ridder, 2021) for a discussion on scarcity vs. overabundance of information in epistemic environments.

> argument being offered, seeking to understand its substance and evaluate its cogency. In case of a positive evaluation, this may lead to a change in view for the receiver (though even at this stage the receiver may still balk at revising her beliefs). It may also lead to a mutually beneficial exchange where both arguer and addressee improve their respective epistemic stances, as posited by Mill, and in some cases even go on to create new epistemic resources together (as in Lakatos' 'proofs and refutations' model of mathematical practice (Lakatos, 1976)).

Figures 1 to 3 represent the three tiers.[10] For simplicity, a main agent is depicted with other agent around her, but the model in fact focuses on complex networks of agents who are interconnected to different degrees. The topologies of such networks crucially determine how these socio-epistemic processes come to unfold.[11]

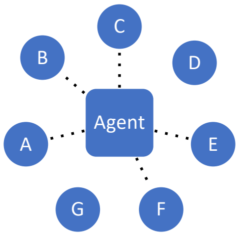

Figure 3: Attention

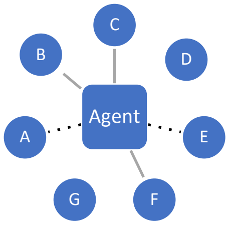

Figure 4: Choices

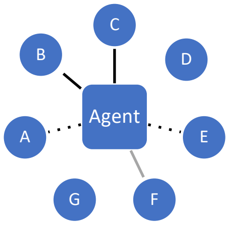

Figure 5: Engagement

Attention: Agent does not 'see' sources D and G, the other sources catch her attention (dotted lines).
Contrastive choices: Agent deems B, C and F as worth exchanging with (grey lines), but not A and E.
Engagement: Agent eventually engages substantively with B and C (black lines), but not with F.

Millian conceptions of argumentation tend to focus primarily on tier 3—the 'force' of an argument alone should suffice to change minds—and to downplay some of the structural obstacles to a truly free and equal exchange of ideas.[12] Indeed, stages 1 and 2 crucially determine

[10] The model can also be understood in terms of set containment: at a given point in time, the set of people I actually engage in epistemic exchanges with is a subset of those who I deem worth exchanging with (above a certain threshold), which in turn is a subset of those who, due to our respective positions in the network, are potential exchange partners for me.

[11] Notice that there are a number of interesting structural similarities between the three-tiered model that I present here and the *network epistemology* research program, as developed by Zollman (Zollman, 2013), Olsson (Olsson, 2013), O'Connor and Weatherall (O'Connor & Weatherall, 2019), among others. For reason of space, I do not develop this point further here, which will remain a topic for future research.

[12] Mill's own emphasis on freedom of speech is aimed at creating a maximally inclusive informational environment, and thus at increased exposure to various views (phenomena belonging to Tier 1). Mills mentions some factors as having a central role in making exchanges more likely to succeed, such as the importance of education. However, he

if and when someone will seriously engage with the epistemic resources being offered by someone else at all. Just as the original SET models, the three-tiered model is neither purely normative nor purely descriptive. It is not purely normative because it does not consider ideal or idealized agents: instead, it considers agents with limited cognitive resources, and who are susceptible to what Levy describes as 'bad beliefs' (Levy, 2021). Moreover, the model is empirically robust as it draws on decades of SET's experimental and observational findings pertaining to exchanges more generally. However, the model is not purely descriptive or predictive either, as it seeks to *explain* the mechanisms that lead different people to engage in epistemic exchanges with some sources but not with others; this is done on the basis of a few foundational principles such as reciprocity and fairness, and by highlighting in particular the roles of attention and trust in such processes. As such, the model is perhaps best understood as an *explanatory model*, in the sense that it seeks to represent some of the causes of the target phenomenon and the mechanisms responsible for bringing it about (Ivani & Dutilh Novaes, 2022). (It may also lead to *prescriptive* recommendations on how to facilitate certain types of epistemic exchanges.)

The three-tiered model offers an explanation for *why* arguments often fail to change minds, as it highlights some of the necessary conditions for this to occur. First, a suitable relation of attention and exposure must emerge between sender and receiver—which is far from obvious, especially in highly saturated informational environments such as the ones we currently inhabit (Gershberg & Illing, 2022). Secondly, a knower must make choices regarding whom to engage with, among the different possibilities: this is where considerations of trustworthiness—understood as related to both *competence* and *benevolence* (Dutilh Novaes, 2020b) (Dutilh Novaes, 2023)—arise. If I already suspect that a given source does not hold benevolent attitudes towards me, should I really spend my precious time and energy engaging with their arguments? Maybe not (Köymen & Dutilh Novaes, forthcoming). For example, the refusal to engage with scientific arguments supporting the efficacy and safety of vaccines on the part of so-called 'anti-vaxxers' is often justified by the (not entirely unreasonable) suspicion that spurious interests are involved (e.g., the 'evil Big Pharma' narrative (Dutilh Novaes, 2020b) (Ivani & Dutilh Novaes, 2022)). Finally, the exchange itself requires that agents with very diverse epistemic backgrounds find enough common ground and suitable means of communication rather than talking past each other, which is far from obvious especially in situations of ideological/political disagreement (Talisse, 2019). If the (potential) exchange fails at any of these three levels, then arguments will not prompt a change of mind.

4. Real-life examples

does not provide a detailed analysis of the conditions under which successful epistemic exchange may occur or fail to occur; in particular, they may fail even in contexts where (religious or otherwise) persecution is not present.

Despite all these challenges, the conclusion that arguments *never* change minds is also unwarranted: arguments sometimes *do* change minds. The question then becomes, under which conditions is this (more) likely to happen? The three-tiered model provides suitable conceptual tools to address this question. Instead of discussing it in the abstract or with toy examples, I here present two recent concrete examples of people who underwent radical epistemic transformations where arguments (presumably) played a significant role: Megan Phelps-Roper, formerly a prominent member of the Westboro Baptist Church,[13] and Derek Black, formerly a prominent proponent of white supremacy in the USA.[14]

The Westboro Baptist Church is a hyper-Calvinist congregation based in Topeka, Kansas, often described as a hate group. It is known for engaging in inflammatory homophobic pickets, as well as hate speech against atheists, Jews, Muslims, transgender people, and numerous Christian denominations. Megan Phelps-Roper is a granddaughter of founder Fred Phelps, and was raised to be a prominent member of the group. As such, she grew up immersed in their stern ideology, and from early on participated in pickets at funerals of gay men (with signs featuring slogans such as 'GOD HATES F*GS') and later of soldiers killed at war (as Westboro members believe that the wars that the US has been involved in in recent decades are God's punishment for the country's tolerance of homosexuality).

Despite the extreme positions of Westboro members, their children, including Megan, typically attended Topeka public schools. At school, she was presumably exposed to other, more tolerant worldviews, but this did not substantially affect her own conviction in the Westboro belief system. Many members had received higher education and some, including Megan's mother, worked as lawyers. Thus, they did not exactly live in an epistemic bubble in the sense of not being exposed to alternative belief systems; in fact, they believed that Westboro members could best preach to the 'wicked' by living among them. The thought was that, if you really knew the 'truth' in your heart, exposure to the world of the wicked would not affect your devotion. In practice, however, there was no room at all for epistemic autonomy or dissent: the supreme value that was instilled in children was that of complete obedience.

[13] My discussion of Megan Phelps-Roper's trajectory draws primarily on the 2015 *New Yorker* profile of her: https://www.newyorker.com/magazine/2015/11/23/conversion-via-twitter-westboro-baptist-church-megan-phelps-roper. She also wrote a memoir, tellingly titled *Unfollow: A Journey from Hatred to Hope* (Phelps-Roper, 2020).

[14] My discussion of Derek Black's trajectory relies primarily on an interview for the *New York Times* podcast 'The Daily' (transcript here: https://www.nytimes.com/2017/08/22/podcasts/the-daily-transcript-derek-black.html). There is also a book narrating Black's journey: *Rising Out of Hatred: The Awakening of a Former White Nationalist*, written by journalist Eli Saslow (Saslow, 2018).

In 2009, Megan joined Twitter to further spread Westboro's views. Some of her homophobic tweets were picked up on and re-tweeted (in the spirit of mockery) by large accounts, which resulted in her receiving many angry replies but also gaining a significant number of followers. She thereby came to be in contact with a wider range of critics, to whom she diligently replied citing biblical passages (along with pop culture references and emojis). She was used to giving interviews to journalists, but on Twitter she could engage with many people directly, with no journalistic filter.

But Megan had by then also started having doubts about some of the Westboro teachings. In particular, around the time she joined Twitter, Westboro was preparing for the end of the world. There were very specific predictions on how Westboro members would lead a hundred and forty-four thousand Jews who repented for killing Jesus through the wilderness of Israel, until Christ would finally come to save them all. Megan felt there was no proper scriptural support for many of these predictions, and turned to Twitter for answers. More specifically, she started following and engaging with Jewish Twitter users, in particular with a Jerusalem-based web designer called David Abitbol.

And thus, even if through bitter debate, she began to forge deeper connections with other Twitter users. Until then, interactions with 'the wicked' had remained superficial and fleeting, such as with counter-protesters at pickets. On Twitter, however, she got involved in extended debates with specific people (such as Abitbol) with whom she developed fierce but friendly patterns of interaction. To her surprise, for the first time in her life she started *caring* about what people outside of Westboro—in particular, some of her Twitter acquaintances—thought of her; the connections with some of her Twitter interlocutors became increasingly meaningful. (In fact, she ended up marrying one of them years later.)

And so, as a result of some small seeds of doubt concerning Westboro's preaching (as well as concerns pertaining to changes in how the church was run and the role of women therein), but mostly through her Twitter connections and interactions, Megan embarked in a long and painful process of questioning everything she had been brought up to believe. About three years after joining Twitter, she started seriously considering leaving the church. That would, of course, entail tremendous social and emotional costs; she would basically lose all contact with her immediate and extended family. She eventually made the consequential decision to leave the church (together with her younger sister Grace) on November 2012, and began connecting again, including in offline environments, with some of her Twitter contacts such as Abitbol. She has since become an advocate for dialogue between groups with conflicting views, and has spoken on multiple venues about her experiences (including the inevitable TED talk[15]).

If we are to believe her own account of the process, arguments played an important role in Megan's (slow but profound) 'epistemic breakthrough' of coming to realize that she could no longer endorse the Westboro belief system. There was much deliberation involved, both with herself and with many of her Twitter contacts (some of whom were also knowledgeable on sources she considered authoritative, in particular the Bible). Through these processes (which at times resembled Socratic dialogues), inconsistencies and discrepancies in the Westboro doctrines became apparent to her, leading to a thorough revision of her own convictions. However, two necessary conditions had to be in place for these arguments to do their work: naturally, she had to be exposed to them (through Twitter, she could be exposed to a wide range of sources and interlocutors); but more importantly, these arguments were coming from people she had grown to respect and care about. She had had exposure to ideas that clashed with the Westboro doctrines before (e.g., at school), but this time the sources of these ideas were people she had forged deeper connections with. This time, she paid more attention and engaged in earnest with the substance of their arguments. What is perhaps remarkable about Megan's trajectory is the fact that the process of recalibration of attributions of respect and trust to different people (away from Westboro members and towards 'outsiders') happened primarily by means of online interactions rather than face-to-face ones. (Her *New Yorker* profile describes the process as 'conversion via Twitter'.) Online connections can become 'real' connections after all, and may offer a much wider net of potential epistemic exchange partners.[16]

Derek Black's trajectory bears interesting similarities to Megan Phelps-Roper's, but in his case the 'conversion' took place primarily through face-to-face interactions rather than online. Derek is the son of Don Black, prominent white supremacist and founder of Stormfront, one of the most influential white supremacist online communities in the US. His godfather is David Duke, one of the most visible Ku Klux Klan leaders in recent decades (as shown in the 2018 Spike Lee movie *BlacKkKlansman*). Both Duke and Don Black are Ku Klux Klan Grand Wizards. Derek was raised to be the 'crown prince' of white supremacy in the United States, and from early on was deeply involved in promoting this worldview, including producing a radio show with his father.

[15] https://www.ted.com/talks/megan_phelps_roper_i_grew_up_in_the_westboro_baptist_church_here_s_why_i_left/transcript

[16] See (Lewiński & Dutilh Novaes, Forthcoming) for an account of online communication drawing on the three-tiered model of epistemic exchange.

Different from Megan Phelps-Roper, Derek was homeschooled, and so had limited exposure to worldviews other than his family's white supremacist beliefs during his youth. His whole socio-emotional world while growing up consisted of people espousing the same ideology. At age 21, he decided to enroll at the New College of Florida in Sarasota, a four-hour drive away from home; this was the first time he left the insular world of white supremacism he had grown up in. He began to live what might be described as a 'double life': recording the radio show with his father in the morning, then attending classes and socializing with students who were (left-leaning) social justice advocates during the rest of the day. Initially, his identity as a white supremacist had not been revealed.

But inevitably, at some point a fellow student exposed his identity, and his racist beliefs and ongoing activism became public knowledge at the college. Unsurprisingly, this led to him becoming ostracized among students. The one exception was a small group of Jewish students who began to invite him to their Shabbat dinners. (By then, Derek had already had a brief relationship with a Jewish woman, which had come to an end when his white supremacy persona became public knowledge.) Perhaps because these were the only people still willing to socialize with him, he became a regular at their dinners.

While some of the dinner-goers did not seek to confront Derek in his beliefs openly, others engaged in heated intellectual discussions with him. Here is an account of these discussions in his own words:

> "I would say, "This is what I believe about I.Q. differences, I have 12 different studies that have been published over the years, here's the journal that's put this stuff together, I believe that this is true, that race predicts I.Q. and that there are I.Q. differences in races." And they would come back with 150 more recent, more well researched studies and explain to me how statistics works and we would go back and forth until I would come to the end of that argument and I'd say, Yes that makes sense, that does not hold together and I'll remove that from my ideological toolbox but everything else is still there. And we did that over a year or two on one thing after another until I got to a point where I didn't believe it anymore."[17]

These conversations went on for years, during which Derek gradually moved away from the white supremacist ideology he had grown up with. Eventually, in 2013, Derek wrote a public statement to the Southern Poverty Law Center, publicly renouncing his previous views. He had much to lose socially and emotionally by distancing himself from white supremacy, including his close

[17] Source: https://www.nytimes.com/2017/08/22/podcasts/the-daily-transcript-derek-black.html

relationship with his family: changing one’s mind can be not only cognitively but also socially costly (an aspect also explored in (Gordon-Smith, 2019)). As with Megan Phelps-Roper, Derek’s epistemic breakthrough did not happen overnight: it was the result of a long process where his beliefs were dispelled one by one, at least partially through the force of arguments (that is, at least if we are to believe his own account of this process). However, once again the fact that arguments came from people whom Derek had come to respect on a personal level (despite the ethnicities of some of them being considered as ‘inferior’ according to the white supremacist worldview he had espoused until then) was a crucial element in the process. He truly listened and engaged with the substance of their arguments because of this favorable interpersonal setting, which in turn was facilitated by his vulnerability and the fact that these were the only people still willing to interact with him on campus. (As Megan, Derek also ended up in a long-term romantic relationship with one of the people who challenged his beliefs early on.)

Until he went to New College, Derek’s exposure to other worldviews had been limited (tier 1 phenomenon), and he had been raised to trust only those who espoused similar ideas as his family’s (tier 2 phenomenon). In Nguyen’s (2020) terms, he was both in an epistemic bubble and in an echo chamber (whereas Megan Phelps-Roper was primarily caught in an echo chamber but not as much in an epistemic bubble). The rewiring of circuits of attention and trust prompted by his experiences on campus is what enabled arguments to do their mind-changing work on Derek.

Naturally, arguments can also change minds on specific issues. The two cases described here correspond to complete overhauls of whole belief systems, but arguments can also, and likely more easily, cause localized revisions (which may require some accommodations but not as radically as in these two cases). The point here is that, if *even* in these two extreme cases arguments appear to have changed the minds of Megan and Derek, then a fortiori in more mundane cases this can occur as well.

One topic I’ve investigated in previous work is the change in public opinion regarding the folk character of ‘Black Pete’ in the Netherlands (Zwarte Piet) (Dutilh Novaes et al., 2020). Black Pete is the assistant of St. Nicholas at the hugely popular St. Nicholas festivities in early December, and was traditionally portrayed in highly racialized ways (blackface, curly hair, thick red lips). In recent years, there has been a significant shift in public opinion regarding the purported racist nature of the character; while until some 10 years ago, the character was viewed by 95% of the population in a positive light, currently at least one third of the population (and rising) came to see it as unacceptable. This has led to important changes in how the character is portrayed, most significantly a sharp decline in the use of blackface makeup. Arguments seem to have played an important role in this shift in public opinion, in particular by confronting and dispelling some of what Charles Mills has aptly described as ‘white ignorance’ (Mills, 2015).

5. Clarifications

Before concluding, a few clarifications of the picture sketched so far seem to be required. Firstly, from an egalitarian-progressive perspective, Megan's and Derek's 'conversions' are viewed as positive because they came to renounce what many of us take to be wrong and problematic worldviews. They attained what *we* take to be significantly improved epistemic states.[18] But the general mechanisms described by the three-tiered model—pathways of attention, trust, and engagement—do not favor specific ideologies (Dutilh Novaes, 2023). Indeed, the spread of 'unsavory' positions such as vaccine rejection and various conspiracy theories follows similar patterns. In particular, propensity to espouse conspiracy theories seems to be strongly associated with distrust towards established institutions such as governments, the press, and the scientific establishment (van Prooijen et al., 2022). In the end, whether we come to espouse 'good' or 'bad beliefs' (in Levy's terminology) is very much a result of the epistemic environments we find ourselves in, including our attributions of credibility and trustworthiness to different sources; the processes leading to 'good' or to 'bad' beliefs are not fundamentally different (Levy, 2021). Bad beliefs can also be supported by arguments—'bad' arguments perhaps (though not necessarily!), but arguments nevertheless, and they too can change minds if the conditions are suitable. (Notice that this is also a thorny point for the optimistic Millian who maintains that truth will eventually prevail, provided that all views can be openly expressed and discussed.)

A second clarification pertains to whether there are jointly *sufficient* conditions for arguments to change minds. What the three-tiered model describes are *necessary* conditions pertaining to attention and attributions of credibility. But even if these are in place, there is no guarantee that arguments will indeed change minds; 'stubborn' thinkers may, and indeed often do, still stick to their prior beliefs, especially beliefs that are thoroughly enmeshed with their ways of living.[19] While the idea of a fool-proof method to change minds by means of high-quality arguments may seem appealing, in practice arguments alone cannot *force* an epistemic update to occur.[20]

[18] At least, I am assuming that most readers of this piece will reject homophobia, racism, and white supremacy.

[19] Compare Sally Haslanger's notion of *cultural technē*, understood as a collection of social meanings "that provides a 'stage-setting' for action and is a constituent part of the local social-regulation system" (Haslanger, 2021) (p. 23). A cultural technē 'gone wrong' will organize social structures in unjust ways, and for Haslanger this is exactly what *ideology* is. To dismantle a cultural technē 'gone wrong', rational arguments by themselves will have little to no effect; instead, the cultural technē in question must first be 'disrupted' to open up possibilities for contestation. In the cases of Megan and Derek, the disruption in question was caused by inhabiting different discursive and affective environments (Twitter for Megan, college for Derek). But at a broader, societal level, more significant disruptions seem necessary, following Haslanger's notion of cultural technē. They may however still partially involve arguments, in for example what is known in the Marxist tradition as *consciousness raising*.

[20] This is a point related to what some authors identify as the intrinsically *coercive* nature of arguments (Nozick, 1981) (Casey, 2020), which is however not always effective. As Wittgenstein pointed out, even a correct mathematical proof may fail to persuade, despite 'the hardness of the logical must' (see (Wright, 1990)). A related

Relatedly, even when a change of mind apparently prompted by arguments occurs, it may well be that the efficacious causes are ultimately non-epistemic factors such as social factors (e.g., a desire to belong to a certain group) or economic incentives. In other words, we cannot be sure that the arguments were persuasive for the *right* (rational) reasons, i.e., pertaining to their quality *qua* arguments.[21] Indeed, given the human propensity for *rationalization* (Cushman, 2020), it is often not transparent to the agent herself what exactly prompted a change of mind.

Finally, Megan and Derek were (presumably) swayed by arguments only because they previously accepted the basic rules of the language-game of argumentation.[22] Megan was skilled at the practice of arguing in support of religious beliefs on the basis of careful scriptural analysis, and came to respect the scriptural knowledge of some of her Twitter interlocutors. Derek referred to 'scientific' studies himself to support his views (e.g., that there are racial IQ differences), but then came to realize that there were much *better* scientific studies supporting opposite views. They were thus receptive to the very practice of supporting positions with arguments and evidence; that is, there was at least a certain degree of meta-level agreement on the 'rules of the game' between them and their interlocutors. Had this not been the case—for example, if they thought that everything was really a matter of 'what feels right to me', or that all opinions are equally valid—then it is unlikely that arguments would have had any grip on them.

In sum, the cases just discussed still do not offer full reassurance to the optimistic Millian that, under the right conditions, good arguments will indeed change minds for the better. 'Bad' arguments may also change minds for the worse; good arguments may fail to prompt a change of mind, even under the right circumstances; even when it looks like a change of mind was caused by engagement with high-quality arguments, we cannot be sure that the actual causes for the change were truly argumentative; and arguments by themselves will have little to no effect in cases where people reject the very idea of updating their beliefs in view of arguments and evidence. Still, the cases discussed offer at least a plausible account of the mechanisms through which good arguments may change minds, and thus partially vindicate Mill's view that engaging with dissenters may allow for the correction of errors.

6. Conclusions

question, not addressed here, is whether it is ethically acceptable to try to change someone's mind (be it by arguments or other interventions); does it not constitute a problematic infringement of someone's intellectual autonomy?

[21] Thanks to James Owen Weatherall for raising these two worries.

[22] I owe this point to Harvey Siegel.

We started with the view that arguments can change minds by the force of reason alone. In practice, however, and certainly in situations where values and political views play a significant role, arguments do not seem to be particularly suitable to change minds; on the contrary, people typically either outright refuse to engage with or else are not moved by arguments that clash with their deep-seated beliefs. But in some circumstances, arguments may in fact succeed in changing minds. I've argued that two important but often underappreciated factors, attention and trust, need to be taken into account to explain the persuasiveness (or lack thereof) of arguments in specific situations. Arguments can only change minds if they catch the receiver's attention, and if the receiver chooses to give them careful consideration, which in turn is significantly (but not completely) determined by attributions of credibility and trustworthiness to the source. If these conditions are in place, then it may well happen (though again, no guarantee!) that arguments will change someone's mind.[23]

We've looked into the real-life cases of Megan Phelps-Roper and Derek Black. In both cases, they came to renounce the worldviews they were brought up with thanks in part to argumentative engagement (at least if we are to believe their own accounts of these processes). However, they had first both come to respect the *sources* of these arguments, and this is why they engaged with their substance in earnest rather than dismissing them outright. Also, in both cases, it was a lengthy process: arguments need time to truly change minds. It was the cumulative effect of many such argumentative interactions that eventually led them to a complete abandonment of their original positions; changing minds through arguments does not happen overnight. How representative these two cases are on a broader scale is difficult to establish; but since the main claim of this paper is merely an existential one—arguments can *sometimes* change minds—they offer support to this modest claim and help illustrate the mechanisms involved.

These two cases also show that changing minds through arguments is costly (Casey, 2020). It can be costly for the person who changes their mind, as it may entail the loss of their most meaningful social and affective connections; and it is costly for those trying to change minds through arguments, as they must invest significant resources (time, energy) to catch the receiver's attention and to gain enough of their trust so that they will engage in earnest with the substance of the arguments. Moreover, the argumentative process itself can be slow and require many iterations. Thus, a plausible conclusion to be drawn is that arguments are not very efficient tools to change minds (as opposed perhaps to e.g., narratives or propagandistic discourse). Still, we

[23] McIntyre (McIntyre, 2021) presents a very similar picture of how science deniers (sometimes) change their minds in view of argument and evidence: "All of these stories are basically the same. They happen within the context of a trusting, personal relationship. As I've said all along, facts and evidence can matter, but they have to be presented by the right person in the right context." (p. 73)

need not go as far as concluding that arguments are pointless and futile, as some like to say; in the right circumstances at least, they may in fact change minds for the 'right' reasons.

Acknowledgments: This research was generously supported by the European Research Council with grant ERC-2017-CoG 771074 for the project 'The Social Epistemology of Argumentation'. I am grateful for the feedback received from various audiences: the 'Epistemic Breakthroughs' workshop in St. Andrews, ECA in Rome, and seminars at Bochum and VU Amsterdam.

References

Bell, D. E., Raiffa, H., & Tversky, A. (1988). Descriptive, normative, and prescriptive interactions in decision making. In D. E. Bell, H. Raiffa, & A. Tversky, *Decision making: Descriptive, normative, and prescriptive interactions* (pp. 9–30). Cambridge University Press.

Casey, J. (2020). Adversariality and Argumentation. *Informal Logic*, *40*, 77–108.

Cohen, D. H. (1995). Argument is War...and War is Hell: Philosophy, Education, and Metaphors for Argumentation. *Informal Logic*, *17*, 177–188.

Cook, K. S. (2013). Social exchange theory. In J. DeLamater & A. Ward (Eds.), *Handbook of social psychology* (pp. 6–88).

Coppock, A. (2022). *Persuasion in parallel: How information changes minds about politics*. The University of Chicago Press.

Cushman, F. (2020). Rationalization is rational. *Behavioral and Brain Sciences*, *43*, 1–69.

Dutilh Novaes, C. (2020a). *The Dialogical Roots of Deduction*. Cambridge University Press.

Dutilh Novaes, C. (2020b). The role of trust in argumentation. *Informal Logic*, *40*, 205–236.

Dutilh Novaes, C. (2021). Who's Afraid of Adversariality? Conflict and Cooperation in Argumentation. *Topoi*, *40*(5), 873–886. https://doi.org/10.1007/s11245-020-09736-9

Dutilh Novaes, C. (2023). The (higher-order) evidential significance of attention and trust—Comments on Levy's Bad Beliefs. *Philosophical Psychology*, *Online First*. https://doi.org/10.1080/09515089.2023.2174845

Dutilh Novaes, C., & de Ridder, J. (2021). Is fake news old news? In S. Bernecker, A. K. Flowerree, & T. Grundmann, *The Epistemology of Fake News*. Oxford University Press.

Dutilh Novaes, C., Sullivan, E., Lagewaard, T., & Alfano, M. (2020). Changing minds through argumentation: Black Pete as a case study. In *Reason to Dissent: Proceedings of the 3rd European Conference on Argumentation* (Vol. 2, pp. 243–260). College Publications.

Fantl, J. (2018). *The Limitations of the Open Mind*. Oxford University Press.

Fogelin, R. (1985). The logic of deep disagreements. *Informal Logic*, *7*, 3–11.

Franck, G. (2019). The economy of attention. *Journal of Sociology*, *55*, 8–19.

Gershberg, Z., & Illing, S. D. (2022). *The paradox of democracy: Free speech, open media, and perilous persuasion*. University of Chicago Press.

Goldman, A. I. (1994). Argumentation and social epistemology. *Journal of Philosophy*, *91*, 27–49.

Goldman, A. I. (2004). An Epistemological Approach to Argumentation. *Informal Logic*, *23*, 49–61.

Gordon-Smith, E. (2019). *Stop Being Reasonable: How we Really Change Minds*. Public Affairs.

Grimm, S., Baumberger, C., & Ammon, S. (2016). *Explaining understanding: New perspectives from epistemology and philosophy of science*. Routledge.

Halliday, D., & McCabe, H. (2019). John Stuart Mill on Free Speech. In D. Coady & J. Chase, *The Routledge Handbook of Applied Epistemology* (pp. 72–87). Routledge.

Hannon, M. (2019). Empathetic Understanding and Deliberative Democracy. *Philosophy and Phenomenological Research*, *Early View*, 1–20.

Haslanger, S. (2021). Political Epistemology and Social Critique. In *Oxford Studies in Political Philosophy Volume 7* (pp. 23–65). Oxford University Press. https://doi.org/10.1093/oso/9780192897480.003.0002

Hawley, K. (2019). *How To Be Trustworthy*. Oxford University Press.

Isenberg, D. J. (1986). Group Polarization: A critical review and a Meta-Analysis. *Journal of Personality and Social Psychology*, *50*, 1141–1151.

Ivani, S., & Dutilh Novaes, C. (2022). Public engagement and argumentation in science. *European Journal for Philosophy of Science*, *12*(3), 54. https://doi.org/10.1007/s13194-022-00480-y

Kahan, D. M. (2017). *Misconceptions, Misinformation, and the Logic of Identity-Protective Cognition*. Yale Law School.

Köymen, B., & Dutilh Novaes, C. (forthcoming). Reasoning and trust: A developmental perspective. In *Why and How We Give and Ask for Reasons*. Oxford University Press.

Kuhn, T. S., & Hacking, I. (2012). *The structure of scientific revolutions* (Fourth edition). The University of Chicago Press.

Lakatos, I. (1976). *Proofs and Refutations: The Logic of Mathematical Discovery*. Cambridge University Press.

Laughlin, P. R. (2011). *Group problem solving*. Princeton University Press.

Levy, N. (2021). *Bad beliefs: Why they happen to good people* (New product). Oxford University Press.

Lewiński, M., & Dutilh Novaes, C. (Forthcoming). The many-to-many model: Communication, attention, and trust in online conversations. In *Conversations Online*. Oxford University Press.

Longino, H. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

McIntyre, L. C. (2021). *How to talk to a science denier: Conversations with flat Earthers, climate deniers, and others who defy reason*. The MIT Press.

Mercier, H. (2018). Reasoning and Argumentation. In L. Ball & V. Thomson, *Routledge International Handbook of Thinking and Reasoning* (pp. 401–414). Routledge.

Mercier, H. (2020). *Not Born Yesterday*. Princeton University Press.

Mill, J. S. (1999). *On Liberty*. Broadview Press.

Mills, C. (2015). Global White Ignorance. In M. Gross & L. McGoey, *Routledge International Handbook of Ignorance Studies* (pp. 217–227). Routledge.

Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme*, *17*, 141–161.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 175–220.

Nozick, R. (1981). *Philosophical Explanations*. Harvard University Press.

O'Connor, C., & Weatherall, J. O. (2019). *The Misinformation Age*. Yale University Press.

Olsson, E. J. (2013). A Bayesian Simulation Model of Group Deliberation and Polarization. In F. (Ed) Zenker (Ed.), *Bayesian Argumentation: The practical side of probability* (pp. 113–333). Springer.

Phelps-Roper, M. (2020). *Unfollow: A memoir of loving and leaving extremism*.

Quine, W. V. (1951). Main Trends in Recent Philosophy: Two Dogmas of Empiricism. *The Philosophical Review*, *60*(1), 20. https://doi.org/10.2307/2181906

Roloff, M. (2015). Social Exchange Theories. In *International Encyclopedia of Interpersonal Communication*. Wiley.

Saslow, E. (2018). *Rising out of hatred: The awakening of a former white nationalist* (First edition). Doubleday.

Sunstein, C. R. (2002). The Law of Group Polarization. *Journal of Political Philosophy*, *10*, 175–195.

Taber, C. S., & Lodge, M. (2006). Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science*, *50*, 755–769.

Talisse, R. (2019). *Overdoing Democracy*. Oxford University Press.

van Prooijen, J.-W., Spadaro, G., & Wang, H. (2022). Suspicion of institutions: How distrust and conspiracy theories deteriorate social relationships. *Current Opinion in Psychology*, *43*, 65–69. https://doi.org/10.1016/j.copsyc.2021.06.013

Wright, C. (1990). Wittgenstein on Mathematical Proof. *Royal Institute of Philosophy Supplement*, *28*, 79–99.

Young, I. M. (2000). *Inclusion and Democracy*. Oxford University Press.

Zollman, K. (2013). Network Epistemology: Communication in Epistemic Communities. *Philosophy Compass*, *8*, 15–27.