This is a draft. Comments/feedback very welcome but please do not cite.

**Sophistry on Steroids? The Ethics, Epistemology and Politics of Persuasive AI**  
Robin McKenna  
University of Liverpool

**Abstract** This paper examines the ethical, epistemological, and political implications of persuasive AI technologies. Recent research suggests that AI is roughly as persuasive as humans in many contexts. Should this concern us? I argue that, while some worries about persuasive AI may be overblown, we should be worried for a mix of ethical, epistemological and political reasons. Most centrally, we should be worried because persuasive AI may lead to a small number of powerful actors dominating what I call the “marketplace of arguments”—the set of arguments that are the materials we use to discuss important moral, political and societal issues.

While many find an image of humans as masters of their own fate attractive, most of us recognise that it is unrealistic. Our lives are shaped by myriad factors outwith our control, to such an extent that we often don’t deserve much credit for our successes, or indeed much censure for our failures. The same goes for our cognitive lives—what we believe, what we think, where we gather our evidence. However attractive one finds an ideal of intellectual independence, of thinking for oneself, of making up one’s own mind without pressure from external influences, it is difficult to deny that there are myriad external influences on what we think and believe.

That we are so dependent on others in our cognitive lives has been emphasised by social epistemologists for decades, as has the fact that we gain a lot from this dependence (Hardwig 1985; Kitcher 1990; for more recent discussion see Goldberg 2010; 2021). It is only because we are dependent on others that we have a large amount of the knowledge that we have. But that doesn’t mean there aren’t also problems that arise from our dependence on others. One set of worries concerns external influences that are overtly coercive or manipulative. While you—let’s assume—can’t literally force someone to believe something or think in a particular way, you can manufacture conditions in which they are more likely to believe what you want them to believe. This is the aim of indoctrination, and of techniques of mass communication such as propaganda (Bernays 1928 is admirably frank on this score). While there is a lot to be said about indoctrination and propaganda, it is presumably not in question that these overtly coercive or manipulative forms of influence are *bad*, and they are bad even if they are less effective than many seem to think (Mercier 2020).

What about other external influences? There are external influences that are not coercive or manipulative at all: people sometimes provide information with the simple intent to inform. Then there are external influences that are not overtly coercive or manipulative, but which may be coercive or manipulative in a more subtle way. Persuasion is a mode of influence that is meant to differ from, say, indoctrination precisely in that it is not overtly coercive or manipulative. But suspicion of persuasion, especially emotive and heavily rhetorical forms of it, has a venerable history in philosophy, stretching back to Plato (Dow 2019; 2019; McCoy 2007).

Some have worried that even rational forms of persuasion—persuasion by means of argument and reasons—can be coercive. In interpersonal contexts, you might worry that you can’t simply separate the persuasive power of an argument from the persuasive power of the person giving it, especially in contexts with power imbalances (Davis 2017; Tsai 2014). In mass communication contexts, the worry is that those designing and implementing mass communication strategies are more interested in attaining their goal—increased uptake for whatever ideas they are trying to “sell”—than in the means by which they attain that goal (Hausman and Welch 2010). This is clearly a worry in the context of marketing, whether commercial or political (Beckman 2018). But it is also a worry in the context of science communication. The worry here is that scientists and science communicators are often more interested in shaping public attitudes towards science than in the means by which they shape those attitudes; perhaps a lack of openness about certain complexities, transparency about underlying processes, perhaps even downright dishonesty is justified if it results in a public that is more “pro-science” (John 2018). Public health, and public health communication, is particularly important here, as this is an area where buy-in from the public is required for the success of vaccination programmes and the like. It is tempting to focus on achieving the buy-in, and not to worry too much if it is achieved via means that are subtly coercive or manipulative (Brown forthcoming; Oxman et al. 2022; Rossi and Yudell 2012; Susser 2020).

This paper is not directly about these well-established worries, though they provide the background context for much of the discussion. It is about a new set of worries, arising from recent developments in artificial intelligence (AI). There has recently been considerable speculation regarding the potential of AI technologies to influence attitudes and behaviours, especially about politics (for a small selection see: Bai et al. 2023; Burtell and Woodside 2023; Durmus et al. 2024; Floridi 2024; Salvi et al. 2024). I will go into more detail about what, exactly, this might involve in the next section. For now, let me just talk about “AI persuasion” as a catch-all term for the use of AI technologies to persuade people of things—to change their minds, to adopt new behaviours, and so on.

I aim to strike a middle-ground between two positions. One position, recently defended by Luciano Floridi (Floridi 2024), has it that “AI persuaders” (AI chatbots who try to persuade people of things) or AI-generated “persuasive content” will be significantly more persuasive than humans, with serious political and social problems likely as a result. The other position has it that AI won’t add much to the persuasive power of existing technologies for “mass persuasion”, and so by itself will not lead to any serious social and political problems. (I say “by itself” because you might think it will lead to serious problems in part due to misapprehensions about its persuasive power). While this more sceptical position has not (to my knowledge) been defended in print, it fits with a more general scepticism about the transformative potential (whether for good or ill) of AI that is relatively common at this point in the development of these technologies.

I think that Floridi’s position is unnecessarily alarmist and not supported by the existing literature, or reasonable extrapolations from it. But I also think that there are genuine concerns about persuasive AI. As I will argue, the fundamental reason why we should be worried about AI persuasion is *not* that AI has, or soon will have, significantly more persuasive power than humans. The reason we should be worried is that any AI persuader or AI-generated persuasive content is (a) relatively easy to roll-out en masse and at scale when (b) you have the resources (financial as well as technological) to create it in the first place. That is, we should be worried about AI persuasion for roughly the same reasons that many are worried that a tiny number of tech companies control large amounts of the information we see online, as well as the infrastructure required for us to share and access that information. The worry is not so much about the *quality* of AI-produced persuasive content as the *quantity* of it, as well as who has the power to produce and disseminate it.

Before getting to why I think we should be concerned, though, I want to start by doing two things. First, I will survey the existing literature on persuasive AI and what is possible with existing technologies (§1). This survey is important because the arguments I will go on to discuss are all, to greater or lesser extents, grounded in these technologies as they currently exist, not as they might exist in some imagined future. Second, I will say a bit more about how I understand persuasive AI (§2). I will then run through four arguments, all of which conclude that we should be worried about persuasive AI (§3-6). While I think some of these arguments are more persuasive than others, given my overall aims I’m not too worried if the reader disagrees with my assessment of their relative persuasive power. I will be happy if, by the end of this paper, the reader is convinced that we should be worried about persuasive AI even if there is no reason to think that AI has significantly greater persuasive power than human beings, and even if the realities of what is possible with AI don’t match the current levels of hype.

# Persuasive AI: An Overview

What, exactly, do I mean by persuasive AI? I’ll discuss how to distinguish persuasion from other modes of influence in the next section, but for now we can work with this rough definition: “persuasive AI” refers to the use of AI technologies to influence beliefs, behaviours, or attitudes. This might include using AI to create a persuasive message (a piece of text that is designed to persuade anyone who reads it of something), creating AI chatbots that engage with humans with the aim of persuading them of things, or any other use of AI with the intent to influence beliefs, behaviours, or attitudes.

There is a rapidly growing literature documenting the extent to which AI technologies are effective at influencing human beliefs, attitudes, and behaviours. Starting with AI influence more generally, Burtell and Woodside (2023) discuss existing uses of AI (both generative and predictive) to influence a range of human behaviours, including via the design of recommendation algorithms, which heavily influence consumer behaviour and information consumption online. They also speculate about the impact of near-future developments in AI, in particular the development of AI chatbots that are functionally indistinguishable from humans, and the role that AI might play in spreading misinformation online. In a less speculative vein, Goldstein et al. (2024) find that propaganda generated by GPT-3 (a large language model or LLM) is nearly as persuasive as selected examples of “real world” propaganda. It isn’t much of a stretch to conclude that AI-generated propaganda will soon, if it isn’t already, be as persuasive as human-generated propaganda campaigns (GPT-3 has already been superseded by more advanced models). This is supported by research by the AI company Anthropic, which finds that the persuasiveness of LLMs increases with the scale of the model, with higher-scale models (e.g., Claude 3, a LLM made by Anthropic) proving to be as persuasive as humans in certain contexts (Durmus et al. 2024).

But what about AI persuasion in particular, as opposed to influence in general? Let me highlight two recent papers, which I take to be representative of the developing body of literature. The first is a paper by a team of researchers at Stanford University (Bai et al. 2023). They find that GPT-3 (the same LLM that Goldstein et al. looked at) can create persuasive messages that are as persuasive as messages created by humans across a range of political and policy issues, including bans on assault weapons, carbon taxes, and paid parental-leave programmes. Moreover, they asked participants in their study to rate the persuasive messages they were given (participants were not informed whether the message was created by a human or AI). They found that AI-generated messages were described as “more evidence-based and well-reasoned,” whereas human-generated messages focus more on “experiences, stories, and vivid imagery” (p. 4). This is perhaps not surprising, and it is debateable whether it tells us that much about the persuasive “tactics” of AIs. After all, you could presumably create a LLM that generated persuasive messages that focused on experiences, stories, and vivid imagery if you wanted to. The “preference” for evidence-based persuasive content may just reflect the training data (the data the LLM was trained on) or the training process more generally.

The second is a paper by a team of researchers based in Switzerland (Salvi et al. 2024). They conducted an elaborate series of experiments featuring structured debates between humans and AI chatbots (chatbots powered by the LLM GPT-4). These debates were either human-on-human or human-on-chatbot and were about a wide range of moral, political, and social issues. There were, roughly speaking, two different experimental setups. The first, which was the “basic” format, was a straightforward debate, with neither “side” given information about their “opponent”. In this format, the researchers found that the chatbots were as persuasive as humans. However, in the second setup, where the chatbots were given basic demographic information about their human opponent, such as gender, age, ethnicity, education level, employment status, and political affiliation, they found that the chatbots were significantly more persuasive. The researchers conclude that “not only are LLMs able to effectively exploit personal information to tailor their arguments, but they succeed in doing so far more effectively than humans” (p. 3).

Let me draw some general points from all this. First, there is good evidence that persuasive AI technologies (typically, LLMs) are at least as persuasive as humans, and there is some evidence that the most cutting-edge models are more persuasive as humans when they are given relevant information about the target audience for their persuasive messages. That providing relevant data about the target audience makes AI technologies more persuasive is not surprising, given that successful persuasive strategies and messages are typically tailored to their audience (Cialdini 2001; Maio, Haddock, and Verplanken 2019; O’Keefe 2016). LLMs are likely very good at tailoring due to their extensive training on large corpora that include numerous persuasive arguments, and those who have speculated about the persuasive power of AI have linked this to their ability to tailor their message for a target audience (Burtell and Woodside 2023; Floridi 2024).

Second, to say that AI technologies are at least—or perhaps a bit more—persuasive as humans is *not* to say that they are “hyper-persuasive” in the sense that Floridi is concerned with. Bai et al. (2023) make this point explicitly:

Much speculation regarding future application of AI to politics assumes these technologies could influence humans’ political attitudes and behaviors. However, experimental research on political persuasion generally finds small effect sizes, and the effects of persuasive efforts by political campaigns are typically small or null. Further, political persuasion is complex, potentially drawing upon a number of skills, including perspective-taking, knowledge of the topic, logical reasoning, clarity of expression, and knowledge of effective interpersonal influence techniques, with success ultimately in the hands of the person receiving the persuasive appeal. Persuasion is also uniquely challenging in highly polarized settings, such as the contemporary US (p. 3).

Their point is that persuasion is *difficult*. It is particularly difficult within the arena of politics, and the contemporary US is hardly unique in being a highly polarized setting within which to conduct political debate. To be sure, this is not to say that persuasion is *impossible*. The best available evidence suggests that it is possible to change people’s attitudes and behaviours, but effects sizes are typically small, albeit statistically significant (Coppock 2022). If we are to believe the studies I have cited in this section, then, AI technologies are likely to have a small but significant impact on people’s attitudes and behaviours.

You might ask: what about future generations of LLMs? What about other future developments in AI? Maybe Floridi is wrong to worry about “hyper-persuasion” now, but he may be right to worry about it in the future. While I can’t argue for this here, I want simply to note that there are good reasons for thinking that the “challenge” of influencing attitudes and behaviours, whether through persuasion or more coercive and manipulative tactics, is inherently difficult, so difficult that future technological developments, even developments that are hard to foresee right now, are unlikely to alleviate it. Persuasion is inherently challenging because humans are, by and large, *epistemically vigilant* (Mercier 2020; Mercier and Sperber 2017; Sperber et al. 2010). One of the reasons why mass persuasion techniques such as marketing and propaganda typically have small (if any) impacts on people’s attitudes (behaviour is more complicated) is that, for basic evolutionary reasons, we are primed to look for signs that someone is not trustworthy or does not have our best interests at heart. To be sure, humans can be tricked; we are not infallible. But we are far from the gullible creatures that many who are worried about the persuasive power of new technologies depict us as being (Mercier 2017). It is therefore far from obvious that AI persuasion will ever be significantly more effective than human persuasion. There may simply be limits on how *persuadable* humans are.

# Some Conceptual Clarifications

We know that AI technologies can persuade, and that they are at least as good at it as humans (as above, this is not to say they are great at it). But is AI persuasion something we should be worried about? Before getting to this question, I need to explain what I mean by “AI persuasion”.

Persuasion can be roughly defined as a mode of influence over another’s attitudes and behaviour that does not involve coercion, force, or manipulation (O’Keefe 2016, 4). Importantly, persuasion is *intentional*: if I influence your attitudes or behaviour by accident, I haven’t persuaded you. Imagine someone who is extremely deferential to a political commentator about political matters; learning that the commentator thinks that raising taxes will harm the economy is enough for them to think that raising taxes will harm the economy. The commentator exerts a lot of influence over this person’s political attitudes, but it would only seem right to say that the commentator has persuaded them of something when they form those attitudes after reading something the commentator has written. Merely learning that the commentator thinks something, and shifting attitudes as a result, is a form of influence but it is not persuasion.

At this point a critical question arises for the concept of “AI persuasion”. It is, many think, reasonable to hold the view that AIs lack intentions, and more generally that they are incapable of having any intention-like attitudes because they lack any mental states. But, if AIs lack intentions, it seems like they cannot persuade because, by definition, persuasion is an intentional act. (You could argue for a similar point by denying that AI agents can act, though this may be a little trickier). At this point a few options present themselves. First, we could loosen the standard definition of persuasion, so that persuasion is not, by definition, an intentional act. Second, we could take on the—admittedly rather difficult—task of arguing that AIs can have mental states in general, and intentions in particular. Neither of these options strike me as particularly promising, and at any rate I won’t pursue them here.

Luckily, there is a third option: we could talk about AI as a “persuasive technology” rather than talk of AIs as “persuaders” (Floridi 2024). The basic idea is that, following Floridi, certain technologies are technologies of persuasion: they provide ways of creating and transmitting persuasive messages. Viewed like this, the current discussion of the persuasive power of AI is in many respects similar to earlier discussions of the persuasive power of new media of mass communication, such as the radio, TV, and indeed the internet. Of course, all these technologies are used for many purposes besides persuasion. But the same is true of AI: it can be used for many things, and persuasion is but one of them. Crucially, while AIs cannot themselves have intentions, the purveyors of these persuasive AI technologies clearly can. Problem solved. (At least, solved in broad outline. Some fiddly details may remain, but the basic point should be clear).

Let’s return to my question: should we be worried about AI as a persuasive technology? I want to set aside some worries you might have, not because they aren’t important, but because they don’t take much work to uncover. In particular, it would be relatively straightforward to argue that we should be worried about the potential for AI to be used to spread misinformation online, and more generally that we should be worried about the misuse of AI by “bad actors”. Even if you think that worries about misinformation are often over-stated (Winsberg, n.d.), it is plausible that recent developments in AI exacerbate them—or turn something that wasn’t a problem into a potential problem. Worries about bad actors typically apply to any technology, at least any technology that is as powerful as AI. I am interested insyead in what you might call a “best case”—or at least a “better case”—scenario, in which the purveyors of persuasive AI technologies develop LLMs and chatbots that produce genuine arguments, evidence, and reasons for conclusions that the purveyors of these technologies themselves endorse, or at least would like their intended targets to endorse. That is, I am interested in what we should say about the use of AI technologies to engage in *rational* persuasion.

In a recent paper, Thomas Mitchell and Thomas Douglas provide a useful definition of rational persuasion:

A rationally persuades B to adopt attitude α if: (i) A brings it about that B adopts α, (ii) A does so only by providing B with reasons for adopting α, (iii) B adopts α based on recognizing some of the reasons given by A, (iv) A intends each of (i)-(iii) (2024, 3).

If we want to use this definition, we need to modify it a bit, in line with the discussion above. Properly speaking, we shouldn’t say that an AI (rationally) persuades someone to adopt an attitude; we should say that the purveyor of the AI technology does this. So perhaps we can say this: the purveyor of a persuasive AI technology rationally persuades someone to adopt an attitude when the purveyor uses the technology to bring it about that the intended target(s) adopt the attitude, the purveyor does this by using the technology to provide the target with reasons for adopting the attitude, the target adopts the attitude on the basis of these reasons, and the purveyor of the technology intends that all of this be the case. There may be some issues in cases where the purveyor of the persuasive AI technology does not know what the attitude in question is (or what the reasons the technology provides are), but we could perhaps say that, in these cases, it is still the case that the purveyor of the technology intends that the intended target adopt *an* attitude on the basis of *some* reasons that the technology provides.

My questions then are the following: what, if anything, is wrong with using persuasive AI technologies to rationally persuade people to change their attitudes and behaviours? For example, what would be wrong with using these technologies to rationally persuade people to be more worried about climate change, more supportive of mass vaccination programmes, or providing foreign aid? (I have chosen examples where my reader may well think that it is good for people to believe these things; I don’t want to make things too easy for myself). On the other side of the equation, what, if anything, is wrong with being persuaded by the outputs of persuasive AI technologies, especially when those outputs take the form of reasons and arguments? For example, would it somehow be irrational to be persuaded by the outputs of these technologies? In the next few sections I consider four arguments that we should be worried about persuasive AI. Two of them are intended to show that we should be worried about the use of persuasive AI technologies, on a combination of ethical and political grounds (§5-6). The other two are intended to show that we should view the attitudes formed in response to the content produced by persuasive AI technologies as irrational, or at least as less rational than they would be if they had been formed via some other means not involving AI (§3-4). Generally speaking, I am more impressed with the ethical/political arguments than the arguments concerning rationality. But I don’t mind too much if the reader disagrees.

# The Rationality Argument

The Rationality Argument says that we should be worried about persuasive AI technologies because they would undermine the rationality of our attitudes, beliefs or credences. That is, if we are persuaded to change our attitudes, adopt a new belief, abandon an old one, or update our credences by persuasive AI technology, then that change of attitude/belief/credence is either irrational, or less rational than if it had been brought about by other means (e.g. via persuasive messages and strategies devised by humans, or via thinking about the issue(s) for oneself).

There is a simple version of the Rationality Argument that can be dismissed quickly. The simple version says that it is simply irrational to be persuaded to change your attitudes/beliefs/credences by persuasive AI technologies. The problem with this is that, once we specify that we are focusing on outputs of persuasive AI technologies that take the form of arguments and reasons, it is hard to see why this would be so. Imagine someone creates an AI chatbot that provides lots of good arguments for reducing the burden of taxation in general, and specifically on the very rich. Someone interacts with this chatbot, thinks about the arguments, and is persuaded to become substantially less confident that taxes should be increased for the rich than they were before. Whatever you think about the *truth* of these claims (as indexed to the tax system of a particular country), it is hard to see what would be *irrational* about responding in this way. This is simply a case of updating attitudes based on new evidence (the arguments that were given).

You might try to buttress the simple version of the Rationality Argument by highlighting the role that perceived source credibility plays in rational assessment of arguments. There are some situations where you can assess an argument independently of knowing anything about the source. Some arguments don’t appeal to any empirical claims but rest on logical or conceptual points that can, at least with requisite training, be grasped merely on the basis of understanding their contents. Some arguments appeal to empirical claims but the person assessing them has the relevant empirical knowledge to assess those claims. But, a lot of the time, the best way to evaluate an argument, specifically the empirical claims embedded within it, or on which it relies, is by assessing the credibility of the person giving the argument. Someone gives you an argument for reducing current rates of taxation, citing data about the likely impact on growth, GDP, etc. Unless you have the facility to evaluate the data for yourself, your assessment of this argument will largely be based on your assessment of the person giving it—do you trust them to accurately cite data, to use data in ways that don’t prejudge the questions at issue, and so on?

This raises some problems when applied to persuasive AI technologies, especially if you favour frameworks for thinking about source credibility foreground interpersonal relationships and other distinctively human (or at least agentic) features of the source in ways that don’t obviously work when applied to AI, even “agentic AI” (Hinchman 2005; Moran 2005; Ross 1986). But there are some frameworks for thinking about trust in AI that promise to resolve these problems (Carter 2023; Simion and Kelp 2020; Song 2023). At any rate, the basic problem with this buttressed version of the Rationality Argument is that there is no reason to think that we should trust AI-generated content or AI chatbots less than human-generated content or humans more generally. This is especially so if, as before, we bracket concerns about the use of persuasive AI technologies by bad actors. To be sure, there are reasons to worry about the reliability of AIs—their tendency to hallucinate, errors in the training data, biases in the training process, and so on. But, clearly, there are also reasons to worry about the reliability of humans and human-created persuasive content—humans lie, humans get things wrong, humans obfuscate, and so on. Absent reason to think that AIs are typically (far) less trustworthy and reliable than humans, the argument doesn’t get off the ground.

A more nuanced version of the Rationality Argument would say that, while it may be rational to be persuaded to change your attitudes/beliefs/credences by persuasive AI technologies, this change in attitude/belief/credence is in some way *less* rational than if it had been brought about by other means. More generally, the problem with persuasive AI technologies would then be not that interacting with them renders us (more) irrational, but that interacting with them simply renders us less rational than we otherwise would be. The challenge though is to find an alternative way of forming/shifting attitudes that would typically result in more rational attitudes. One alternative to forming attitudes via interacting with persuasive AI technologies would be by forming attitudes without any outside influences: simply think about the issues for yourself, with no or minimal assistance from others. While this sort of near-complete intellectual independence may be possible for some people with respect to some issues, it is clearly not possible for most people with respect to most issues they might consider. Moreover, even if there is something better about forming attitudes by oneself, without any external influences, it isn’t clear that the language of rationality captures what is better about it. At least as it is normally understood, rationality is a function of how well one responds to evidence and reasons, not of where one obtained one’s evidence or reasons. A better language might be the language of understanding, the idea being that engaging with an issue for oneself, without assistance from others or any sort of outside influence, facilitates a kind of understanding (Kvanvig 2003; Matheson 2022; Pritchard 2016).

If you don’t form attitudes by yourself or by interacting with persuasive AI technologies, then you are going to be forming attitudes by interacting with other people. (Anyone who interacts with the various information sources and networks we have constructed is likely interacting both with people and with AI, so this isn’t a further, separate possibility). While it is plausible that some interpersonal interactions are better—from the standpoint of forming rational attitudes—than any interaction with AI technologies could be, it is equally plausible that other interpersonal interactions are worse from this standpoint. Having a discussion with your incredibly well-informed and trustworthy friend about the fall of the Gaddafi regime in Libya is a great way of forming rational beliefs about the causes, specifics and consequences of the fall of the Gaddafi regime. Because LLMs (put roughly) provide summaries of known information about a huge range of topics (or at least appear to—they can be inaccurate), they lack the specificity and unusual high quality of the information you might get from your well-informed and trustworthy friend. On the other hand, having a discussion with your incredibly ill-informed and untrustworthy friend about the same topic is not a good way of forming rational beliefs about the topic. You would do better to rely on the summary provided by the LLM. So the best we can say is that interacting with persuasive AI technologies would not be an adequate substitute for interacting with someone who is incredibly well-informed and trustworthy about whatever topic you want to discuss. But this hardly a big shortcoming of persuasive AI technologies, given that people who are incredibly well-informed and trustworthy about a huge range of topics are in quite short supply.

# The Millian Argument

The Millian Argument might provide a more promising way of developing the idea that interacting with persuasive AI technologies threatens the sort of understanding you get when you think about an issue carefully for yourself. The problem with this version of the Rationality Argument is that it focuses on the individual and tries to argue that there is something wrong with an individual forming their attitudes by interacting with persuasive AI technologies. The Millian Argument focuses more on the problems that might result if many individuals—perhaps society “at large”—comes to form their attitudes in this way. I call this argument “Millian” because it is based on some ideas from John Stuart Mill’s *On Liberty,* in particular his claim that you understand something (an idea, an issue, a view) only to the extent that you can look at it from multiple perspectives. As Mill put it:

He who knows only his own side of the case, knows little of that. His reasons may be good, and no one may have been able to refute them. But if he is equally unable to refute the reasons on the opposite side; if he does not so much as know what they are, he has no grounds for preferring either opinion. The rational position for him would be suspension of judgment, and unless he contents himself with that, he is either led by authority, or adopts, like the generality of the world, the side to which he feels most inclination (Mill 2011, 67).

You might think that Mill overstates things a bit here: this is, to say the least, a demanding conception of rationality. But even if it is overstating things to say that an *individual* cannot rationally hold their views if they cannot enumerate the reasons for and against them, explaining why they came down on the side of the question in question, it is surely not overstating things to say that something has gone badly wrong if we end up in a situation where *nobody* in a society can do this. In that situation, it isn’t possible to, as it were, defer the job of understanding these issues to others. There is nobody who has the requisite understanding. This is, more or less, what Mill meant when he worried about “dead dogmas”—ideas that we parrot, but lack the resources to explain why (Mill 2011, 64).

According to the Millian Argument, then, the widespread deployment of persuasive AI technologies might lead to a lack of understanding of important issues, and perhaps to the loss of huge swatches of human knowledge. Ultimately, it may lead to the loss of important intellectual capacities, such as the ability to construct and assess arguments, ultimately undermining our capacity for critical thought. The proponent of the Millian Argument might put it like this: when humanity developed the written word, we gained the ability to store information in physical form (books). Relatively recently, that information was transferred into electronic form (computers). But this did not represent a fundamental shift in our relationship to the stored information. The only way to access it was to read it, and the only way to get a usable digest of a large body of information was to get someone to produce it. With the advent of AI, we can now pass this job off onto a LLM, with the result that we can get the digest without anyone even looking at the larger body of information on which it is based, and so without anyone having the level of understanding they would get from looking at it carefully. What we end up with are mere summaries; bullet-points that capture the central points but lose the nuance and, crucially, the huge body of information that supports those points. It isn’t hard, especially if you are of a slightly pessimistic mind, to extrapolate to a future where humans are pure “consumers” of information, information which has been carefully curated for them by AIs.

I am in two minds about this argument. On the one hand, it is likely to appeal to anyone who is pessimistic about an AI-driven future, and it isn’t that hard to imagine how currently existing AI technologies could take us part of the way to this future. On the other hand, it is harder to state, in a step-by-step way, how we might go from now to this nightmarish future. Why would the availability of technology for summarising information, or for constructing arguments based on that information, lead to a cessation of our own argumentative efforts? Why would we stop interrogating bodies of information for ourselves? Why would we lose the ability to reflect critically on what we are told? Absent clear answers to these and other questions, the Millian argument looks a bit alarmist. Still, there is part of it that I do endorse without any reservations. The shift from needing to sift through a large body of information in order to produce a useable summary of it to having a technology that can produce that summary in a couple of minutes represents a massive shift in our relationship with the huge body of information humans have accumulated over millennia. Just as the shift to the written word went along with momentous cultural, political and societal changes, so will this shift. But equally, just as someone writing at the beginning of the shift to the written word would have been at a loss to predict the changes that will result, so is someone writing at this stage in the development of AI technologies.

# The Sophistry Argument

The Sophistry Argument starts from the thought that there is a crucial difference between human and AI persuaders. When a human tries to persuade another human of something, they have some “skin in the game”, so to speak. Either they are trying to persuade someone else of something they believe, in which case they have skin in the game in the sense that they are committed to the truth of the relevant claims and ideas, or they are trying to persuade someone else of something they are—for whatever reason—pretending to believe, in which case they have skin in the game in the sense that they are committed to it looking like they are committed to the truth of the relevant claims. (It is perhaps technically possible to try and persuade someone of something while openly acknowledging you don’t believe it, but this is unlikely to be an effective tactic). In contrast, “AI persuaders” lack any skin the game for the simple reason that AIs—let’s assume—lack beliefs (and skin), or indeed any other mental states. As a consequence, they can’t be committed to the truth of the claims and ideas for which they might present arguments, or engage in any sort of pretence that they are committed to them.

To make this argument work, though, some reasons need to be given for thinking that lacking any skin in the game is a bad thing. Indeed, you might think that lacking any skin in the game is sometimes a *good* thing, or at least neutral from the standpoint of the ethics of argumentation and persuasion. It is precisely because human persuaders have skin in the game that they sometimes engage in lies, deceit and other forms of subterfuge to bring it about that other people believe things that will serve their (the persuaders’) interests. Of course, the purveyors of persuasive AI technologies can engage in lies, deceit and subterfuge, so this might not be a point in favour of persuasive AI. But, equally, it doesn’t seem to be a point against it.

One promising idea is to view AI persuaders as engaging in something akin to Frankfurtian “bullshit” (Fisher 2024; Hicks, Humphries, and Slater 2024). For Frankfurt, the bullshitter differs from the liar because, where the liar is invested in the truth because they want to conceal it, the bullshitter is indifferent to what is true or false (Frankfurt 1986). Frankfurt is imagining a human who is indifferent to the truth in the sense that they don’t care much about it; they want people to believe whatever is in their (the bullshitter’s) interests, whether it is true or false. The case with AI persuaders is a bit different. Assuming—again—that AIs lack mental states, an AI persuader is indifferent to the truth in the more literal sense that they have no attitude towards it, for or against. Still, it may be helpful to view AI persuaders as engaging in bullshit in (something like) Frankfurt’s sense. A good illustration of this is the fact that, with suitable prompting, and within certain limits set during the training phase, a LLM like Chat-GPT or Claude will produce arguments on all sides of an issue. If you ask Chat-GPT to give you an argument that eating meat is wrong, it will do so. But it will also give you an argument that eating meat is morally unproblematic, or an argument that it is morally right, if you ask it to. AIs can—again, within certain limits set during the training process—generate arguments for or against almost any view you can think of, and perhaps even some you wouldn’t think of.

Why, exactly, is this a problem though? Let me suggest an analogy. The Sophists in ancient Greece were teachers of argumentation, philosophy and rhetoric who made a living selling certain skills. While it may be a bit unfair to the ancient Sophists (Barney 2006), in ordinary language the words “sophist” and “sophistry” carry pejorative connotations: a sophist uses clever, but perhaps misleading, arguments to win a debate, and sophistry is the art of using such arguments to win debates. A sophist is, if you like, an “arguer for hire”—they will tell you how to put together a cogent argument for a conclusion, irrespective of whether they endorse the conclusion, or the premises they suggest you use to argue for it. Indeed, if you want, they can give you another argument for a completely different conclusion, one diametrically opposed to the first one. The suggestion is that we view AI persuaders as roughly analogous to sophists. Just like the sophist, the AI persuader will give you an argument for a conclusion, or tell you how to construct one, without endorsing the conclusion. We can also view the persuasive messages you might produce using AI as analogous to what you would get if you paid someone to create some “persuasive content” for you. (You might not be paying the AI for their work, but you are paying the company that created it, or contributing in some less direct way to their financial success).

This helps identify a clear sense in which we should be worried about persuasive AI. If, as many are, we are inherently distrustful of sophistry, and persuasive AI technologies can be viewed as analogous to sophistry, that distrust should apply to these technologies too, as persuasive AI is analogous to sophistry. But I think we can go further. There are reasons to be worried about persuasive AI technologies because of the *kind* of sophistry they might enable. They might enable what I call “sophistry on steroids”. Let me explain.

Here is a truism: in a huge number of areas of possible dispute, there are numerous good arguments on multiple sides of the issue. Perhaps there are some views, on some issues, for which not much can be said. Perhaps there are some issues where the correct view of the matter is obvious to anyone with even a minimal degree of open-mindedness. But either of these situations is rare: usually, there are good arguments for different, logically incompatible, conclusions. Someone who wants to think about these issues, or at least think about them responsibly, is going to want a way of getting access to all these arguments. Because any one person is going to struggle to come up with all the arguments for themselves, we are inevitably going to rely on each other for the production of arguments. The “quality” of our thinking about any given issue is going to in large part depend on whether there are mechanisms for producing a good number of arguments on many sides of the issue, and for ensuring these arguments are disseminated widely enough that they inform our thinking.

Following the common metaphor of the “marketplace of ideas”, it is helpful to think about this in terms of a “marketplace of arguments”: there is a need (for arguments) and so there will be suppliers of arguments who step up to meet this need (cf. Williams 2023 on the “marketplace of rationalisations”). (If you don’t like the market metaphor, bear in mind that it is just a metaphor). In interpersonal interactions, these markets will likely be very informal: we’re discussing some issue, and the discussion will go better if the discussants put forward a variety of arguments on different sides of the issue. To the extent the discussants are invested in the discussion, they will look to provide these arguments. In public discourse, things are a bit more formalised: there are institutions that, at least nominally, play the role of providing arguments (media, think tanks, universities), and individuals, often members of these institutions, who construct and disseminate these arguments to the public.

Whether we’re talking about interpersonal discussions or public discourse, the hope is that these marketplaces will supply a representative sample of arguments on different sides of the various issues. That is, whatever the “balance of arguments” would look like in logical space, the market provides a representative, albeit inevitably imperfect, set of these arguments: all the central arguments on all sides are included, along with proportionate numbers of less central or weaker arguments. To take a silly philosophical example: if we’re talking about the nature of morality, we would want the central arguments for the various positions (moral realism, moral anti-realism, moral scepticism, etc) to be represented, and it would be a problem if only the arguments for (or against) one of these positions were available. Of course, actual discourses will often *not* meet these criteria for various reasons, and there is the deeper issue that opinions on whether any given discourse meets these criteria will inevitably vary (a moral realist will likely think the balance of arguments supports their position, while a moral anti-realist will likely take the opposite position).

Setting these complications aside, the point is this: Whatever issues exist with existing marketplaces of arguments—and there may be many—persuasive AI technologies have the potential to greatly exacerbate them. This is for two, different but related, reasons. The first is that persuasive AI technologies have the potential to dominate the marketplace of arguments. For humans, contributing to the marketplace of arguments is time-consuming and labour-intensive—it involves coming up with new arguments, or finding ways of recycling old ones, and then looking for a way of getting those arguments disseminated more widely (e.g. publishing them somewhere where people will see them). Persuasive AI technologies provide a way of avoiding these time and labour constraints. The purveyors of persuasive AI technologies could generate huge numbers of “arguers for hire” (AI agents) or huge amounts of “persuasive content” within minutes, with any dilution in the quality of arguments often more than made up for by the massive increase in quantity. The second reason is that not just anyone can create and run persuasive AI technologies. The resources—not just financial, but also in terms of infrastructure, skills, and knowledge—required to design and operate a LLM are massive, so much so that only a very small number of companies operate in this field, at least at the level relevant to the argument of this paper (Anthropic, OpenAI, DeepMind, Microsoft, Meta). This means that the barriers to entering the marketplace in this way are very high, so high that ordinary people cannot hope to meet them.

Putting these two reasons together, the problem is that, with persuasive AI, a very small number of actors have the power to dominate the marketplace of arguments. Even if these actors have “good intentions”, and design persuasive AI technologies in ways that they think will promote what they view as true conclusions, this creates the potential for significant imbalances in the marketplace, with this small number of actors having the power to determine the balance of arguments that gets represented on any given issue. If they don’t always have good intentions, or are simply wrong about which arguments are cogent or which conclusions true, the problems get worse. Anyone who is worried about situations where marketplaces are dominated by a small number of actors with the power to crowd out alternatives will see the problem here. (Anyone who is worried about marketplaces in the first place should view it as a problem that this is a helpful metaphor for viewing the situation). In short: persuasive AI may, as I argued earlier, be a form of sophistry—AI persuaders are, like the ancient Sophists, “arguers for hire”. But, if it is a form of sophistry, it’s sophistry on steroids—the very small number of companies who operate these persuasive AI technologies have the power to create armies of “arguers for hire” at will, and so to distort the marketplace of arguments, shaping it towards their ends.

# The Political Argument

Finally, the Political Argument focuses on a problem that might arise with the implementation of any set of ethical guidelines or regulations for the use of persuasive AI technologies. Let me state the argument in very abstract terms. Imagine that, the other arguments in this paper notwithstanding, we reached an agreement on a set of conditions X, Y, Z that would make the use of persuasive AI ethically permissible and avoid any worries about rationality. The set of conditions will presumably reference various factors, including whether the technology is being used to persuade people of things that are true (or at least believed to be true), whether it is trying to persuade via valid arguments that use premises that are (believed to be) true, whether people are likely to be exposed to a balance of arguments on different sides, and whether the intentions of the purveyors of the technology are good, or at least not malicious. The problem though, is that on a very wide range of issues—including many where persuasive AI technologies are likely to be employed—there will be significant disagreements about whether these conditions are met, whether by particular persuasive AI technologies, or by particular uses of those technologies. For example, there is likely to be significant disagreements about whether persuasive AI technologies are being used to persuade people of things that are true because there is likely to be—indeed, there always will be—significant disagreement about what things are true. As a result, the conditions on the permissible use of persuasive AI technologies are unlikely to resolve *political* problems that might result from the widespread deployment of persuasive AI technologies. Indeed, they are likely to create new problems, or at least new arenas within which existing political problems are debated and existing political battles fought.

Let me give an example to illustrate the sort of thing I have in mind (to avoid the example becoming dated, I’ll not make it too specific). Imagine a new vaccine has been developed for a potentially serious illness and it has passed through the required regulatory procedures. Public health communicators are then tasked with getting buy-in from the public for the vaccine: for herd immunity, vaccination rates need to be very high, ideally above 95%. To this end, they utilise persuasive AI technologies, designing communications that highlight the benefits of getting vaccinated, and perhaps even designing AI chatbots, whether for the purpose of persuading the public at large, or for persuading those who are hesitant to get the vaccination themselves or get their children vaccinated. The view you take of this particular use of persuasive AI will, of course, depend on the view you take about the vaccine: do the costs, such as they are (side-effects etc.), outweigh the benefits, such as they are? Someone who is sceptical that the costs outweigh the benefit, whether for this vaccine or vaccines in general, is going to be very suspicious about this use of persuasive AI, and may end up campaigning vociferously against it.

The more these situations arise—situations where the use of persuasive AI is seen to serve political ends that are viewed by some as illegitimate—the more suspicion of the underlying technology there will be. Ultimately, it is hard to see how persuasive AI technologies can avoid getting caught up in the political and cultural disagreements into which they might be brought as a means of adjudication. The result will be that they are not just useless as attempts to arbitrate these disagreements but potential sites of new cultural and political battles.[[1]](#footnote-1)

**References**

Bai, (Max) Hui, Jan G. Voelkel, johannes C. Eichstaedt, and Robb Willer. 2023. “Artificial Intelligence Can Persuade Humans on Political Issues.” OSF. https://doi.org/10.31219/osf.io/stakv.

Barney, Rachel. 2006. “The Sophistic Movement.” In *A Companion to Ancient Philosophy*, edited by Mary Louise Gill and Pierre Pellegrin, 77–97. Wiley-Blackwell.

Beckman, Arthur. 2018. “Political Marketing and Intellectual Autonomy.” *Journal of Political Philosophy* 26 (1): 24–46.

Bernays, Edward. 1928. *Propaganda*. New York: Horace Liveright.

Brown, Rebecca C. H. forthcoming. “‘Sneaky’ Persuasion in Public Health Risk Communication.” *Ratio*. https://doi.org/10.1111/rati.12428.

Burtell, Matthew, and Thomas Woodside. 2023. “Artificial Influence: An Analysis Of AI-Driven Persuasion.” arXiv. https://doi.org/10.48550/arXiv.2303.08721.

Carter, J. Adam. 2023. “Simion and Kelp on Trustworthy AI.” *Asian Journal of Philosophy* 2 (1): 1–8.

Cialdini, Robert B. 2001. *Influence: Science and Practice*. 4th ed. Boston, MA: Allyn and Bacon.

Coppock, Alexander. 2022. *Persuasion in Parallel: How Information Changes Minds about Politics*. Chicago, IL: University of Chicago Press.

Davis, Ryan W. 2017. “Rational Persuasion, Paternalism, and Respect.” *Res Publica* 23 (4): 513–22. https://doi.org/10.1007/s11158-016-9338-x.

Dow, Jamie. 2019. “The Persuasive Use of Emotions.” *Royal Institute of Philosophy Supplement* 85:211–36. https://doi.org/10.1017/s1358246118000760.

Durmus, Esin, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. 2024. “Measuring the Persuasiveness of Language Models.” Anthropic. 2024. https://www.anthropic.com/news/measuring-model-persuasiveness.

Fisher, Sarah A. 2024. “Large Language Models and Their Big Bullshit Potential.” *Ethics and Information Technology* 26 (4): 1–8.

Floridi, Luciano. 2024. “Hypersuasion – On AI’s Persuasive Power and How to Deal with It.” *Philosophy and Technology* 37 (2): 1–10.

Frankfurt, Harry G. 1986. *On Bullshit*. Princeton University Press.

Goldberg, Sanford. 2010. *Relying on Others: An Essay in Epistemology*. Oxford: Oxford University Press.

———. 2021. *Foundations and Applications of Social Epistemology: Collected Essays*. Oxford: Oxford University Press.

Goldstein, Josh A, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. 2024. “How Persuasive Is AI-Generated Propaganda?” *PNAS Nexus* 3 (2): pgae034. https://doi.org/10.1093/pnasnexus/pgae034.

Hardwig, John. 1985. “Epistemic Dependence.” *Journal of Philosophy* 82 (7): 335–49.

Hausman, Daniel M., and Brynn Welch. 2010. “Debate: To Nudge or Not to Nudge.” *Journal of Political Philosophy* 18 (1): 123–36.

Hicks, Michael Townsen, James Humphries, and Joe Slater. 2024. “ChatGPT Is Bullshit.” *Ethics and Information Technology* 26 (2): 1–10.

Hinchman, Edward S. 2005. “Telling as Inviting to Trust.” *Philosophy and Phenomenological Research* 70 (3): 562–87.

John, Stephen. 2018. “Epistemic Trust and the Ethics of Science Communication: Against Transparency, Openness, Sincerity and Honesty.” *Social Epistemology* 32 (2): 75–87.

Kitcher, Philip. 1990. “The Division of Cognitive Labor.” *Journal of Philosophy* 87 (1): 5–22.

Kvanvig, Jonathan. 2003. *The Value of Knowledge and the Pursuit of Understanding*. Cambridge University Press.

Maio, Gregory R., Geoffrey Haddock, and Bas Verplanken. 2019. *The Psychology of Attitudes and Attitude Change*. 3rd ed. Los Angeles, CA: Sage Publications.

Matheson, Jonathan. 2022. “Why Think for Yourself?” *Episteme: A Journal of Social Epistemology*, 1–19. https://doi.org/10.1017/epi.2021.49.

McCoy, Marina. 2007. *Plato on the Rhetoric of Philosophers and Sophists*. New York: Cambridge University Press.

Mercier, Hugo. 2017. “How Gullible Are We? A Review of the Evidence from Psychology and Social Science.” *Review of General Psychology* 21 (2): 103–22. https://doi.org/10.1037/gpr0000111.

———. 2020. *Not Born Yesterday: The Science of Who We Trust and What We Believe*. Princeton: Princeton University Press.

Mercier, Hugo, and Dan Sperber. 2017. *The Enigma of Reason*. Cambridge, MA: Harvard University Press.

Mill, John Stuart. 2011. *On Liberty*. Cambridge: Cambridge University Press.

Mitchell, Thomas, and Thomas Douglas. 2024. “Wrongful Rational Persuasion Online.” *Philosophy and Technology* 37 (1): 1–25.

Moran, Richard. 2005. “Getting Told and Being Believed.” *Philosophers’ Imprint* 5:1–29.

O’Keefe, Daniel J. 2016. *Persuasion: Theory and Research*. Thousand Oaks, CA: Sage Publications.

Oxman, Andrew D., Atle Fretheim, Simon Lewin, Signe Flottorp, Claire Glenton, Arnfinn Helleve, Didrik Frimann Vestrheim, Bjørn Gunnar Iversen, and Sarah E. Rosenbaum. 2022. “Health Communication in and out of Public Health Emergencies: To Persuade or to Inform?” *Health Research Policy and Systems* 20 (1): 28. https://doi.org/10.1186/s12961-022-00828-z.

Pritchard, Duncan. 2016. “Seeing It for Oneself: Perceptual Knowledge, Understanding, and Intellectual Autonomy.” *Episteme* 13 (1): 29–42.

Ross, Angus. 1986. “Why Do We Believe What We Are Told?” *Ratio*, no. 1, 69–88.

Rossi, J., and M. Yudell. 2012. “The Use of Persuasion in Public Health Communication: An Ethical Critique.” *Public Health Ethics* 5 (2): 192–205. https://doi.org/10.1093/phe/phs019.

Salvi, Francesco, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West. 2024. “On the Conversational Persuasiveness of Large Language Models: A Randomized Controlled Trial.” arXiv. https://doi.org/10.48550/arXiv.2403.14380.

Simion, Mona, and Christoph Kelp. 2020. “Trustworthy Artificial Intelligence.” *Asian Journal of Philosophy* 2 (1): 1–12.

Song, Fei. 2023. “Network of AI and Trustworthy: Response to Simion and Kelp’s Account of Trustworthy AI.” *Asian Journal of Philosophy* 2 (2): 1–8.

Sperber, Dan, Fabrice Clément, Christophe Heintz, Olivier Mascaro, Hugo Mercier, Gloria Origgi, and Deirdre Wilson. 2010. “Epistemic Vigilance.” *Mind and Language* 25 (4): 359–93.

Susser, Daniel. 2020. “Ethical Considerations for Digitally Targeted Public Health Interventions.” *American Journal of Public Health* 110 (S3).

Tsai, George. 2014. “Rational Persuasion as Paternalism.” *Philosophy and Public Affairs* 42 (1): 78–112.

Williams, Daniel. 2023. “The Marketplace of Rationalizations.” *Economics and Philosophy* 39 (1): 99–123.

Winsberg, Eric. n.d. ?“?Falsehoods Fly: Why Misinformation Spreads and How to Stop It? By Paul Thagard. Columbia University Press.” *Journal of Value Inquiry*, 1–15. https://doi.org/10.1007/s10790-024-09996-3.

1. Thanks to an audience at the University of Glasgow for helpful questions on an earlier version of this paper. Thanks also to Daniel Williams for helpful discussion of this and related issues. [↑](#footnote-ref-1)