Routledge
Taylor & Francis Group

Social Epistemology

A Journal of Knowledge, Culture and Policy

ISSN: 0269-1728 (Print) 1464-5297 (Online) Journal homepage: www.tandfonline.com/journals/tsep20

AI and Epistemic Agency: How AI Influences Belief Revision and Its Normative Implications

Mark Coeckelbergh

To cite this article: Mark Coeckelbergh (18 Mar 2025): AI and Epistemic Agency: How AI Influences Belief Revision and Its Normative Implications, Social Epistemology, DOI: 10.1080/02691728.2025.2466164

To link to this article: https://doi.org/10.1080/02691728.2025.2466164

Published online: 18 Mar 2025.

SOCIAL EPISTEMOLOGY
https://doi.org/10.1080/02691728.2025.2466164

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

AI and Epistemic Agency: How AI Influences Belief Revision and Its Normative Implications

Mark Coeckelbergh

Department of Philosophy, University of Vienna, Vienna, Austria

ABSTRACT
In the ethics of artificial intelligence literature, there is increasing attention to knowledge-related issues such as explainability, bias, and epistemic bubbles. This paper investigates epistemic problems raised by AI and their normative implications through the lens of the concept of epistemic agency. How is epistemic agency impacted by AI? The paper argues that the use of artificial intelligence and data science, while offering more information, risks to influence the formation and revision of our beliefs in ways that diminish our epistemic agency. Using examples of someone who struggles to revise her beliefs, the paper discusses several intended and non-intended influences. It analyses these problems by engaging with the literature on epistemic agency and on the political epistemology of digital technologies, discussing the ethical and political consequences, and indicates some directions for technology and education policy.

ARTICLE HISTORY
Received 17 January 2025
Accepted 2 February 2025

KEYWORDS
Epistemic agency; social epistemology; belief revision; artificial intelligence

Introduction

In the ethics of artificial intelligence (AI) literature there is increasing attention to knowledge-related issues such as explainability (Wachter, Mittelstadt, and Floridi 2017), bias (Ntoutsi et al. 2020), and epistemic bubbles (Nguyen 2020). In this paper, I propose to look at epistemic problems raised by AI through the lens of the concept of epistemic agency. Epistemic agency 'concerns the control that agents may exercise over their beliefs' (Schlosser 2019) and relates to the question of how our beliefs are formed and revised. To what extent do we have control over the formation and revision of our beliefs? Can we make epistemic changes our lives, and are we able to take responsibility for it (Gunn and Lynch 2021, 390)?

There is a long-standing discussion about whether the formation of beliefs can plausibly be regarded as voluntary in any sense (see, for example, Heil 1983), and what it means to say that we are responsible for our beliefs and ought to reflect on them. Epistemic agency has been conceptualized in various ways (for a critical discussion see Kornblith 2016; Setiya 2013; Sosa 2013), and there is a discussion of epistemic control and epistemic responsibility (Steup 2018). Do I have the same kind of control over my belief formation as I have over my actions (Schlosser 2019)? Do we have voluntary control over our beliefs, as doxastic voluntarism (Vitz 2008) claims?

The question regarding epistemic agency is important also for normative reasons, since it is doubtful if we can be responsible for our beliefs if we did not voluntarily form them and if their formation was influenced in ways beyond our control. Moreover, our beliefs, in turn, shape our actions towards others. If this is so, then epistemic agency is a necessary condition for moral

CONTACT Mark Coeckelbergh mark.coeckelbergh@univie.ac.at Department of Philosophy, University of Vienna, Universitaetsstrasse 7(NIG), Vienna 1010, Austria

responsibility. In any case, whether we as a matter of fact have *full* voluntary control over our beliefs or not (see again the debate about doxastic voluntarism, for an overview see Vitz 2008) and regardless of the implications for responsibility, it is clear that as human beings we have the capability to detach ourselves from our beliefs to some extent and to reflect on them, and that most of us tend to value reflective knowledge (Sosa 1991). Most importantly, we like to take an active role in belief acquisition: we *wish* to be epistemic agents.

Furthermore, neither epistemic agency nor, in general, is a merely individual matter; my beliefs are part of a knowledge community; my social and political environment influences my beliefs. This is the terrain of social epistemology. At least since the work of Goldman and Fuller (for an overview of the literature see Goldman and Whitcomb 2011; however, there are earlier beginnings, for example in the sociology of science and knowledge), social epistemology has taken into account this social dimension and studies how social-epistemic practices shape what and how agents know and how collective knowledge is made. I will engage with recent literature below.

Assuming that we wish to be epistemic agents and that epistemic agency is influenced by the social, in this paper, I am interested in analysing and understanding cases in which an agent is, in principle, free to form her beliefs, but in which this formation is rendered more difficult because of *technologically* shaped conditions. In other words: I'm interested in cases in which epistemic agency is compromised and diminished due to the use of AI.

To ask this question regarding technology is important as discussions in epistemology usually ignore the fact that most of our epistemic lives take place in environments that are very much mediated by digital technologies, including AI and data science. Our acquisitions and constructions of knowledge is not only a social matter but also often takes place via search engines, in social media environments, via the use of electronic devices, and so on. It is therefore advisable to analyse the epistemic consequences of the use of digital technologies for the epistemic work individuals do within social-epistemological environments such as digital social media. Epistemology can learn here from philosophy of technology, media studies, and related fields.

An interesting exception in recent work in epistemology that addresses these issues and cites some works in philosophy of technology is Freiman's (2023) paper: in order to analyse how conversational AIs may influence people's beliefs, Freiman proposes the conception of 'technology-based belief', which is meant to acknowledge that non-human agency can function as the originator of our beliefs and their content (without assuming that the technology has human-like agency). He argues that technology-based belief is different from testimony-based beliefs, which have a human as the source of knowledge, but also different from instrument-based beliefs, since in the case of conversational AI, the content is delivered in natural language. This is an example of how conceptualizing the influence of technology – here AI – can be integrated with (social) epistemology.

Moreover, in contemporary philosophy of mind it is increasingly recognized that the ways in which we conceptualized minds and knowledge need to be sensitive to the ways in which new technologies influence our everyday lives and our thinking about human beings and their minds. Clowes, Gärtner, and Hipólito (2021) have argued that we now face a mind-technology problem, as a successor of the mind-body problem: new conceptualizations about the nature of mind and its relationship to artefacts have given rise to 'a new constellation of basic philosophical problems about the very nature of mind'. (2) For instance, today computers and AI challenge us to reframe and rethink what human minds are. If AI can carry out tasks that previously were seen as belonging to human thought, what does that mean for our understanding of ourselves? Do minds work like computers or not, and why? For example, are minds computational? If not, is it nevertheless a productive metaphor or not? How, exactly, should we conceptualize the relationship between humans and their artificial creations? Are we becoming cyborgs? How are our minds changing as a result of our interactions with the technologies? And what kind of knowledge and self-knowledge does AI create? Does AI distort our epistemic environments? Is knowledge in crisis?

In this paper, I focus on AI in the form of machine learning. I am interested in exploring what AI and the forms of knowledge it creates do to the formation and revision of our beliefs, and explore its normative consequences. My thesis in this paper is that the use of artificial intelligence data science may shape the formation and revision of our beliefs in intended or unintended ways and thereby make the exercise of epistemic agency, and, in particular, control over the formation and *revision* of beliefs, harder. I will discuss this question by responding to the relevant literature on epistemic agency and on the ethics and politics of AI. This will involve use of examples and belief revision that are instructive for how AI might influence belief formation and belief revision in problematic ways. I will unpack this claim and structure the discussion by distinguishing between the intended manipulation of beliefs and non-intended influences such as epistemic bubbles and the defaulting of statistical information. I will show how this is also relevant to issues concerning freedom, responsibility, and personal identity. I will then explore some normative implications of this analysis – including ethical and political aspects – and suggest a direction for technology and education policy to address the problems raised.

My focus here is on beliefs, but note that the use of AI might also create problems for taking responsibility for our actions – a possibility that also relates to the concerns of epistemology. For example, if we are not able to explain how an AI system (in particular, a deep learning system) formulates its outputs, we suffer from a particular type of ignorance: we don't know why we are doing what we are doing. This is not only a problem for accepting responsibility for our actions, but also makes it more difficult, if not impossible, to act responsibly towards others (Coeckelbergh 2020).

There is also the question of whether in some cases and situations AI itself can be considered an epistemic agent. This question relates to the issue concerning differences in knowledge and reasoning between humans and machines, and more generally to the question of whether AI has epistemic agency status, and, if so, what status. To ascribe epistemic agency to AI systems is certainly not obvious. For example, if epistemic agency requires human-like understanding, then if AI lacks this kind of understanding, as Wang (2021) has argued, it has no chance of acquiring epistemic agency status. Swanepoel (2021) has even cast doubt on the assumption that AI has agency status as such, since according to her agency is linked to intentional action. Nevertheless, as Freiman (2023) has argued, non-human technology can still be the source of knowledge, without having human-like agency or human-like understanding. However, I will *not* further discuss these questions regarding agency and type of beliefs here in order to limit the scope and length of my paper; my focus concerns AI's impact on the epistemic *agency* of *humans*.

AI and the Revision of Beliefs: Three Types of Influences

Can and does AI shape the formation and revision of our beliefs, and if so, why, how does it do so, and what are the normative consequences?

An interesting example that may be useful to discuss this question is offered by Bondy (2015), albeit not in the context of discussions about AI. It concerns a person, Claire, who was raised in a racist environment and came to believe that people with her white skin colour are superior to others, but later learns that there is no scientific evidence for this belief. Will she be able to exercise epistemic agency and responsibility, and either no longer believe this, or at least not use it in future deliberations about how to treat people? Bondy's example does not consider the use of technology and the technological environments in which we exercise epistemic agency. If we change the example and add AI to the picture, the new question is: will Claire be able to exercise epistemic agency under conditions in which AI is used? More specifically, how could AI render it more difficult for Claire to change her beliefs?

I propose to distinguish between at least three ways in which the use of AI – in particular the use of AI in digital social media environments – may render the formation and revision of our beliefs more difficult:

(1) direct manipulation of beliefs
(2) epistemic bubbles
(3) defaulting of statistical knowledge

Direct Manipulation of Beliefs

In the case of direct manipulation of beliefs, some people intend to influence the beliefs of others. For example, in Claire's case, this could be a political manipulator who defends racist beliefs and wants to maintain the flourishing of these beliefs in her society. It is in his or her interest that Claire has difficulties to change her beliefs and that those who have not yet formed, their beliefs form racist beliefs. For that purpose, the manipulator uses AI in social media environments to profile and target people. In Claire's case, that could mean for example that she is targeted with racist propaganda. This epistemic environment renders it harder for Claire to change her beliefs. In addition, a manipulator might use deception and create a situation in which have difficulties distinguishing between truth and falsehood. As MacKenzie, Rose, and Bhatt (2020, 695) suggest, the design of digital platforms, which use of AI in the form of deep learning algorithms, can make these problems worse. Formation and revision of beliefs become harder if we do not know what is true or not. This uncertainty diminishes our epistemic agency and in this case, it is done on purpose.

However, next to manipulation of beliefs through propaganda and advertising, fake news (Croce and Piazza 2021; Rini 2017), or deceit (MacKenzie, Rose, and Bhatt 2021), there are also non-intended ways in which AI may influence epistemic agency, which involve a more complex interaction between technologies and the social-epistemic worlds they create.

Epistemic Bubbles

Consider the phenomena Nguyen (2020) has identifies with the terms 'echo chambers' and 'epistemic bubbles':

> An epistemic bubble is a social epistemic structure in which other relevant voices have been left out, perhaps accidentally. An echo chamber is a social epistemic structure from which other relevant voices have been actively excluded and discredited. Members of epistemic bubbles lack exposure to relevant information and arguments. Members of echo chambers, on the other hand, have been brought to systematically distrust all outside sources. In epistemic bubbles, other voices are not heard; in echo chambers, other voices are actively undermined. (abstract)

If we only hear one voice and if that happens to be in line with our current belief, it will be harder to change that belief. Moreover, if we distrust outside sources, this will further undermine the exercise of our epistemic agency. Consider Claire again: if she hears only white supremacist echoes and voices, and comes to distrust sources that come outside her white supremacist social media environment, then this creates problems for the exercise of her epistemic agency with regard to her beliefs. An earlier concept was 'filter bubbles' (Pariser 2011): through personalization through digital technologies, we find ourselves surrounded by views we agree with; we are not sufficiently exposed to opposing views. The Internet we encounter is personalized and influences what we think, and this has consequences for epistemic agency. Such bubbles do not only reduce opposing views – which is by itself already problematic, for example if one believes that in a democracy citizens should be exposed to, and discuss, opposing views – but also undermine the exercise of epistemic agency. If one feels very comfortable in one's bubble, one feels less need to reflect on, or revise, one's beliefs. Other beliefs are distrusted or are not even encountered. In so far as AI contributes to the creation of such (social-)epistemic structures, for example, by powering a search engine such as Google's that tends to create such epistemic filters or by enabling social media platforms like Facebook in which people like Claire can nest themselves in white supremacist bubbles, AI presents a risk for epistemic agency.

The Defaulting of Statistical Information

Another issue has to do with the nature of the knowledge presented through AI. There is a risk that AI defaults statistical knowledge, at the expense of other kinds of knowledge, for example, causal knowledge. Consider the example again. In principle, Claire can consult scientific information and learn that there is no scientific basis for her white supremacism. However, if AI offers her statistical knowledge of a correlation between skin colour and success in her particular society, then she might be tempted to keep her white supremacist belief rather than change it, even if there is no causal link. While this kind of statistical knowledge might also have been available before AI was used, the widespread use of AI in society – as transformed by digital technologies such as current digital social media – is likely to make this kind of knowledge more readily available. It becomes the default.

Again, this type of influence is usually not intended; it is not necessarily a case of manipulation. In the case of epistemic bubbles and the defaulting of statistical knowledge it might even be that the use of AI for belief formation is user-driven: users may intentionally wish to seek to influence the formation of their beliefs, but then bump against the limits presented by these phenomena – often without being aware of these limits. For example, Claire may actively search for other beliefs that counter her white supremacism, but at least initially or without help from others experience difficulties to leave her epistemic bubble or get different kinds of information. The algorithms and the epistemic-technological environments they enable or support hinder this.

Further Discussion: Freedom, Responsibility, and Personal Identity

These claims about diminished epistemic agency due to the use of AI and other digital technologies do not deny people's basic capacity for epistemic agency and *freedom* as human beings. They remain free in a metaphysical or ordinary psychological sense. For example, we could say that Claire has free will with regard to her beliefs. She can, in principle, choose her beliefs. Nothing or nobody takes away her capacity for freedom in this thought experiment, not even AI or a racist manipulator. She can potentially change get rid of her racist belief or not use it anymore in her deliberations. However, the problem is that she cannot actualize her potential for epistemic agency under these conditions, because it becomes more *difficult* to *exercise* her epistemic freedom and agency with regard to that belief and in general. This is so partly because she has not learned to do so in her environment in which that belief was widespread and easily available, and partly because this belief revision is rendered difficult because of the dominant presence of statistical knowledge offered by content producers using AI and data science (which is then remediated via digital social media), the unintended influence of epistemic bubbles and echo chambers, or even the intended and direct manipulation of her beliefs by means of AI. While Claire remains, in principle, entirely free to change her beliefs, the issue is that AI shapes belief formation and revision in such a way that it becomes harder for her to change her beliefs. Her epistemic agency is not (sufficiently) manifested; it is hindered and compromised because of the social-technological epistemic environments she finds herself in.

Interestingly, this problem at least partly resembles those raised by the case of nudging, which is about influencing one's choice of architecture without taking away one's freedom. A nudge organizes one's choice and action in such a way that agents are gently pushed in a particular direction. According to Sunstein and Thaler (2008), this has the advantage of preserving freedom while helping humans make better decisions. For example, the design of a supermarket might be changed in such a way that organic products are presented in prominent places. The choice architecture is changed and people's choices are subtly influenced, without giving people explicit directives or prohibiting certain courses of action.

Nudging theory is about choice and action, but we could also apply a similar kind of argument to the epistemic problem at hand. The formation of a non-racist belief is not directly hindered or forbidden by AI or by a racist manipulator who uses AI, and in this sense, Claire keeps her basic epistemic freedom: the freedom to believe what she wants. But because of the technology and its use in a particular social environment, the formation of her new belief (compare: her choice) is hindered and influenced by changing the epistemic architecture. AI presents statistical information as default and status quo and/or contributes to an epistemic bubble in which the white supremacy belief is also defaulted. It is, in principle, possible to choose the other belief, to change one's belief. Claire *could* reflect, do research about evidence for her belief, and embrace non-racist beliefs. She is not prevented from doing so. In the example, she even wishes to do so and perhaps tries to do so. But through her background and education, and in addition through the epistemic architecture co-constructed by AI (her epistemic bubble, the manipulation of her beliefs via social media, and so on), the white supremacist belief presents itself as the default and the norm. To get rid of that belief is possible but requires more effort. In the economy of belief formation, keeping the belief is the cheaper option. It requires less or no work. Revising beliefs involves higher costs.

Perhaps this socio-technological creation of a kind of 'epistemic laziness' already happened with the use of the Internet in general (and continues to happen). Lynch (2016) has argued that the net has given us the impression that getting information is enough. We become lazy, passive knowers who prefer to use search engines and fail to acquire other types of knowledge, in particular knowledge based on reasoning and knowledge about how information connects together. Lynch analyses what he calls 'Google-knowing'. It is easy and fast. And we trust that kind of knowledge and become dependent on it. Googling is believing. Similarly, one could say: using AI is believing. We are discouraged to do more epistemic work and do not exercise our epistemic agency. AI and data science give us the statistics; why would we look further and obtain causal knowledge? We have the knowledge offered by recommender algorithms, by personalized Google search, by the Facebook algorithm. And we feel good in the epistemic bubbles of the social media we use, in which our beliefs are confirmed on a daily basis. Or worse: other people, companies, and governments using these technologies and the methods of this science might work behind our backs to manipulate our choices and shape our beliefs. The formation and revision of beliefs are taken care of by others, and with no more reasoning needed, our epistemic agency is thus diminished or at least severely at risk.

Recently, Gunn and Lynch (2021) have emphasized that our *responsibility* to exercise such agency is also diminished. For example, responsible agency requires us to update beliefs in the light of new evidence, and also respect and appreciate others, avoiding what the authors call 'epistemic arrogance' (392). The Internet, so the authors argue, makes this more difficult, for example, by facilitating in-group/out-group sorting. While the Internet may also expand epistemic agency in some ways, for example by making information more widely available, the responsible exercise of epistemic agency becomes more difficult through information personalization (see again the Google example) and because it becomes more difficult to evaluate testimonial knowledge and to judge trustworthiness. And sometimes exercising epistemic agency is also difficult simply because there is *too much* information. AI and data science might increase these problems. By categorizing people, AI may encourage drawing in-group/out-group lines, it may produce knowledge that cannot be easily evaluated by humans as trustworthy or not (because humans cannot do the analysis done by deep learning algorithms, for example), and it may offer so much information that we become epistemically confused.

How hard it is to change a belief also depends on how personal these beliefs are. There is a link between the way AI affects epistemic agency and *personal identity*, including narrative identity (for general work on narrative identity (see, for example, Atkins (2004) and the work of Ricoeur and Schechtman). When Claire tries to change her beliefs and is hindered in this by knowledge and manipulation through AI, she is not only trying to do something to her beliefs – as if those beliefs were entirely external to her personal identity. Her identity, in the form of a story about herself, is linked to her beliefs, including her political beliefs. If we assume that her political beliefs also constitute her 'internal' narrative identity (for more discussion of this claim see Ulatowski and Lumsden 2023), including the normative aspects of that narrative identity, it is plausible that Claire does not only want to drop her belief in white supremacy; she also wants to be *the kind of person* that does not hold these beliefs and she wants to be able a story about herself in which that belief does not play a role. She may want to replace it with another belief, for example, a belief in basic human equality, or the belief that people with a black skin colour in the U.S. deserve a positive discriminatory treatment given the racist dimension of history and current practices. *That* is the kind of person Claire wants to be. The question of epistemic agency is thus linked to that of narrative identity and, as we will see later, virtue. One could then ask questions such as: Are we still the same person when we manage to shake off an earlier belief? Can the exercise of epistemic agency – here in an AI context – really change who we are as a person, and what is the precise link between epistemic agency and personal identity? And how does AI influence this change in personal identity – through its effect on epistemic agency and otherwise? For example, is Claire a different person when she no longer believes in a link between skin colour and superiority, and what else would be needed to change her (completely?) as a person?

Note that this focus on 'shaking off' beliefs does not mean that we can always do this (there might be other reasons why we cannot do this, reasons that have nothing to do with technology as such), nor that we should always *want* to take distance from a particular belief or that we can and want to take distance from *all* beliefs. There is no need to go the full sceptical or extreme voluntarist route, and there is a separate question about the architecture of beliefs in the light of personal identity. It might be that there are some hinge beliefs that one may want to embrace since one has come to understand them as part of oneself, as Modesto interprets Wittgenstein's anti-scepticism (Modesto 2013; Wittgenstein 1974), whereas other beliefs can – and, in some cases, probably should – be shaken off more easily. In the former case, maintaining a particular belief can also be seen as an exercise of epistemic agency. Beliefs also relate to one another; we may assume that there is a kind of ecology of beliefs, in which some beliefs depend on other beliefs. And there is the question about the (degree of) integration of beliefs. Claire could try to get rid of her white supremacist and related beliefs but keep others she feels define the person she is and wants to be, and these other, non-supremacist beliefs are well integrated with one another.

Thus, epistemic agency need not imply extreme scepticism or voluntarism, and the exact way in which beliefs hang together is a separate question. The arguments about the potential influence of AI also work with a concept of epistemic agency that is 'medium sceptic', so to speak: one that is about the agent questioning her beliefs but not necessarily all beliefs and with no requirement to completely transform her identity. Nobody asks Claire if she always questions her beliefs or shakes off all the beliefs she acquired during her youth. In fact, it might well be that, at least seen from a first-person point of view, nobody asks anything from her: it might be that she has the experience that she herself decides this personal-epistemic change. She has a basic capacity for epistemic agency and wants to use it to change what she believes and the person she is. She has come to the conclusion that she no longer wants to identify with white supremacist beliefs, and then tries to change that – and encounters difficulties within a particular techno-social environment in which digital technologies such as AI and power structures (see below) play a role. At the same time, because of this socio-technological influence and in spite of Claire's voluntarist thinking, whether she actually manages to shake off her beliefs or manages to not use it anymore in relevant deliberations does not totally depend on her will to change, but is also influenced on her social and technological environment. The concept of epistemic agency constructed here is in that sense also 'medium voluntarist': what we believe and the extent to which we can change our beliefs depends to some extent on our socio-economic environment and other factors, not just on what we want. The discussion about epistemic agency and voluntarism in the case of AI-mediated environments is thus connected to the more general discussion about epistemic agency and external influences.

Moreover, the examples constructed here assume that Claire wishes to get rid of her belief, but there may be reasons why one does not want to change that particular belief here and now. There could also be reasons for not exercising one's epistemic agency in particular circumstances. Nothing said here requires an obsession with epistemic agency; it may well be that in some circumstances it is recommended not to even question a particular belief. For example, if I feel very much in love with my partner and show that to her at a particular time, then in that moment it seems not to be a good moment to exercise my epistemic agency and question my belief that I love my partner, even if there might be other moments, circumstances, or relationships in which such an exercise might be entirely appropriate or even highly recommended. The point I am making is about the socio-technological barriers that are in place *if and when* one wishes to change one's belief and wishes to exercise one's epistemic agency in this way (a wish that in turn may be influenced by for example being exposed to contradictory information coming from the internet).

This analysis shows that epistemic agency as related to personal identity clearly has a practical and normative dimension; in particular, it also links to virtue ethics. More generally, we need to discuss the normative implications of the analysis presented here.

Normative Implications: Moral and Political Problems

So far, this discussion does not directly touch upon the link between epistemic agency and *morality*. It does not appeal to morality to make claims about the influence of AI on epistemic agency. But exploring this route offers an interesting way to further develop the argument. If it is true that not only evidence but also 'morality has a role to play in doxastic deliberation', as Dandelet (2021) argues, then what goes wrong (or could go wrong) in Claire's case is not only a failure to, in the formation of her beliefs, relying on evidence about there not being a causal relation between skin colour and superiority, but also a failure to let that formation of beliefs be untouched by moral demands, which harm others. This could also be seen as a failure of epistemic agency in a broader sense, with the term now understood as not only involving the formation of beliefs but also as involving moral knowledge such as moral demands. We could thus add to the thought experiment (and hence to the account I am developing here) the idea that it will be easier for Claire to get rid of her false belief (white supremacy) if she were to let the formation of her beliefs be influenced by the moral belief that racism is wrong. This use of morality might make it easier for her to counter the influence of AI on her beliefs and her epistemic agency. More generally, moral knowledge may help to develop, sustain, and maintain epistemic agency.

If this is the case, then it is recommendable to add this moral dimension to education (see also the next section). But helping people to exercise their epistemic agency is not only an individual matter; we also need to look at the political side of epistemic agency problems. The socio-epistemic and techno-epistemic issues discussed here are related to wider, problematic socio-political structures. For example, we may consider what Dotson (2014) calls 'epistemic oppression': 'a persistent and unwarranted infringement on the ability to utilize persuasively shared epistemic resources that hinder one's contribution to knowledge production' (116). However, as suggested throughout this analysis, there is also a problem with knowledge consumption. Extending this concept of structural epistemic oppression to moral knowledge and to the topic under discussion, one could argue that it is important to increase access to scientific knowledge and evidence and remove socio-technological barriers to using it to revise one's beliefs, but also to avoid epistemic oppression by making available forms of moral knowledge. In the example, this could mean: giving people access to moral and political ideas that counter racism.

In the initial thought experiment, Claire was epistemically oppressed in at least two senses: she did not have sufficient access to scientific knowledge in her environment (in particular, knowledge of a lack of evidence for a causal link between skin colour and superiority) and she found herself in socio-technological environments that discouraged changing her belief and exercising her epistemic agency. However, she was also not sufficiently exposed to moral and political theory – even in rudimentary or popular form – that could have countered her belief and helped her to change it. This kind of oppression can also occur in social media environments, in which one becomes locked in an epistemic bubble and is mainly exposed to one cluster of moral and political beliefs, to the exclusion of others. AI and data science can be used on purpose to achieve such exclusion and oppression of others, although often the effect is unintended.

In the example of Claire and AI, epistemic oppression is continued through (1) the dominant presence of statistical knowledge about a correlation between skin colour and success (for example, in social media) – to the exclusion of other kinds of knowledge, in particular (lack of) causal evidence for a causal link between skin colour, which shaped the early formation of her belief and made epistemic re-formation difficult – and (2) insufficient exposure to moral and political beliefs that would make it easier for Claire to get rid of her white superiority belief (for example, because she would be living in a social media bubble in which those superiority beliefs were not present).

A related way to frame and elaborate the normative importance of increasing epistemic agency in a more political way is to use the term 'epistemic justice' (see, for example, Fricker 2007; Pohlhaus 2020) in a way that goes beyond the epistemology of testimony and applies more broadly to the silencing of knowledge, which leads to exclusion: if and in so far AI leads to a systematic exclusion in the sense of silencing of specific kinds of knowledge (including knowledge about morality and politics) and harms those who suffer from this exclusion, then socio-technological barriers to epistemic agency also constitute an injustice. One could argue that the mentioned potential influences of AI constitute an injustice because of at least two reasons that have to do with knowledge: because they lead to a lack of diversity of knowledge, which is a form of exclusion and therefore problematic in itself, and because they lead to problems of exercising epistemic agency with regard to one's beliefs, which could be seen as a politically relevant form of harm, wrongdoing, and unjust disempowerment.

The exclusion of types of knowledge and the diminishing of epistemic agency are linked to power and the exclusion of people in various and complex ways. As Lynch notes, technologies such as the Internet and big data change not only how we know but also who gets to know and who decides what counts as knowledge. There is thus a relationship between epistemic agency and power in that sense. The wide availability of information has not always and certainly does not necessarily lead to the democratization of knowledge (especially knowledge based on evidence and reasoning), let alone to the spread of wisdom. Furthermore, epistemic agency might itself be a question of power. Do some people have the privilege of being well-trained knowers, whereas others are often mainly the passive objects of knowing, epistemic *patients* rather than epistemic agents? It seems that there is no epistemic equality, globally or locally. With regard to epistemic agency, one could say that some well-educated people and some powerful actors in the tech world have more power to shape their own beliefs *and* those of others. In that sense, some people have more epistemic agency than others, which can be analyzed with political-philophical concepts such as oppression, justice, and power.

Thus, the point is not just that truth is in danger, which creates problems for democracy and public debate (Snow and Vaccarezza 2021), but that some people and organizations have more power to define what is accepted as true within a particular context and to shape their own beliefs and those of others. This is a problem of epistemic inequality and epistemic justice. In the case of Claire, some people were (are?) in a position of power in her society to define what is true about social relations, for example, people from a white supremacist community who were allowed/had the power to promulgate their beliefs and who benefited from the low degree of epistemic agency Claire previously had, and used social media to try to maintain that situation. Moreover, Claire has less power to shape her beliefs and those of others than other people, including those who benefit from her lack of epistemic agency.

Furthermore, from the angle of social-epistemological theory that focuses on race, Claire's struggle specifically can be interpreted as an attempt to overcome what Mills (2007), in his contribution to racial epistemology, has called 'white ignorance': Claire's delusion of racial superiority and problems regarding epistemic agency related to trying to overcome this are then not seen as contingent but put in the social context of racial domination. There are also other accounts that relate epistemology to the politics of identity. For example, Adam (2000) has argued that AI 'follows classical versions of epistemology in assuming that the identity of the knowing subject is not important' (abstract) and has proposed a feminist reading which shows that the representation of knowledge in AI privileges some (male developers) and denies a voice to less powerful groups such as women. In our example, one could use this to argue that Claire's efforts to exercise her epistemic agency are rendered difficult because, among other reasons, she is influenced by a male-dominated epistemology of AI that is not only racist but also systematically privileges men when it comes to supporting the development of agency and reflection.

Finally, the normative implications of the analysis presented above can be elaborated by using the concept of virtue. The mentioned change in narrative identity clearly has a normative dimension and a virtue dimension in the sense that it concerns what kind of person one should be. In particular, Claire's evolution from a person who grew up in an environment where white superiority was the norm to a person who questions this and tries to get rid of this belief by exercising her epistemic agency can also be framed as a voyage towards becoming a more virtuous person for two reasons: (1) exercising one's epistemic agency can itself be seen as a virtue, an epistemic virtue (to reflect on one's beliefs makes you a better person) and (2) the result can be framed in terms of virtue: changing one's belief can make one into a better person. If Claire manages to shed her racist belief by exercising her epistemic agency, she will become more epistemically virtuous and also become a better, more virtuous person.

However, sometimes the link between epistemological agency and virtue is not so clear. For example, Rini (2017) has argued, somewhat controversially, that some forms of online partisanship are not *necessarily* epistemically vicious: the author thinks that from an individual's point of view, it sometimes makes sense to assign more credibility to a testifier when she has the same political affiliation and makes politically relevant claims (E-50). To the extent that AI encourages partisan epistemological behaviour, such discussions are relevant to the questions regarding epistemic agency: is partisan trust necessarily a good way of exercising one's epistemic agency, if it is one at all, and why and under which circumstances? (More work could be done on this virtue theme by further linking it to the tradition of virtue epistemology. Here, Sosa (1991) is again relevant. Moreover, the issue of partisan trust also links to discussions about division of labour and reliance on expertise. However, here I will not further develop these themes, which deserve their own treatment.)

Conclusion and Brief Policy Recommendations

To conclude, the use of AI in the form of machine learning, especially in digital social media contexts, creates risks for the exercise of epistemic agency. In this paper, I have focused on belief revision and the ways that may be hindered by AI: through direct manipulation, but also via epistemic bubbles and the defaulting of statistical information. I have also shown how this is related to issues concerning freedom, responsibility, and personal identity, and how it has normative consequences – both moral and political. The influence of AI on belief formation and revision is not only about fake news and deceit, but also about the creation of other epistemic processes and architectures that render the exercise of epistemic agency more difficult. Moreover, the problem regarding epistemic agency is not AI in itself (if this makes sense at all), but the interplay of AI with specific epistemic-technological environments (I focused on digital social media) and the users. 'Users' can mean those who are influenced through AI but also those who influence by using AI on purpose. How AI shapes belief formation and revision also depends on the identities of the users and on their political context, for example regarding existing forms of epistemic and political oppression that promote the interests of some at the expense of others.

Given these problems and their normative implications, what are the policy implications of the analysis offered above? In order to address the problem of AI shaping (particularly hindering) the exercise of epistemic agency, policymakers could intervene in at least three domains:

The first is the education of the general public in ways that ensure capabilities to get access to, and use, diverse types of knowledge, including scientific knowledge, and strengthen their epistemic agency in relation to knowledge offered by AI and manipulation by AI. This includes the right use and interpretation of this knowledge and belief formation based on a diversity of sources including scientific evidence. In existing literature on education policy with regard to digital transformation, the emphasis is often on the acquisition of digital skills, but not on belief formation and epistemic agency under technological conditions. An exception is Croce and Piazza (2021), who argue for enhancing epistemic agency through educational interventions in order to address the problem of fake news. Users should be educated about the benefits of a more varied information diet, although this may not work for users trapped in echo chambers. A similar argument could be made with regard to belief revision (more generally); addressing the echo chambers difficulty is then also key. Moreover, one could explore the broader question of how to support people to develop a narrative practical identity under these techno-epistemic conditions.

A second intervention is regulation of technology use and development, which aims at preventing the widespread use of AI for the manipulation of beliefs and for diminishing epistemic agency (or at least for taking advantage of an already low level of epistemic agency). Here, the focus is not so much on educating people but on regulating powerful actors in the world of technology development and use, including large corporations and powerful states. The political justification for this could be that the development and exercise of epistemic agency is an individual good that needs to be protected and fostered and that contributes to the epistemic health of society.

Third, given the link between the exercise of epistemic agency and systemic forms of oppression and injustice indicated above, I suggest that educational and technological changes can only be successful with regard to increasing epistemic agency if they are linked to broader efforts to address power issues and structural issues in society. This includes addressing socio-economic issues and responding to identity-based political claims. Ultimately, this requires us to question our descriptive and normative beliefs about the order of society. Social epistemology needs to include a political epistemology. That is, our analysis of the way knowledge is organized in society needs to include an analysis of political interests and powers vested in propagating some beliefs and narratives rather than others, and in maintaining epistemic-technological environments that support these interests and powers.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

Open access funding provided by University of Vienna. This research was funded in part by the Knowledge in Crisis project, supported by the Austrian Science Fund (FWF) under the Clusters of Excellence programme (https://doi.org/10.55776/COE3).

Notes on contributor

Mark Coeckelbergh is a full Professor of Philosophy of Media and Technology at the Philosophy of Department of the University of Vienna. He is also ERA Chair at the Institute of Philosophy of the Czech Academy of Sciences in Prague and Guest Professor at WASP-HS and University of Uppsala. Previously he was the President of the Society for Philosophy and Technology (SPT). His expertise focuses on ethics and technology, in particular robotics and artificial intelligence. He is a member of various entities that support policy building in the area of robotics and artificial intelligence, such as the European Commission's High Level Expert Group on Artificial Intelligence, the Expert Council Ethics of AI of the Austrian UNESCO Commission, the Austrian Council on Robotics and Artificial Intelligence, and the Austrian Advisory Council on Automated Mobility. He is University of Vienna's Circle U. Academic Chair for Artificial Intelligence. He is the author of 17 philosophy books and numerous articles and is involved in several national and European research projects on AI and robotics.

ORCID

Mark Coeckelbergh http://orcid.org/0000-0001-9576-1002

References

Adam, A. 2000. “Deleting the Subject: A Feminist Reading of Epistemology in Artificial Intelligence.” *Minds and Machines* 10 (2): 231–253. https://doi.org/10.1023/A:1008306015799.

Atkins, K. 2004. “Narrative Identity, Practical Identity and Ethical Subjectivity.” *Continental Philosophy Review* 37 (3): 341–366. https://doi.org/10.1007/s11007-004-5559-3.

Bondy, P. 2015. “Epistemic Deontologism and Strong Doxastic Voluntarism: A Defense.” *Dialogue - Canadian Philosophical Association* 54 (4): 747–768. https://doi.org/10.1017/S0012217315000487.

Clowes, R. W., K. Gärtner, and I. Hipólito. 2021. “The Mind-Technology Problem and the Deep History of Mind Design.” In *The Mind-Technology Problem: Investigating Minds, Selves and 21st Century Artifacts*, edited by W. Robert, KlausGärtner Clowes, and Inês Hipólito, 1–45. Cham: Springer.

Coeckelbergh, M. 2020. “Artificial Intelligence, Responsibility Attribution, and a Relational Justification of Explainability.” *Science and Engineering Ethics* 26 (4): 2051–2068. https://doi.org/10.1007/s11948-019-00146-8.

Croce, M., and T. Piazza. 2021. “Consuming Fake News: Can We Do Any Better?” *Social Epistemology* 37 (2): 1–10. https://doi.org/10.1080/02691728.2021.1949643.

Dandelet, S. 2021. “Doxastic Wronging and Evidentialism.” *Australasian Journal of Philosophy* 1 (1): 1–14. https://doi.org/10.1080/00048402.2021.1982999.

Dotson, K. 2014. “Conceptualizing Epistemic Oppression.” *Social Epistemology* 28 (2): 115–138. https://doi.org/10.1080/02691728.2013.782585.

Freiman, O. 2023. “Analysis of Beliefs Acquired from a Conversational AI: Instruments-Based Beliefs, Testimony-Based Beliefs, and Technology-Based Beliefs.” *Episteme* 21 (3): 1031–1047. https://doi.org/10.1017/epi.2023.12.

Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Goldman, A., and D. Whitcomb. 2011. *Social Epistemology: Essential Readings*. Oxford: Oxford University Press.

Gunn, H., and M. P. Lynch. 2021. “The Internet and Epistemic Agency.” In *Applied Epistemology*, edited by Jennifer Lackey, 389–408. Oxford: Oxford University Press.

Heil, J. 1983. “Doxastic Agency.” *Philosophical Studies* 43 (3): 355–364. https://doi.org/10.1007/BF00372372.

Kornblith, H. 2016. “Epistemic Agency.” In *Performance Epistemology*, edited by Miguel A. F. Vargas, 167–182. Oxford: Oxford University Press.

Lynch, M. P. 2016. *The Internet of Us: Knowing More and Understanding Less in the Age of Big Data*. New York: Liveright.

MacKenzie, A., J. Rose, and I. Bhatt. 2020. “Dupery by Design: The Epistemology of Deceit in a Postdigital Era.” *Postdigital Science & Education* 3 (3): 693–699. https://doi.org/10.1007/s42438-020-00114-7.

MacKenzie, A., J. Rose, and I. Bhatt. 2021. *The Epistemology of Deceit in a Postdigital Era – Dupery by Design*. Cham: Springer.

Mills, C. 2007. “White Ignorance.” In *Race and Epistemologies of Ignorance*, edited by Shannon Sullivan and Nancy Tuana, 13–38. Albany: State University of New York Press.

Modesto, M. 2013. “Wittgenstein on Epistemic Agency.” In *Mind, Language and Action*, edited by Daniele Moyal-Sharrock, Volker Munz, and Annalisa Coliva, 160–163. Berlin: De Gruyter.

Nguyen, C. T. 2020. “Echo Chambers and Epistemic Bubbles.” *Episteme* 17 (2): 141–161. https://doi.org/10.1017/epi.2018.32.

Ntoutsi, E., P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, et al. 2020. “Bias in Data-Driven Artificial Intelligence Systems—An Introductory Survey.” *WIREs Data Mining and Knowledge Discovery* 10 (3): e1356. https://doi.org/10.1002/widm.1356.

Pariser, E. 2011. *The Filter Bubble: How the New Personalized Web is Changing What We Read and How We Think*. London: Penguin Books.

Pohlhaus, G. 2020. “Epistemic Agency Under Oppression.” *Philosophical Papers* 49 (2): 233–251. https://doi.org/10.1080/05568641.2020.1780149.

Rini, R. 2017. “Fake News and Partisan Epistemology.” *Kennedy Institute of Ethics Journal* 27 (S2): E43–E64. https://doi.org/10.1353/ken.2017.0025.

Schlosser, M. 2019. “Agency.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/archives/win2019/entries/agency/.

Setiya, K. 2013. “Epistemic Agency: Some Doubts.” *Philosophical Issues* 23 (1): 179–198. https://doi.org/10.1111/phis.12009.

Snow, N., and M. Vaccarezza. 2021. *Virtues, Democracy, and Online Media: Ethical and Epistemic Issues*. New York: Routledge.

Sosa, E. 1991. “Knowledge and Intellectual Virtue.” In *Knowledge in Perspective: Selected Essays in Epistemology*, 225–244. Cambridge: Cambridge University Press.

Sosa, E. 2013. “Epistemic Agency.” *The Journal of Philosophy* 110 (11): 585–605. https://doi.org/10.5840/jphil2013110116.

Steup, M. 2018. “Doxastic Voluntarism and Up-To-Me-Ness.” *International Journal of Philosophical Studies* 26 (4): 611–618. https://doi.org/10.1080/09672559.2018.1511148.

Sunstein, C., and R. Thaler. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven: Yale University Press.

Swanepoel, D. 2021. “Does Artificial Intelligence Have Agency?” In *The Mind-Technology Problem: Investigating Minds, Selves and 21st Century Artifacts*, edited by Robert W. Clowes, Klaus Gärtner, and Inês Hipólito, 83–104. Cham: Springer.

Ulatowski, J, and D. Lumsden. 2023. “Do Political Convictions Infect Every Fibre of Our Being?” *Social Epistemology* 38 (5): 560–576. https://doi.org/10.1080/02691728.2023.2186752.

Vitz, R. 2008. “Doxastic Voluntarism.” *The Internet Encyclopedia of Philosophy*. https://iep.utm.edu/doxastic-voluntarism/.

Wachter, S., B. M, and L. Floridi. 2017. “Transparent, Explainable, and Accountable AI for Robotics.” *Science (Robotics)* 2 (6): eaan6080. https://doi.org/10.1126/scirobotics.aan6080.

Wang, J. 2021. “Is Artificial Intelligence Capable of Understanding? An Analysis Based on Philosophical Hermeneutics.” *Cultures of Science* 4 (3): 135–146. https://doi.org/10.1177/20966083211056405.

Wittgenstein, L. 1974. *On Certainty*. Edited by G.E.M. Anscombe and G. H. von Wright, Denis Paul and G. E. M. Anscombe. Oxford: Blackwell.