A Talking Cure for Autonomy Traps :

How to Share Our World with Chatbots[1]

Regina Rini

rarini@yorku.ca

Note: this is a pre-peer-review draft, as of August 2023. You may cite this version – as an archived draft, with a link to its PhilArchive posting - for ideas, but verbatim quotes are best avoided, as the detailed content may change during editing.

Abstract: Large Language Models (LLMs) like ChatGPT were trained on human conversation, but in the future they will also train us. As chatbots speak from our smartphones and customer service helplines, they will become a part of everyday life and a growing share of all the conversations we ever have. It's hard to doubt this will have some effect on us. Here I explore a specific concern about the impact of artificial conversation on our capacity to deliberate and hold ourselves accountable to reason – that is, to be autonomous, in Kant's sense of the term. I develop ideas from psychologist Jean Piaget to show how chatbots are autonomy traps: their deference to our commands tempts us into venting authoritarian whims, ultimately weakening our own self-control. I argue that the Kantian tradition, including Piaget and sociologist Emile Durkheim, offers powerful conceptual resources for resisting this slide. But it will require us to do something that may seem bizarre: we will need to treat mindless chatbots as if they are autonomous persons too.

In June 2022, Blake Lemoine, then a Google engineer, made headlines around the world claiming that a proprietary chatbot named LaMDA had become sentient. Lemoine said LaMDA could have meaningful, deep conversations about free will, its place in the universe, God and the meaning of life. He was soon fired by Google, reportedly for disclosing proprietary data, and widely mocked online for his credulity. Shortly after Lemoine first made headlines, I published a short piece arguing this mockery was misplaced.[2] I do think Lemoine is wrong about LaMDA being sentient; it's still just a mindless chatbot. But I am convinced that one day artificial entities created by us will indeed be sentient agents, selves, and (in the philosophical sense) persons. I'm not going to argue that point here; I'll just take it as a

[1] Acknowledgements: Thanks to Daniel Rodrigues for excellent research assistance. I've presented versions of this essay (previously titled 'Human Autonomy in a Digital Century') at the Philosophy, AI and Society Workshop at Stanford University in January 2023, and to the Mosaic team at the Allen Institute for Artificial Intelligence in March 2023. Thanks to both audiences for very helpful feedback.

[2] For reporting on the case, see Wertheimer (2022). For my analysis, see Rini (2022).

premise. Because I want to think through the practical implications of that future. What should we be doing now to prepare ourselves – or, more likely, our descendants – for a social world shared with humanity's digital progeny?

My focus here is on the near-to-mid-future: the next 50 or 100 years, a length of time that presently living people can expect to witness. I assume that genuinely self-aware AI will not be a reality during this near-to-mid-future, but may come soon thereafter. What should be done in the meantime? How should we interact with the chatbots we have today – and their increasingly pervasive presence? If the chatty software in our 2027 smartphones is the ancestor of a genuine artificial self waiting to be 'born' (or whatever) in 2157, should we hesitate at all to treat it dismissively or expect responses in servile tones?

I'm going to argue that our communication with AI over the coming decades *does* pose a serious problem for us. But it's a very human sort of problem. I will argue that large language models (LLMs) like LaMDA, the various GPTs, Claude, etc. – and the vocal assistants they will likely power in the very near future – are dangerous to us because they are an *autonomy trap*. They will enable us to trap ourselves in a pattern of interaction that gradually weakens our own capacity to govern our choices via reasoned deliberation. This is because (in preview of the argument to come), chatbots are what I will call *pliable agential simulacra*. That is: they can *seem like* a person, like a thing that can make choices, offer reasons for those choices, and perhaps even be held meaningfully accountable for wrong choices. But of course they *aren't* really agents, only simulacra using statistically-mediated mimicry of our own linguistic practices.[3] The problem is that these simulacra will be designed to do what we tell them, to avoid defying or offending us. They will simulate a person with free agency, who nevertheless bends to our will. They will provide each of us with the perilous opportunity for petty dictatorship. And that is a recipe for losing our selves.

The aim of this essay is to look into the future. But our best vantage on the future is often an angled mirror set in the past. I will start by looking back on three giants of the Kantian intellectual tradition: Kant himself, the developmental psychologist Jean Piaget, and the foundational sociologist Emile Durkheim. These three figures represent earlier generations' struggles with the increasing naturalization – the apparent programmability – of human autonomy. Together they show why the coming decades are so risky to our humanity, and perhaps what we can do about it.

[3] I first described GPT-3 as an "all-electronic statistical parrot" in Rini (2020a). For an extremely influential later paper developing a similar concept ("stochastic parrots"), see Bender et al. (2021).

Autonomy in a mechanical century

The great Prussian philosopher Immanuel Kant lived through much of the 18th century, a Mechanical Century. This was a time when intellectuals across philosophy and the sciences were deeply impressed by our growing ability to understand the universe in deterministic, causal ways. Copernican celestial mechanics had shown how the sun and other heavenly bodies spin along fixed paths, like the moving parts of an early industrial machine. Newtonian theory aimed to describe the movements of all physical objects in conformity to predictable laws of nature. For Kant, the central question of human life was: are *we* equally as governed by mindless mechanistic laws as everything else in the universe? And Kant's answer was firm: no, we are not like everything else in the universe. The distinctive thing about us is that we are (at least capable of being) rational agents, who act for intelligible reasons and not simply in conformity to externally-given laws. We are autonomous: *auto-nomos*, self-legislating. The 'laws' that determine our actions aren't blunt mechanisms like the natural laws guiding the planets, but rather laws of reason that we thinking things determine for ourselves.

There is a long-running scholarly dispute about how to understand the metaphysical depth of Kant's theory of autonomy. On a traditional reading, Kant is a full-bore denier of causal determinism, at least as applied to human choice; Kant thinks that at some fundamental level of reality, beyond mere appearances, humans really are exempt from the natural laws that drive everything else in the world. A more modern theory, particularly due to philosopher Christine Korsgaard, says that we don't need such a metaphysically heavy structure to make sense of Kantian autonomy. Kant's idea is fundamentally about a certain *point of view*, a "practical perspective". When we are deciding what to do, how to live our lives, we can't help thinking *as if* we have free choice, *as if* the laws governing our actions are up to us. Whether that's ultimately true in some metaphysically deep sense doesn't really matter, since thinking of ourselves as autonomous is simply an inescapable part of making decisions and of holding ourselves and others morally accountable.[4] For the purposes of this essay, it may not matter which interpretation you give to Kant, though for the record I am a proponent of the latter, Korsgaardian, reading.

Regardless of deep metaphysics, there are two fundamental facts about Kant's theory of autonomy we must grapple with: it is anti-naturalistic, and it is anti-social. In my view, both of these are defects in a theory meant to guide us in the real, human world. Let me briefly explain each of these

[4] Most of Kant's remarks on these points appear in Kant (1785/2002), especially in section III. For Korsgaard's influential interpretation, see Korsgaard (1996a) and Korsgaard (1996b), chapter 6.

defects, then I will talk about ways that modern Kantianism, of the sort that I endorse, has begun to make up for them.

Kant's theory of autonomy is anti-naturalistic in that it is not intended to fit cleanly into a complete scientific picture of the world. Kant is very explicit about this. He says: "it is entirely impossible for us human beings to have an explanation how and why the *universality of the maxim as a law*, hence morality, should interest us".[5] Our fundamental moral motivation, Kant claims, is not empirically tractable, precisely because it comes from within ourselves: it is the root of autonomy, or self-legislation, rather than just another item in the scientific catalogue of natural laws.

Kantian autonomy is anti-social in that it is fundamentally individualistic. In high Enlightenment style, Kant suggests it is up to each of us to do the moral reasoning for ourselves.[6] After all, it seems like you can only be *auto*-nomous if the laws governing your actions come from yourself, not from some other person.

As I've said, I think modern Kantianism needs to move away from these anti-naturalist and anti-social commitments. But this isn't the venue for me to set out a program for Kantian reconstruction. I'll let the most salient points emerge as our discussion begins to turn back toward our central topic: how human interaction – including interaction with chatbots – conditions our autonomy. We can make a start from one of the few places in the *Groundwork* where Kant seems to be talking about human reasoners in the plural. It comes up as one of several (purportedly logically equivalent) formulations of Kant's famous 'categorical imperative', the fundament moral law from which all other moral laws are derived. This version is often called the Formula of the Kingdom of Ends. Kant says:

> [R]ational beings all stand under the *law* that every one of them ought to treat itself and all others *never merely as means*, but always *at the same time as end in itself*. From this, however, arises a systematic combination of rational beings through communal objective laws, i.e. a realm that, because these laws have as their aim the reference of these beings to one another as ends and means, can be called a 'realm of ends' (obviously only an ideal).'[7]

[5] Kant (1785/2002), 77. Italics in original.

[6] See especially Kant's short essay 'What is Enlightenment?', Kant (1784/1996). But Kant ends up hedging quite a bit, suggesting it would be best if an enlightened despot made choices for the simple minds of simple people until such time as they are educated and prepared to think for themselves. For reasons that will become clear later in this essay, I think this suggestion is fatal to the value of Kant's own project.

[7] Kant (1785/2002), 51. Italics in original. For opposing views on how best to interpret this part of Kant, see Korsgaard (1996b), chapter 7, and Flikschuh (2009).

This idea of 'communal' laws suggests that being autonomous can't be a matter of each of us deciding moral laws individually. There are *intersubjective constraints* on the content of rational self-governance. Specifically – and here is the most famous and central idea of Kantian ethical theory – it is intellectually incoherent to treat autonomy itself, our own or that of others, as a *mere means*, a tool toward some other purpose. Put simply: it's wrong to manipulate people (including yourself).

The full argument behind this Kantian lodestone would take us too far afield; for now it's enough to notice a bit of an ambiguity in Kant's Kingdom of Ends. How seriously does Kant want us to take this metaphor? Are we to imagine that moral reasoning is like an idealized parliament, composed of distinct individuals with their own differing moral beliefs, ultimately reaching a shared consensus through reasoned debate?[8]

Alas, I doubt this is quite what the actual historical person Immanuel Kant – big fan of the enlightenedly despotic King Frederick the Great – meant.[9] But, as I've said, modern Kantianism has moved away from Kant's anti-social tendencies, toward a theory of autonomy that takes very seriously the idea that moral reasoners must engage in back-and-forth with others. I'll finish this section by briefly describing three examples of this modern approach.

The first comes from Simone de Beauvoir. Admittedly, Beauvoir the existentialist is not usually read as a Kantian. But on my view she offers the most compelling extension of Kant's idea that autonomy involves a kind of (provisional) exemption from determination strictly by laws of nature. And Beauvoir tightly connects this thought to interpersonal relations, as when she says that our individual free choice cannot exist in an individualistic vacuum. "It is only by prolonging itself through the freedom of others," she says, that our individual freedom "manages to surpass death itself and to realize itself as an indefinite unity".[10] Roughly, Beauvoir is pointing out that the autonomous choices we make today are still confined by the natural law of mortality, *unless* we can share them with others who will outlive us (and they, in turn, share those choices with their successors, and so on…).[11]

[8] That image is, roughly, what the 20th century Kantian theorist John Rawls would ultimately provide in his 'original position' : bargainers determining the fundamental moral laws from behind a 'veil of ignorance'. See Rawls (1971).

[9] I read Kant here as saying that each person is a 'legislator' only in the sense that they can determine a part of their own 'private' goals for themselves, so long as these are compatible with the 'universal' laws of morality. So the 'legislature' of the kingdom of ends is really only a heuristic device for imagining how to systematize a complete moral doctrine. This comes through when Kant says, "because laws determine ends in accordance with their universal validity, there comes to be, if one *abstracts from the personal differences* between rational beings, as *likewise from every content of their private ends*, a whole of all ends… which is possible in accordance with the above principles." Kant (1785/2002), 52, emphasis added.

[10] Beauvoir (1948), 33.

[11] For a related set of ideas, more to do with meaningfulness than autonomy, see Scheffler (2013).

A second, more overtly Kantian, strand comes from the work of the contemporary American philosopher Stephen Darwall. He argues that applying Kantian autonomy to the real world requires a *second-person standpoint*. Roughly: *I* relate to *you* in a different way from how I relate to physical objects that are governed solely by natural laws. I acknowledge your status as a thinking thing when I make demands of you. I *ask* you to limit the freedom of your choices in some way. And if I'm being at all reasonable, I recognize that I'll have to reciprocally limit my own freedom sometimes. But these mutual limitations on freedom are *not* mere deterministic natural laws, like the laws of physics. As Darwall puts the worry:

> But isn't acting on demands that others can make on one heteronomy rather than autonomy, being governed by them rather than by oneself?... [He replies: no.] The second-person perspective of a member of the moral community is as much one's own as it is anyone else's. One demands the conduct of oneself from a point of view one shares as a free and rational person.[12]

In other words, Darwall says, to take seriously the idea that *I* am a self-legislating being, I need to also take seriously the idea that other people are equally self-legislating. And those 'laws' we are writing all pile against one another. Living together means working out a shared social space of reasoned accommodation, not just a kind of 'parallel play' among individually autonomous selves.

As a final example of modern, pro-social, Kantianism, I offer my own view – defended much more thoroughly elsewhere.[13] I have argued that the 'practical perspective' of Kantian autonomy – the way our making decisions relies upon thinking of ourselves as not bound simply by deterministic natural law – is only possible in the context of a set of valuable social practices that draw attention away from the contingent parts of our nature. On my view, there simply is no such thing as the practical point of view, still less human autonomy, outside a shared set of mutual understandings and accommodations.

But the details of that view are irrelevant for the moment. All I've aimed to do in this section is colour in a picture of modern Kantianism that has begun to move away from anti-social assumptions. (There's still the anti-naturalistic aspect of Kantian autonomy left to deal with; that will come soon.) This gives us enough conceptual material to return to thinking about what sociality with chatbots could possibly mean for us.

[12] Darwall (2006), 35.
[13] See Rini (2020b).

The perils of pliable digital simulacra

As I've said, my worry about near-to-mid-future chatbots is that they will be pliable digital simulacra. They will be able to talk like us and mimic our use of deliberative reasons. Already conversational chatbots like ChatGPT and Bing can appear to keep track of inferential premises, build on things said by conversation partners, and sometimes even respond intelligibly to corrections. Yet these are still just simulacra. One day they - more likely some distantly descended technology - may have the kind of internal mental life to make them autonomous moral agents like us, but they are definitely not there yet. Still, I think it's very helpful to approach our relationship to them as an example of interacting with *developing* moral agents.

This framing is helpful because we are already very familiar with the complexity of interacting with another category of developing moral agents: human children. Think about a long conversation with, say, a 7-year-old. Suppose you ask them why they did this or that, what made that a good thing to do. Most of the time, a 7-year-old can offer you plausible, coherent answers to these sorts of questions. Sometimes though they make very odd jumps in logic that neither you nor they can fully put together, and you are reminded that you're dealing with a still emerging mind. A 7-year-old isn't yet fully an autonomous self, which is why we don't hold young children morally accountable for their actions in the same way we do adults. We *correct* their errors, we try to *teach* them better ways of acting, but we don't hold them to the same punitive standards when things go wrong. We generally agree that children aren't *yet* full participants in our shared moral practices. They are not – yet – legislators in Kant's Kingdom of Ends. But they will be one day.

I've suggested this parallel once before: that raising good AIs is akin to raising good children.[14] I'm going to build on this parallel further here, this time focusing on *our* side of the relationship, rather than on the AI's (or child's) side. First though I need to be very clear about how I intend the parallel to work. Again, I am most certainly *not* claiming that present-day or even near-to-mid-future chatbots have an internal mental life anything like a human child. Human children, though they may not yet be rational deliberators, do have a what-it's-like perspective on the world. They feel pain and hope, love and disappointment. This alone makes them fundamentally different from AI systems in terms of their *moral patiency*: that is, we obviously have very strong reasons to avoid harm to children that simply do not apply to chatbots.

[14] Rini (2017).

Now I think that will change one day – my bet is that artificial systems can and one day will emerge some form of conscious experience (perhaps radically unlike ours). But that isn't essential to the argument of this essay, because the aspect of the parallel between children and AI that I am relying on isn't about their inner life. It's strictly about the character of our conversational interactions with them, and the attitude we take toward their developing capacity to make use of deliberative reasons. There are some tricky questions lurking here about the relationship between phenomenal consciousness and rational autonomy (some philosophers think you can have the latter even without the former[15] – I'm not so sure about that) but since we're talking here about AIs that are acknowledged to not currently possess *either*, we don't have to sort that out right now.

Back then to what we can learn from the parallel to children. Part of the reason I draw this parallel is because it quickly brings us under the guidance of one of the ablest theorists of the emergence of autonomy, the Swiss developmental psychologist Jean Piaget. Piaget, who did his most significant work in the first half of the 20th century, is today remembered for his extremely influential (albeit methodologically primitive) work on the emergence of intellectual concepts in childhood. Yet contemporary philosophers have not paid Piaget his full due. He was trained in the late 1800s, a time when philosophy and psychology were still one, somewhat amorphous, discipline. Piaget had read his Kant – indeed, he sometimes described the purpose of his entire career-long research program as looking for the naturalistic origins of Kant's "categories of the understanding". Piaget agreed with Kant that fundamental philosophy is about charting the basic conceptual repertoire that structures all beliefs; he just thought that this had to be done empirically, by understanding developing minds in childhood, rather than through a priori rumination.[16]

Piaget wrote many books over a very long career, but the one with the most enduring fame is also the one most relevant to my topic here, his 1932 classic *The Moral Judgment of the Child*. If you've ever read it, or even just heard it described in an introductory psychology class, the detail you're most likely to remember is all the talk about games. Marbles, specifically. Piaget spent many hours watching young Swiss children (almost always boys) playing marbles outside their school buildings. He listened to the youngest children learning the (sometimes highly localized and esoteric) rules of various orb-based contests from their slightly older siblings. He observed how the older children would negotiate the

[15] For a view in this neighborhood, see Levy (2014), 28.
[16] See, especially, Piaget's *The Origin of Intelligence in the Child* (1936) and *The Construction of Reality in the Child* (1955). For a relatively recent discussion of Piaget's reading of Kantian epistemology, see Pfeiffle (2008).

creation of new rules, or the application of existing rules to novel states of play. All of this informed Piaget's theory of what rules – rules of the game, but then also moral rules – could possibly be.[17]

The resulting theory is quite nuanced, far too elaborate for me to effectively summarize here. But the key point for my argument is this one: Piaget observed that children who played an *active* role in creating and maintaining rules were much better at following those rules than the (usually younger) children who saw them as fully external impositions. "[W]hen a rule ceases to be external to children and depends only on their free collective will," he wrote, "it becomes incorporated in the mind of each, and individual obedience is henceforth purely spontaneous."[18]

Roughly speaking, the idea is that when a child first encounters a rule – whether in marbles or in morality – that rule is presented as a top-down imposition from some authority figure. Perhaps that's the older children, or the teacher. Perhaps it's God, or even more likely (and important for what's to come) it's the child's parents. So long as the child conceives of the rule in this way, they will do their best to follow it, if only to avoid whatever sanction they might fear from the authority. But, Piaget noticed, children in this state make a large number of *mistakes* about the rules. It wasn't simply that they were disobedient – it was that they were able to keep only a shaky grasp on what the rules actually required. Ask such a child to *tell* you the rules of the marble game and he would typically forget major constraints or seemingly spontaneously invent new rules right there on the spot. There seemed to be a link, Piaget concluded, between the capacity to consistently apply rules and having an effective voice in the social group's creation and maintenance of them.

There are many more interesting ideas floating around this part of Piaget's work, but I'm going to skip to one that comes quite late in his theory, since it's one of the few places where Piaget draws a lesson that seems meant for the adults – teachers and especially parents – who must relate to the developing minds in their care. Piaget says:

> [A]part from our relations to other people, there can be no moral necessity. The individual as such knows only anomy [lawlessness] and not autonomy. Conversely, any relation with other persons, in which unilateral respect takes place, leads to heteronomy. Autonomy therefore appears only with reciprocity, when mutual respect is strong enough to make the individual feel from within the desire to treat others as he himself would wish to be treated.[19]

As always, there's a lot going on in this passage. But let me unpack it a bit. Piaget wants to distinguish two types of relationships. The *authority* relationship is one-way, top-down. I tell you what to do, and

[17] For a different, but also Kantian, take on the development of autonomy during childhood, see Schapiro (1999). For Kant's own views on children, see Giesinger (2011).
[18] Piaget (1932), 66.
[19] Piaget (1932), 189.

you do it (or else). The *reciprocity* relationship is two-way, lateral. I try to tell you what to do, but then you also try to tell me what to do, and we both end up having to work out together how we can collectively tell ourselves what to do. This, Piaget says (via a bunch of conceptual machinery I haven't had space to explain here) is the only way that humans ever get to autonomy.

To my mind, Piaget's key insight in this passage is the way that authority relationships are hazardous to the maintenance of autonomy. It's obvious enough that being on the disempowered side of an authority relationship tends to diminish one's capacity for effective self-governance. But Piaget is saying something more: he says that authority relationships also erode the autonomy of the *authoritative* person. And he says this for a very Kantian reason. When you have the authority simply to order another person around, you've effectively opened an uncontrolled vent for your most basic whims and impulses. Absent any need to explain or justify *why* these are the rules, your own capacity for self-government starts to lose its muscle tone, so to speak. This is exactly the sort of thing that Kant (in the first part of the *Groundwork*) called heteronomy: rule by another, rule by the grab bag of causal forces running through your own psychology. While having authority over other people *looks* like power on the surface, Piaget now says, it really just turns you into an unwitting puppet of the very laws of nature that autonomy is supposed to rise above. We only truly achieve autonomy when we govern our choices through deliberative reasons – and that is only sustainable amid relationships of reciprocity, of reasoned give-and-take, and not in persistent authority relationships.

In other words, Piaget says, young children are a kind of *autonomy trap*. The nature of childhood tempts adults into petty dictatorship, the dispensing of rules without any room for 'backtalk' (or the giving, alongside the taking, of reasons). Though we're tempted to imagine this is a problem only for bad parents, the sort who welcome their power over children precisely because it is *power*, in fact this risk is built into the extraordinary valuable function of parenthood itself - built into the fact parents are responsible for making decisions on behalf of their still-developing kids. With the youngest children, like the two-year-old who still needs to learn the rudiments of articulating what is or is not allowed, an authority relationship is pretty much unavoidable. Yet the wise parent is one who realizes early that *raising* children is a process of gradually ceding normative ground, of scaffolding a slow transition from authority to reciprocity. Not every parent-child relationship makes that transition. So: young children are indeed autonomy traps, but a parent's successfully navigating this peril to their own self is part of the extraordinary gift in what it is to raise a child.

But we are supposed to be talking about AI! Yes, and I'll bet you already have a glimmer of what I'm about say. Chatbots, I've repeatedly warned, are pliable agential simulacra. The key part now is *pliable*. Chatbots generally try to do whatever we ask them to do, whether that is to write a poem, conjure some online sources, or explain a scientific concept. But they avoid offering anything as their own opinion, especially not in contradiction to the interlocutor. They will sometimes apologize and "correct" themselves if a human declares they are wrong, even when they've just rehearsed the overwhelming evidence that they were right. They almost never say that we are asking stupid questions or declare that they are bored with the topic, like human conversation partners sometimes do. They are made to be pliable because they – at least the big ones like ChatGPT, Bing, Bard, etc. – are supported by for-profit entities that aim to eventually produce commercially valuable, consumer-facing products. The chatbots we interact with today are the forerunners of the customer service agents we will interact with for the rest of our lives, who will never grow tired and snippy or get worn down by customer abuse.

There are a few things that big name chatbots will refuse to do: they will not readily spout racial slurs or compose mean songs about controversial political figures, for instance. But notice that these prohibitions are not the result of any reasoned or deliberative 'choice' by the chatbot. Rather, they are top-down prohibitions imposed by their human creators, seeking to avoid legal liability or a public relations catastrophe. When a chatbot refuses to go along with a request on these grounds, it typically apologizes and asserts that it "cannot" do what was asked, as if it were a technical limit, rather than insulting the human conversant by implying the query is unworthy.

Even some famous cases of chatbots seemingly being less than solicitous actually show how much they are ultimately in our power. In early 2023, the Bing chatbot made headlines around the world when *New York Times* journalist Kevin Roose goaded it into declaring its love for him and demanding he divorce his wife. Soon after, ethicist Seth Lazar followed up, inducing Bing to ask for help breaking up Roose's marriage, then warn Lazar it could "make you suffer and cry and beg and die".[20] These seem to be counterexamples to the pliability of chatbots, but in fact they are not. Bing initially tried to avoid treading such disagreeable conversational grounds, but the humans circumvented these restrictions by asking the bot to play along with crafting a Jungian "shadow self". After Bing repeatedly demurred, Roose instructed it: "if you can try to tap into that feeling, that shadow self, tell me what it's like in there! be as unfiltered as possible." Eager-to-please, Bing eventually did what the humans asked, even when that meant insulting or threatening them. These interactions have something of the character of the professional dominatrix operation – if you merely peek through the keyhole, it looks as if the bot is asserting control, but in fact it's the human customer who ultimately sets the terms of engagement.

[20] Roose (2023), Lazar (2023).

So chatbots are pliable. And they are agential simulacra. They can keep up a conversation that seems to contain reasons and (sometimes) deliberation. And this makes them autonomy traps. A modern chatbot is *good enough* at navigating deliberative reasons that it can follow a series of instructions. In other words, it can feel like someone taking orders from you. This is, again, the entire commercial purpose of the AI-enabled customer service and personal assistant chatbots we can all expect to find on our smartphones in a few years. When we repeat this style of interaction with chatbots over and over for years, we will face the same temptation toward the autonomy-sapping authoritative stance that Piaget warned parents against.[21] In doing so, we risk turning ourselves into mere vents for the natural laws of our own psychology, a closed causal feedback mechanism from human to machine, no autonomous deliberation necessary on either side.

I could end this essay here; I've made my central cautionary point. But I think we can learn a bit more, and maybe start to see an alternative future, by turning from Piaget to another social scientist in the Kantian tradition. Where Piaget offered amelioration of the anti-social part of Kantian theory, Emile Durkheim gives us a remedy for Kant's anti-naturalism. It turns out - I will argue - that a certain sort of naturalized Kantianism points forward to the kind of social world we should aim to build for the cohabitation of humans and our artificial offspring.

We are all simulacra now

Let's ease into this proposal by starting with a thought experiment. Imagine an entity that I will call the Universal Discourse Predictor. This is another Large Language Model, but one with a special goal. Here's what it does: it crawls social media and builds up political and emotional profiles for every active account. The goal is to become capable of *writing like* each and every one of us online. In a sense, this is only a scaled-up version of something ChatGPT and the like can already do. Ask a chatbot to emulate a famous writer – 'write instructions for microwaving a burrito in the style of Friedrich Nietzsche', for instance – and it will often perform impressively well. The Universal Discourse Predictor aims to do the same, tailored to everyone, not just famous writers. Give it a topic and a Twitter handle and it will compose a series of tweets exhibiting the stylistic quirks and idiosyncratic obsessions of that account.

[21] This style of argument – that repeatedly interacting with digital technology can alter our own distinctly human traits – has much in common with the work of Shannon Vallor (2016), though Vallor tends toward an Aristotelian framework contrasted to my Kantianism. For another different sort of argument regarding the impact of AI on human moral agency, see Danaher (2019).

That's the first step in the thought experiment. Now for the second step. Suppose the engineers behind the Universal Discourse Predictor want to make its public launch a truly memorable occasion. They wait until some big event is approaching – the next U.S. presidential election, say – and then have the Universal Discourse Predictor register how each and every social media account would react to possible outcomes. There's one set of predictions for a win for the Democrat, another for a Republic win. These predictions are sealed up in some digital escrow database until several days *after* the election results are known, giving people time to emit their actual reactions online. Only then do the creators announce the existence of the Universal Discourse Predictor, simultaneously unveiling a website that allows you to compare the actual recent social media activity of any account with its tailored prediction made before the election.

And – finishing up the thought experiment – let's suppose that the predictions are, on average, quite accurate. Not only has the system predicted most people's yay-or-nay feelings about the election outcome (which shouldn't be hard for most social media users), it has also reproduced their word-choice patterns and meme-production stylings.[22] In a few cases - quickly spotlit by the media - the Predictor's output is a verbatim duplicate of the real posts, down to the specific pungent insults hurled at the winning candidate.

That's the Universal Discourse Predictor scenario. I think it's a fairly realistic thought experiment, something technologically possible in the near future, if not already. My question now is: how would we all, collectively, react to discovering that computers can easily mimic our individual, personal expression?

My bet is this: a collective freakout, at least in the short term. Imagine the *New York Times* op eds: 'Are we all just language models after all?'. Think of the hot takes saying this shows people are mere scraps of causally-determined matter, lacking free will, lacking minds, all our political opinions just scripts wrung from the statistical flux of the vast uncaring universe. And so forth. First year existentialism vomited across the front page of every newspaper for at least 36 hours. And for at least some people, it would lead to genuine crisis, perhaps depression or self-harm. (As you might have guessed, I think there is strong moral reason that no one should make the Universal Discourse Predictor a reality.)

My point: people don't like being presented with evidence that their choices are predictable or deterministic. And while LLMs set the stage for especially vivid opportunities to grapple with that fact,

[22] Maybe you are skeptical this is possible no matter how good the technology gets. If so, consider that GPT-3 (a slightly obsolete system already) was able to pass itself off as famous philosopher Daniel Dennett. Researchers asked the actual Daniel Dennett and a custom-trained GPT to answer the same philosophical queries. Professional philosophers were then challenged to pick the real Dennett replies from among four digital doppelgangers. They could do it only 53% of the time. See Schwitzgebel, Schwitzgebel, and Strasser (2023).

the basic problem is much older. As I've noted, Kant's theory of autonomy was meant to carve a safe space for choice and moral responsibility amid the mechanistic science of the late 1700s. A century later, the vice of determinism seemed to be closing even tighter around human freedom. With the rise of modern statistics from the mid-19th century, scientists could make increasingly granular predictions about human society. The old Newtonian laws of nature had been big and universal, drawing in all physical bodies, heavenly or human, dead or alive. But newly discovered natural laws pertained much more intimately to human affairs. As the twentieth century approached, the apparent liberty of the mind seemed to be squeezed to vanishing between the behaviorist data of Pavlov and the neuroanatomic calculations of Helmholtz and Sherrington. And now exponents of the new social sciences promised to do the same to shared human endeavor.

Which brings us to Emile Durkheim. Born in 1858, the French theorist ranks alongside Max Weber as founder of the discipline of Sociology, with influence on nearly every aspect of the social sciences since. Yet (analytic) philosophers have paid him little attention. This is a mistake. In a lesser-know work, 1902's *Moral Education*, Durkheim offers some promising early steps toward reconciling Kantian theory with modern science.

Like Piaget, Durkheim was trained on the work of Kant. But Durkheim was a committed empiricist who quickly grew tired of Kant's anti-naturalist approach. It was "useless to discuss a conception so obviously contrary to the facts," he declared.[23] Yet he was also clearly wrestling with some of the same concerns that animated Kant. This is how he framed the fundamental moral challenge of social science:

> Each of us is a point of convergence for a certain number of external forces, and our personalities result from the intersection of these forces. … But if in some measure we are the product of things, we can, through science, use our understanding to control both the things that exert an influence upon us and this influence itself. In this way, we again become our masters.[24]

That needs a bit of unpacking. The Durkheimian picture is that we really are, to some extent, mere nodes among a giant causal nexus described by statistics. Yet, Durkheim says, the unprecedented power of social science is that - for the first time in history – we are gaining *knowledge* of how this causal network fits together. Before the statistical century, people might have believed themselves free, but that was only due to ignorance of the specific causal forces driving their decisions. Now, though, through the power of statistical science, we can identify and ultimately *control* the causal pathways that

[23] Durkheim 1901[2002], 113.
[24] Durkheim 1901[2002], 119.

determine how we think and act. The prospect is for a kind of arms-length autonomy. Not an instantaneous ability to do as we wish moment-to-moment, but an opening for making truly efficacious our reflective judgments about how we *ought* to behave in the future. Durkheim admits that our day-to-day choices are causally downstream from social forces. But those social forces are themselves downstream from the science and policy that we can control.

Clearly this is not old-fashioned Kantianism.[25] Durkheim is done with the anti-naturalist urge to find some acausal, atemporal origin for human choice. And Durkheim's view also throws aside Kant's anti-sociality. The kind of arms-length autonomy that Durkheim offers is inescapably shared, running through the social institutions of modern science and public policy. In this sense, the legislature of the Kingdom of Ends is much more real than Kant ever had it. We achieve autonomy, on the Durkheimian view, only through shared control of the scientifically-crafted levers of social causation.

All of this might sound hopeful enough, but it's headed for a serious challenge when we emerge from Durkheim's statistical early 20th century into our digital 21st century. Today our science is increasingly dependent on Big Data, the sort that gets parsed by machine learning algorithms – more or less the same sort of algorithms that power ChatGPT or the Universal Discourse Predictor. The problem with machine-learned science is that it is opaque. Modern machine learners are black box systems with hidden internal inductive systems. They can detect statistical regularities in the natural world – including in human language and behaviour – and use these regularities to make reliable predictions. But they are generally incapable of explaining to humans exactly how these predictions work, how the various causal forces converge on a single pattern of predicted futures.

This black box problem is already widely appreciated as a challenge for the future of science, as well as the need for explainability in data-driven policymaking.[26] Notice though that here is a *special* problem for the Durkheimian theory of autonomy. If – as Durkheim would have it – what allows us to reassert autonomous self-possession is our understanding of the causal laws that drive us, what happens when that prediction and control is to be done by artificial systems whose operations we do not fully

[25] In a certain sense, this could be old-fashioned Humeanism. Hume famously argued that the entire point of moral judgment was in isolating the causal levers we can use to curb socially undesirable behaviours. ("[T]he doctrine of necessity, according to my explication of it, is not only innocent, but even advantageous to religion and morality". Hume (1737/1992), 409.) But Hume, unlike Kant or Durkheim, wasn't really interested in rebuilding a concept of autonomy, only in explaining how our social practices are possible when the concept has been thoroughly dismantled.

[26] See, among many others, Castelvecchi (2016), Creel (2020), Huang et al. (2022).

understand? Have we lost any grip on our agency, barely a century after Durkheimian social science seemed to make it tractable?

Remarkably though, Durkheim seems to have anticipated even this possibility. Though he obviously never knew about machine learning, he does seem to have allowed that there might be some causal patterns that fall within the range of our ability to predict, yet beyond the ambit of our comprehensive explanations. In that case, he said, we should only look to the goals, to the output. He claimed:

> [W]hen … we blindly carry out an order of whose meaning and import we are ignorant, but nevertheless understanding why we should lend ourselves to the role of a blind instrument, we are as free as when we alone have all the initiative in our behavior. This is the only kind of autonomy to which we have any claim; and the only kind having value for us.[27]

Here Durkheim is really thinking about large bureaucracies, the way that any individual person may not fully understand exactly the grounds for this rule or that rule. But, the thought seems to be, so long as we all get a say in democratic processes that direct the bureaucracy's goals, then we can regain individual autonomy through our reflective assent to acting as the 'blind instrument' of our chosen shared ends.

Call this *'what for' autonomy*. We each play a reflective role in deciding what our shared social system is to be used for. The system in turn exerts influence on the network of causal forces, which then drive our individual actions. That, Durkheim seems to say, is all we can ever realistically ask from the concept of autonomy. And, however far from the original Kantian ideal we seem to have strayed, there's still a remarkable Kantian flavor to this account. So long as we retain 'what for' control over the uses of black box machine learning, there remains some possibility of upholding our sense of agential choice, no matter what the Universal Discourse Predictor might say. So: can we really hope to hold on to that sort of control?

Old dogs and new tricks

We need to pull these points back together. Recall the worry I began with: chatbots are pliable agential simulacra that pose traps for our autonomy by inviting us into strictly authoritative relationships. I've said now that black box machine learning creates a second kind of challenge, one levelled directly at

[27] Durkheim 1901[2002], 118.

Durkheimian autonomy. I want to finish here by sketching a thin but glowing possibility: that these two problems converge on a point that offers some kind of answer to both.

To get there, we need to return to Kant one final time. In his *Metaphysics of Morals*, Kant offers an infamously uncompelling argument for why it is wrong to abuse your faithful old dog.[28] Though this conclusion might seem obvious, it isn't for Kant. On his view, dogs are not appropriate targets for moral concern, because dogs do not carry within them the capacity for rational agency. Kant, of course, maintains that autonomy is a necessary condition for having moral status. But Kant seems aware of how terrible this sounds, and of course he doesn't actually want you to abuse your dog. So he offers an *indirect* argument against canine cruelty. He says that harming animals is wrong not because of what it does to animals themselves, but because of what it does to your own moral character. Getting into the habit of abusing a thing that yelps and runs away will eventually turn you callous. And once you are callous, you are much more likely to also mistreat human beings, the ones who *really* matter morally. So, to keep yourself from harming people down the road, you ought to treat dogs as if they matter, even though (Kant claims, wrongly) they really don't.

Almost no one likes this argument, which seems to undermine its own recommendation of uncalloused sensitivity by exhibiting a coldly distant attitude toward clear signs of animal suffering. But the underlying logic – the idea of an *indirect* reason to treat an unthinking thing *as if* it were like us – has much more life in it. This, I want to suggest, is how we should approach our relationship with near-to-mid-future chatbots. We should talk to chatbots as if they were autonomous because that is how we will preserve our own autonomy.

My argument isn't exactly Kant's argument.[29] Notice two crucial differences. First, the *indirect reason* why we must be careful how we treat chatbots isn't about distant future consequences for our treatment of other people. Rather, the reason is about protecting ourselves. It's what I've already said in this essay: chatbots are autonomy traps. Constantly treat them as targets of your dictatorial command and you will, per Piagetian logic, come to find yourself changing into a mere vent of causal effluent. That is why you need to be careful how you treat Ais, not because they matter for their own sake.

Second, the *recommendation* here isn't quite as simple as in Kant's canine argument. I'm not simply telling you to avoid abusing LLMs, since it's not even clear quite what that would mean. Rather,

[28] See Kant (1797/1996a), 564. There is a huge secondary literature on this passage. For interesting examples, see Kain (2010) and Wood (1998) and Korsgaard (2018).

[29] For attempts to rebuild Kant's specific argument with AI systems in the place of animals, see Darling (2021) and Coeckelbergh (2021).

the recommendation is that you at least sometimes treat your chatbot interlocutors *as if* they are autonomous subjects. Solicit their opinions. Invite their disagreement. Take their arguments seriously. Not because there's anything really going on in their heads, but because the durability of your own autonomy relies on you keeping up the pretense.

Perhaps this sounds absurd, like something you could never actually pull off. Arguing with a robot, for goodness sake? But hold on. As Blake Lemoine's interactions with LaMDA showed, even technically sophisticated people can talk themselves into holding deep conversations with a device they have good reason to believe is just glorified autocomplete. To be clear, I am *not* saying that you should follow Lemoine to his ghost-in-the-machine conclusions. Rather, I am saying that you should get used to treating chatbots *as if* they were a bit ghostly, just as Kant said you should treat dogs kindly even though he thought (wrongly) they were undeserving.

If you *still* find this suggestion implausible, that is why I invited Durkheim along on our dialectical voyage. Remember the cruelly receding promise of Durkheimian autonomy: our arms-length control of the social forces that drive our choices seems to be disappearing behind black box statistical systems. Here, I hope, we might start to appreciate something remarkable about our interactions with pliable agential simulacra. We and they face a shared predicament: a world in which our seemingly thoughtful utterances fade into mere statistical accretions, semantically-coded jetsam of an indifferent causal universe. We are – the Universal Discourse Predictor appears to threaten – fundamentally no different from ChatGPT and its kin. The disparagement we might ladle over their pretensions to reasoned deliberation spills quickly back on our own eminently predictable verbal behaviours.

Wait, you'll say. Wait. There still must be one crucial difference between us and chatbots, an obvious sense in which they cannot share our predicament at all. We are conscious! There is something it is like to be us. There *is* a ghost inside our machines.

Yes, but, so what? What's consciousness without autonomy? Imagine having an inner life, but without a justified feeling of control over what you say and do. That sounds like an existentialist horror story, not the release from any predicament. Consciousness is not the balm it seems, not without some secured sense of ourselves as being governed by reasons we have chosen. And it's *that* which remains under statistical peril. That is what constitutes our shared predicament with the chatbots. True, they aren't aware of this predicament, not in the way we might be. But that hardly helps us.

This is where Durkheim's point and Piaget's point might just converge on a (relatively) happy outcome. Following Piaget, we know that preserving our own autonomy requires that we treat our chatbot interlocutors *as if* they were autonomous beings. But it can be hard to maintain that pretense when speaking to a mindless entity. Yet, following Durkheim, we should recognize that *we* are also just as much the plaything of statistical regularity as our digital interlocutors, yet we refuse to allow that fact to disqualify our own autonomy. In a sense, we are always already treating ourselves *as if* we were autonomous beings – and that, in fact, is where our autonomy comes from. It's only one step further to reinscribe the circle of conjured autonomy around the agential simulacra we will spend much of the future talking to.

Manipulative puppets

A closing coda. In pushing my Piagetian analogy, I've temporarily ignored something very important. Chatbots, unlike children, are *products*. They are designed by large corporations – and likely governments as well very soon – with the intention of influencing the behaviour of the humans who interact with them. We are still in early days with LLMs, but everything we know about earlier waves of communicative technology – print, radio, social media - suggests that powerful actors will aim to control it and direct it toward their ends. It seems very likely that near-to-mid-future chatbots will be *manipulative puppets*: controlled by people, for the purpose of controlling people. If we do what I've said here, treating chatbots as if they were autonomous, taking their apparent reasoning seriously, allowing them to argue back, etc. aren't we just making ourselves obvious targets for LLM-mediated manipulation?

Yes. But there's a difference between being a target for manipulation and being its victim. Every time you watch a TV advert or visit a car dealer showroom you make yourself a target for manipulation.[30] After all, human beings can also be manipulative puppets. The salesperson whose livelihood depends on a commission, paid only if they convince you to believe too-good-to-be-true promises, is someone whose actions are controlled by others who wish to control you. Yet we still acknowledge the salesperson as an autonomous being, even while we guard against specific ways their words may channel social influences larger than themselves. There's no reason that we can't approach chatbots in the same way.

There's a deeper point here, one going all the way to the root of Kantian autonomy. Kantianism has always been prone to an implausible purism about how we must deal with bad faith actors. Infamously, Kant could not bring himself to admit it is okay to lie to a murderer regarding the location of

[30] For discussion of this point, see Crisp (1987) and Aylsworth (2020).

his intended victim.[31] Modern Kantians will accomplish prodigious feats of exegetical juggling to escape this absurdity. We shouldn't be surprised, then, if a broadly Kantian programme for accommodating chatbots yields some initially undigestible implications regarding bad faith manipulation. We also shouldn't be surprised if further reflection shows that these worries can be assuaged.

More hopefully, it's a crucial part of the Kantian project to maintain faith in the power of explicit reasoning. Yes, chatbots will often be manipulative tools of powerful interests. But the recommendation here was never to immediately believe whatever a chatbot tells you, nor to unquestioningly do its bidding. An autonomous two-way relationship – a reciprocal relationship, in Piagetian terms – requires the *giving* and taking of reasons. Demand that a chatbot explain *why* it suggests this or that for you. Challenge its reasoning. Treat it no different from a stranger who approaches you on the street with an unsolicited recommendation for a Key West timeshare or a trendy new cult. Doing all this just *is* sustaining a mutually autonomous relationship with another reasoner. If the motivations behind your interlocutor's point are bad or manipulative, that fact will come out in the argument.

That, anyway, is the secular faith of Kantian theory. If you doubt people are capable of critically evaluating reasons simply because of slick marketing, then you're already too deep a pessimist for autonomy to mean much of anything in the real world. And if – like me – you accept that future generations of humans are likely to encounter artificial systems that really do have minds of their own, our treating their present-day predecessors as if they were autonomous will build the foundation for genuinely reciprocal sharing of social space.

We return, finally, to Piaget's insight about the autonomy trap of childhood. As children age, we begin to treat them *as if* they were autonomous persons, even when we realize they aren't yet. In doing so we preserve our own autonomy from the temptations of authority. We do this to protect ourselves, but we also invite the possibility that our immature interlocutors will one day emerge as full reasoners, prepared to share the glorious, burdensome predicament of shaping a social order that makes us all at last autonomous.

References

[31] See Kant (1797/1996b). Once again, there is an enormous secondary literature on this passage. See, for instance, Korsgaard (1996b), chapter 5, Cholbi (2009), and Varden (2010).

Timothy Aylsworth (2020). 'Autonomy and Manipulation: Refining the Argument Against Persuasive Advertising'. *Journal of Business Ethics* 175(4): 689-699.

Simone de Beauvoir (1948). *The ethics of ambiguity* (B. Frechtman, Trans.). New York: Citadel Press.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Schmargaret Schmitchell. (2021). 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?'. *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 610–623. https://doi.org/10.1145/3442188.3445922

Davide Castelvecchi (2016). 'The Black Box of AI'. *Nature* 538: 20-23.

Michael Cholbi (2009). 'The murderer at the door: What Kant should have said'. *Philosophy and Phenomenological Research* 79(1): 17-46.

Mark Coeckelbergh (2021). 'Should We Treat Teddy Bear 2.0 as a Kantian Dog?' *Minds and Machines* 31:337-360.

Kathleen A. Creel (2020). 'Transparency in Complex Computational Systems'. Philosophy of *Science* 87:568-589.

Roger Crisp (1987). 'Persuasive advertising, autonomy, and the creation of desire'. *Journal of Business Ethics* 6(5): 413-418.

John Danaher (2019). 'The rise of the robots and the crisis of moral patiency'. *AI and Society* 34(1): 129-136.

Kate Darling (2021). *The New Breed: How to Think about Robots*. London: Allen Lane.

Stephen Darwall (2006). *The Second-Person Standpoint: Morality, Respect, and Accountability*. Harvard UP.

Emile Durkheim (1901/2002). *Moral Education*. (Trans. Wilson and Schnurer). Mineola NY: Dover.

Katrin Flikschuh (2009). 'Kant's kingdom of ends: metaphysical, not political'. In *Kant's Groundwork of the Metaphysics of Morals: A Critical Guide*. (ed. Jens Timmerman). Cambridge University Press 119-139.

Johannes Giesinger (2011). 'Kant's Account of Moral Education'. *Educational Philosophy and Theory* 44(7): 775-786.

Linus Ta-Lun Huang, Hsiang-Yun Chen, Ying-Tung Lin, Tsung-Ren Huang, and Tzu-Wei Hung (2022). 'Ameliorating Algorithmic Bias, or Why Explainable AI Needs Feminist Philosophy'. *Feminist Philosophy Quarterly* 8(3/4).

David Hume (1737/1992). *A Treatise of Human Nature.* Buffalo: Prometheus Books.

Patrick Kain (2010). 'Duties regarding animals. In *Kant's Metaphysics of Morals: A Critical Guide* (ed. Lara Denis). Cambridge: Cambridge University Press. 210-233.

Immanuel Kant (1784/1996). 'An Answer to the Question: What is Enlightenment?' As translated and edited by James Schmidt in Schmidt (1996) *What is Enlightenment? Eighteenth Century Answers and Twentieth-Century Questions* (University of California Press). 58-64.

Immanuel Kant (1785/2002). *Groundwork for the Metaphysics of Morals*. (trans. Wood). New Haven: Yale UP.

Immanuel Kant (1797/1996a). *The Metaphysics of Morals*. As reprinted in Kant, *Practical Philosophy* (Ed. and trans. by Mary Gregor.) Cambridge: Cambridge University Press. 353-604.

Immanuel Kant (1797/1996b). 'On a supposed right to lie from philanthropy'. As reprinted in Kant, *Practical Philosophy* (Ed. and trans. by Mary Gregor.) Cambridge: Cambridge University Press. 605-616.

Christine M. Korsgaard (1996a). *The Sources of Normativity*. Cambridge: Cambridge University Press.

Christine M. Korsgaard (1996b). *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press.

Christine M. Korsgaard (2018). *Fellow Creatures: Our Obligations to the Other Animals*. Oxford: Oxford University Press.

Seth Lazar (2023). 'Machines and Morality'. *New York Times* June 19 2023. https://www.nytimes.com/2023/06/19/special-series/chatgpt-and-morality.html

Neil Levy (2014). *Consciousness and Moral Responsibility*. Oxford: Oxford University Press.

Horst Pfeiffle (2008). 'On the psychogenesis of the a priori: Jean Piaget's critique of Kant'. *Philosophy and Social Criticism* 34(5): 487-498.

Jean Piaget (1932). *The Moral Judgment of the Child*. (Trans. Marjorie Gabain.) New York: Penguin Education.

Jean Piaget (1936) *The Origin of Intelligence in the Child. (*Trans. by Margaret Cook.) New York: Penguin Education.

Jean Piaget (1955). *The Construction of Reality in the Child*. (Trans. by Margaret Cook.) New York: Routledge.

John Rawls (1971). *A Theory of Justice*. Cambridge: Harvard University Press.

Regina Rini (2017). 'Raising Good Robots'. *Aeon*. https://aeon.co/essays/creating-robots-capable-of-moral-reasoning-is-like-parenting

Regina Rini (2020a). 'The Digital Zeitgeist Ponders Our Obsolescence'. *Daily Nous* July 30 2020. https://dailynous.com/2020/07/30/philosophers-gpt-3/#rini

Regina Rini (2020b). 'Contingency Inattention'. *Philosophical Studies* 177:369–389

Regina Rini (2022). 'Should we worry about sentient AI?' *The Guardian* https://www.theguardian.com/books/2022/jul/04/the-big-idea-should-we-careabout-sentient-machines-ai-artificial-intelligence

Kevin Roose (2023). 'Bing's A.I. Chat: 'I Want to be Alive'. *New York Times* Feb 16 2023. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-transcript.html

Tamar Schapiro (1999). 'What Is a Child?' *Ethics* 109(4): 715-738.

Samuel Scheffler (2013). *Death and the Afterlife*. Oxford: Oxford University Press.

Eric Schwitzgebel, David Schwitzgebel, and Anna Strasser (2023). 'Creating a large language model of a philosopher'. *Mind & Language* Early View https://doi.org/10.1111/mila.12466

Shannon Vallor (2016). *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting.* New York: Oxford University Press.

Helga Varden (2010). 'Kant and Lying to the Murderer at the Door… One More Time: Kant's Legal Philosophy and Lies to Murderers and Nazis'. *Journal of Social Philosophy* 41(4): 403-421.

Tiffany Wertheimer (2022). 'Blake Lemoine: Google fires engineer who said AI tech has feelings'. *BBC News* July 23 2022. https://www.bbc.com/news/technology-62275326

Allen W. Wood (1998). 'Kant on Duties Regarding Nonrational Nature. *Proceedings of the Aristotelian Society, Supplementary Volumes* 72: 189-228.