CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	“Computer says no”: Artificial Intelligence, gender bias, and epistemic injustice
Authors	Walmsley, Joel
Publication date	2023-10-17
Original Citation	Walmsley, J. (2023) '"Computer says no": Artificial Intelligence, gender bias, and epistemic injustice', in Edwards, M. L. and Palermos, S. O. (eds.) Feminist Philosophy and Emerging Technologies. Abingdon: Routledge, pp. 249-263. doi: 10.4324/9781003275992-16
Type of publication	Book chapter
Link to publisher's version	10.4324/9781003275992-16
Rights	© 2023, Joel Walmsley. All rights reserved. This is an Accepted Manuscript of a book chapter published by Routledge in Feminist Philosophy and Emerging Technologies on 17 October 2023, available online: https://doi.org/10.4324/9781003275992-16
Download date	2026-09-16 13:48:58
Item downloaded from	https://hdl.handle.net/10468/15300

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

"Computer Says No": Artificial Intelligence, Gender Bias, and Epistemic Injustice.

Joel Walmsley

Department of Philosophy,
University College Cork

March, 2023

Abstract:

Ever since its beginnings, the field of Artificial Intelligence has been plagued with the possibility of perpetuating a range of depressingly familiar kinds of injustice, roughly because the biases and prejudices of the programmers can be, literally, codified. But several recent controversies about biased machine translation and automated CV-evaluation highlight a novel set of concerns that are simultaneously both ethical and epistemological, and which stem from AI's most recent developments; we don't fully understand the machines we've built, they're faster, more powerful and more complex than us, but we're growing to trust them preferentially nonetheless. This paper examines some of the ways in which Fricker's (2007) concept(s) of 'epistemic injustice' can highlight such problems, and concludes with suggestions about re-conceiving human-AI interaction—along the model of *collaboration* rather than *competition*—that might avoid them.

1. Introduction.

The expression "computer says no" was originally a punchline in the BBC sketch comedy show *Little Britain*, but has now become a name for an attitude which is at once thoroughly modern and all-too-familiar. It is exemplified when, in response to an ostensibly reasonable request from a customer or client, a public-facing representative of an organisation defaults to a response that consists in checking the results of an unknown computer process, and then making an unthinking decision on that basis in a way that is contrary to common sense and often deliberately unhelpful (perhaps as a result of either laziness, malice, or incompetence.) I say "all too familiar" because this combination of factors—not really knowing how or why the machine "says" what it does, but trusting it nonetheless, and using it to guide our social conduct—has become increasingly common in tandem with our growing reliance on AI-powered decision-making and recommender systems.

Ever since its beginnings, the field of Artificial Intelligence (AI) has been plagued with the possibility of perpetuating a range of depressingly familiar kinds of injustice, roughly because the biases of the programmers—and society more generally—can be, literally, codified.[1] In this paper, I argue that the intersection of three major factors in recent computer science and AI gives rise to a novel space in

[1] See Gebru (2020) for a recent survey, as well as Adam (1998), and Kember (2003), for deeper treatments of historical examples.

which additional forms of injustice can also arise. Roughly, machine learning techniques often mean that even experts don't quite know how AI systems do the many things that they can. "Big Data" (and related advances in computer power and speed) mean that AI systems do things that would be practically, epistemically, unmanageable for humans anyway. And so we have seen a recent turn in tides whereby people are more and more willing to trust computers (partly because we *have* to), and therefore to accord AI systems with a great deal of epistemic privilege.

The specifically *epistemic* features of this characterisation of both the "computer says no" phenomenon, and the recent advances in AI, suggest two interesting consequences. On the one hand, the fact that we increasingly trust relatively opaque AI systems underlies a number of recent controversies where the use of AI systems has perpetuated various instances of gender bias. On the other hand, there is an interesting parallel with Fricker's (2007) discussion of *epistemic* injustice. In particular, the concept of *testimonial* injustice could be used to characterise the ethical problem that arises when we prejudicially assign a deflated level of credibility to a human decision in comparison to that of an AI system; our increasing inclination to trust AI systems too much may lead to a corresponding inclination to trust each other too little. Further, if—as is often the case—the design of AI systems inhibits or impairs a *user's* ability to understand or communicate their experience, then we also run the risk of committing what Fricker has described as *hermeneutical* injustice.

I conclude by noting that these problems arise specifically because, as Danaher (2019, p.19, my emphasis) puts it: "Machines are now being designed, built, and implemented to *replace*, not simply complement, their human coworkers." Instead of this competitive or oppositional approach, I suggest human-AI *collaboration*—with a judicious division of labour, following Smith (2019), that recognises the tasks at which both humans and computers excel—would be a better way of approaching the issue, such that, instead of "Computer says no!" it might be appropriate to suggest (albeit less snappily) "Computer says 'Let's work on this together'."

2. Three Trends.

Let's start by outlining three recent developments in computer science, AI, and human-computer interaction that have come to prominence over the last decade or so; machine learning, big data, and the growing reliance on (or trust in) AI systems to make predictions, recommendations, and decisions. As we shall see, all three have significant epistemological consequences, and their intersection therefore provides for the possibility of novel instances of epistemic injustice.

It is instructive to contrast machine learning with a more familiar picture of classical programming, in line with what Haugeland (1985) called "Good Old Fashioned AI" or GOFAI. In this approach, humans provide both the program and the input data; it is the computer's job to apply the former to

the latter in order to generate an output. Perhaps the most significant general insight of the GOFAI approach is the conviction (and partial demonstration) that an enormous variety of tasks—logico-mathematical proof, puzzle-solving, chess, dyadic conversation—can be conceptualised and replicated using this input-processing-output schematic. By contrast, machine learning, as it were, turns this on its head; humans provide a data-set (which may or may not be labelled) and the AI system comes up with a set of rules (or mapping functions) which describe statistical patterns and correlations within the data. The intent is often that those rules can subsequently be used to make predictions or recommendations about future data points that were not in the training set, because the machine learning system may uncover correlations that were not already apparent to the human programmer or end-user. What's crucial to this process (and whence the approach derives its name) is the fact that the system itself can modify these rule-like mappings, as more data is acquired, in order to improve accuracy. The consequence, however, is a double-edged sword. Because the rules track patterns of which we were previously ignorant, machine learning's appeal (and one of its oft-cited advantages) is its ability to facilitate new discoveries. But in parallel, as the underlying mapping rules are incrementally modified in response to new data, and the complexity of the system thus increases, human developers can quickly lose track of an understanding of how the system actually works.[2]

Secondly, for machine learning systems to formulate and update their rules, a large set of training data is required. But it just so happens that we now live in an era where not only is computer hardware powerful and fast enough to process enormous quantities of data, but also human users—with every online click, view, "like," tag, or purchase—are willing to provide copious amounts of it for free. Smith (2019, p.52) for example, notes that the machine learning mentioned above only really became possible as a "post Big Data" development. This big data revolution of recent years is sometimes characterised as a dramatic expansion along the dimensions of the "three Vs" of velocity, volume, and variety; data is being collected faster and faster, in ever-increasing quantities, and in a bewildering variety of forms (text, images, audio, video etc.) The rise of interactive social media and the so-called "web 2.0"—with billions of tweets, posts, shares, and search queries—has generated extremely large data sets of details about users and their connections, on which machine learning systems can be trained to discover patterns, trends, and associations. When coupled with data that is collected and curated for specific purposes as well (e.g., the information divulged when seeking an insurance quotation), the patterns encoded in these rapidly-expanding data sets can subsequently be used to serve up better targeted adverts for products that you might buy, or to recommend movies or music or restaurants that you might like, and to make predictions about your creditworthiness or likelihood of committing a crime.

[2] See Walmsley (2021) for an overview of the consequences for "transparency."

The epistemic upshot of all this, at the intersection of machine learning and big data, is that we have built automated systems that work in ways we do not always fully understand, that are more powerful and much faster than the human mind, and which are trained on data sets that are too large for us to comprehend and process with the use of conventional tools (such as pocket calculators and spreadsheets). As a result, we often see even AI experts and insiders from the tech sector giving TED talks and writing op-ed pieces with click-bait headlines that make oblique reference to alchemy, sorcery, and various forms of "black box" mystery as the "Dark Secret at the Heart of AI."[3] It is tempting to cite Arthur C. Clarke's famous "third law"—*any sufficiently advanced technology is indistinguishable from magic*—and dismiss this sense of awe. But recent calls for transparency that respond to these epistemic limitations (e.g., in current GDPR legislation, or in the 2019 report from the European Commission's High-Level Expert Group on AI) become more pressing when one notes a third, somewhat less well-known, recent development in the field of Human-Computer interaction.

Logg et al. (2019) have investigated a phenomenon that they call "algorithm appreciation," whereby, in a variety of situations, people seem to prefer advice when they think that it is provided by a computer system, rather than by a human. This seems to mark a reversal in a tendency (first documented in the 1950s) towards *distrust* of the output of algorithms in comparison to human judgment (dubbed "algorithm aversion" by Dietvorst et al., 2015). In a series of experiments, Logg and her colleagues asked participants to make a variety of judgments (e.g., a numerical estimate concerning a visual stimulus, a prediction about the popularity of a song, and the likelihood of romantic attraction between two individuals) and found that people were significantly more likely to defer to advice about those judgments when they thought it came from an algorithm rather than from humans. Of course, in some cases, deferring to 'advice' from computers over and above that of fellow humans may be entirely appropriate; one would be correct to trust a computer to produce the correct answer to a complex mathematical calculation, since empirical data plausibly show that computers *are* more reliable on that front. But in Logg et al.'s studies, the output of a computer system seems to be preferred *simply because it is* (believed to be) *the result of an algorithm*. This point will be important below, when we come to discuss epistemic injustice (particularly of the testimonial variety). For now, suffice it to say that the work of Logg et al. seems to demonstrate an increasing willingness to trust automated systems even when we are not sure (or not told) how they work; it is also consistent with the many anecdotes of people who have ended up driving through fields and into rivers as a result of faithfully following the directions of a sat-nav or GPS system.

The implications of the considerations canvassed here are clear: humans are at an epistemic disadvantage when it comes to understanding the systems we have built in the fastest moving areas of contemporary AI. Of course, the history of technology—indeed, the history of science—contains

[3] See https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/

many examples where such opacity is not concerning in and of itself. A microscope or an automobile might be reliable enough for use in certain tasks even if we do not fully understand how they work; indeed, for a long time we did not understand how the human eye worked, but we still, as it were, used them for epistemic gain.[4] A crucial difference here is that although many people do not understand how technologies such as microscopes and automobiles work, there are nonetheless some experts (e.g., mechanics) who *do* understand them very well. In the case of machine learning, as we have seen, even many experts – i.e., the relevant computer scientists – claim not to fully understand how the systems generate the outputs that they do. So the opacity of machine learning systems *per se* may be a problem, especially if it means that in fact *nobody* knows why they are reliable, or under what conditions they may be unreliable.

Furthermore, the worry here is not only the opacity of machine learning systems *per* se, but also, their opacity *in conjunction with* their power and the trust we place in them. At the intersection of machine learning, big data, and algorithm appreciation, we have a situation where we do not fully understand AI systems, they're faster, more powerful, and more complex than us, but we increasingly trust them nonetheless.

3. Machine Gender Bias.

One consequence of the way that machine learning systems are developed is that they can sometimes end up, literally, encoding the more general societal prejudices that are represented in the big data set on which they are trained. Riffing on the expression "machine learning," some theorists have dubbed this "machine bias," although one must be careful to note that the most relevant parallel is something like *implicit* bias, rather than explicitly endorsed prejudice. For reasons that are perhaps unsurprising, such machine bias is often manifested along the familiar lines of race, class, gender, sexuality and (dis)ability and their intersections (see Eubanks, 2018, and Noble, 2018, for detailed surveys), but for present purposes, let us focus on two notorious recent examples concerning gender in particular, which arise precisely because of the technical features discussed in the previous section. In these cases, it is all too easy for AI systems to introduce both gendered *stereotypes* and the "default male bias" that, in the words of Criado Perez (2019), frequently renders women as "invisible" to such systems.

First, consider the news story that broke in 2018 about a sexist bias in a recruitment tool that was developed—and subsequently scrapped—by Amazon.[5] The goal had been to develop an AI system

[4] Thanks to an anonymous referee for reminding me of this point. I expand on it somewhat in Walmsley (2021), noting that complete transparency is often neither desired nor required of technology; we are usually happy to use such systems by adopting a kind of Dennettian intentional stance, as long as the system functions correctly, and the respects in which it is opaque are roughly neutral along ethical or political dimensions.

[5] See: https://www.bbc.com/news/technology-45809919

that could automatically scan the CVs of job applicants in order to come up with a short-list of the most promising candidates to be hired. However, the system was trained on historical data consisting of the CVs of many years' worth of previous applicants and (successful) employees; in a sector that is notoriously male-dominated, the majority of these were men. As a result, the system effectively taught itself to prefer male applicants, and to penalise CVs that contained keywords relating to women; the problematic male-domination of the industry in question was echoed and amplified by the automated recruitment process. As Gebru (2020) notes, "Analyzed within the context of the society it was built in, it is unsurprising that an automated hiring tool like Amazon's would exhibit these types of biases." This example of machine bias is not necessarily representative of any *explicit* prejudices of the human designers or recruiters (even if, evidently, they perhaps should have known better); rather, the system's undeniably sexist recommendations are an apparently unforeseen (and probably unintended) consequence of its training regime.

Relatedly, consider the way in which *Google Translate* handles languages in which the third-person singular pronoun is gender neutral, such as Turkish, Swahili, and Finnish.[6] Until fairly recently, when translating from such languages into English (in which the third-person singular pronoun *is* gendered), for example, the system tended to return answers that conform to stereotypical gender roles and characteristics (for example, the Turkish "O bir doktor" was translated as "he is a doctor", whereas "O bir hemşire" returned "she is a nurse"; "O bir mühendis" was translated as "he is an engineer", whereas "O bir aşçı" gave "She is a cook".) It's important to note that it is not so much that the AI system (or even its designer) displays an explicit prejudice; rather, the repetition of gender bias in the form of sexist stereotypes arises an unintended artefact of the way in which the system is trained. If *Google Translate* works by identifying and matching patterns found in millions of bilingual texts scraped from the internet, and these patterns (*in vivo* as it were) tend to exemplify more general societal prejudices about gender roles, characteristics and occupations, then it is not really surprising (but nonetheless problematic and disappointing) that the translation engine comes to replicate them.

In both of these cases, it is important to note that the problem was identified (and fixed) with human intervention: *Google Translate* now offers multiple translations from languages with gender neutral third-person pronouns, and Amazon has scrapped its automated recruitment tool. But these cases do highlight the problems that can arise—in this case concerning gender, but *mutatis mutandis* for race, class, (dis)ability etc.—when we defer to opaque but powerful AI systems.

4. Epistemic Injustice: Testimonial and Hermeneutical

[6] And, for a more general critique of gender bias in machine translation, see Bolukbasi et al., (2016)

The subtitle of Fricker's (2007) book *Epistemic Injustice* is "Ethics and the Power of Knowing" and therefore immediately suggests a connection between the foregoing considerations about the epistemology of AI, and a novel line of consideration for AI ethics. Given our epistemic limitations with respect to contemporary AI systems—and the problems to which these limitations can give rise—coupled with our tendency to defer to their advice and to trust them over and above other humans, I want to suggest that we run the risk of perpetuating novel instances of epistemic injustice.

Fricker defines epistemic injustice most generally as "… a kind of injustice in which someone is *wronged specifically in her capacity as a knower*."[7] She goes on to distinguish two specific types of epistemic injustice, on which I will focus; *testimonial* injustice and *hermeneutical* injustice.

In cases of testimonial injustice, "prejudice causes a hearer to give a deflated level of credibility to a speaker's word."[8] The idea is not so much that explicitly-endorsed prejudice will cause a hearer to *reject* a specific claim made by a speaker, but rather that, as Fricker puts it, "prejudice will tend surreptitiously to inflate or deflate the level of credibility afforded the speaker."[9] By way of example, Fricker mentions the case of a speaker's accent; this may affect the way in which a hearer assigns credibility insofar as the accent is perceived as carrying information about education, class, and geographic origin. Similarly, in Harper Lee's *To Kill a Mockingbird*, Tom Robinson's trial is a

> "…zero sum contest between the word of a black man against that of a white girl (or perhaps that of her father who has brought the case to court), and there are those on the jury for whom the idea that the black man is to be epistemologically trusted and the white girl distrusted is virtually a psychological impossibility."[10]

On the other hand, in cases of *hermeneutical* injustice, "a gap in collective interpretive resources puts someone at an unfair disadvantage when it comes to making sense of their social experiences."[11] This may be because a culture lacks a critical concept for a distinctive kind of social experience; Fricker cites the examples of *postpartum depression* and *sexual harassment* as cases whereby, prior to the formulation and dissemination of these terms, women suffered a "cognitive disadvantage from a gap in collective hermeneutical resources."[12] Alternatively, similar gaps in hermeneutical resources may inhibit or present a person from making their experiences communicatively intelligible; Fricker cites an example from Ian McEwan's 1997 novel *Enduring Love*, in which Joe Rose's experience of being stalked does not make the legal grade to "count" and he becomes depressingly aware that the official bureaucratic interrogative flowchart does not make room for his distinctive experience. He finds

[7] Fricker (2007) p.20
[8] *Ibid.,* p.1
[9] *Ibid*., p.17
[10] *Ibid.,* p.25
[11] *Ibid*., p.1
[12] *Ibid*., p.151

himself unable to make his distress (legally) intelligible, and is thus told that it is "not a police matter."

Prima facie, something like these two kinds of epistemic injustice may be present when we encounter the "computer says no" phenomenon with which I started. It may be that the output of an AI recommender system is assigned an inflated level of credibility relative to the human (because of "algorithm appreciation") and thus deferred to unjustly. Or it may be that the data points and input categories, on the basis of which the AI system makes its decision, do not contain the requisite level of detail or nuance to adequately capture all the relevant details of a person's lived experience. So the foregoing considerations invite an intervening question: When we willingly and preferentially rely on AI systems that even experts don't fully understand, and when those systems are more complex and powerful than us, do we run the risk of perpetuating epistemic injustices, either by trusting such systems over each other, or by using them in a way that inhibits our ability to understand or communicate our experiences?

5. AI and epistemic injustice.

Let us examine this question with respect to another recent controversial example: Babylon Health's *GP at Hand* service and app; a patient-facing digital symptom checker or a computerised diagnostic decision support programme.[13] In practice, the app/service has two components. First, first there is a "chatbot" symptom-checker designed to assist patients directly by creating differential diagnoses and advising on the need for further care. As per Babylon Health's own published research (Baker et al., 2020), it is explicitly intended as a triage system. Second, there is a component that puts patients face-to-face with human GPs via a videoconferencing service, depending on the outcome of the symptom-checking triage process.

According to a BBC documentary which aired in November 2018 as part of the *Horizon* series,[14] in 2017 Babylon Health partnered with a physical GP surgery such that any adult in the greater London areas could switch their NHS GP to Babylon's surgery. It quickly became the fastest-growing GP practice in the UK, with over 30,000 patients registered (most of whom had never visited the physical location). The potential benefits—24/7 access, reduced waiting time, reduced costs to the NHS—are clear, and often mentioned as part of the publicity and promotion.

The *GP at Hand* service does raise a significant number of "conventional" ethical concerns. For one thing, Babylon Health was criticised by the UK's Medicines and Healthcare products Regulatory

[13] See https://www.gpathand.nhs.uk/ . For Babylon's own published information about the system, see Baker et al. (2020) and for a rebuttal, see Fraser et al. (2018).
[14] https://www.bbc.co.uk/programmes/b0bqjq0q

Agency regarding the promotion of the app and its safety[15]; although the app does have an MHRA classification, it is only in the category alongside Zimmer frames and spectacles, in which manufacturers self-declare the safety and effectiveness of their product. For another, Babylon Health selected Rwanda as a location for the app's initial trial-run; even though that did eventuate in a full roll-out that increased access to GP healthcare in Rwanda, there is an air of colonialism to the practice of conducting a trial run in a developing country, especially if it was done on the estimation that there would be less backlash or legal/financial fallout if something went wrong. Finally, the app could well encounter similar problems to Amazon's automated hiring tool and *Google Translate*, depending on its training data set; one can think of many examples of gender inequality in the healthcare system (e.g., the naming and classification of "hysteria", or the systematic neglect of research on, diagnosis of, and treatment for endometriosis[16]) that ought to give one pause before training an AI system on such a history.

In addition, however, the *GP at Hand* service also raises the possibility of novel instances of the kinds of epistemic injustice mentioned in the previous section. First, in the BBC's *Horizon* documentary, Ali Parsa (CEO of Babylon Health) is quite explicit in framing the company's goal as a comparison or competition between human GPs and AI, with a hint about the possibility of the latter replacing the former. He says:

> "We're saying that we are going to create a machine that can eventually hopefully tell you about your disease as accurately as a doctor can, that can give you your treatment better than most doctors can, that can look faster, cheaper, at variations of symptoms, billions of variations of symptoms, as well as if you had one of the best doctors in the world in your pocket."

This is in keeping with Babylon Health's own attempts to demonstrate the safety and reliability (indeed, *superiority*) of their app: the matter is always framed as a *comparative* question of diagnostic success *in competition against* Human GPs (e.g., in Baker et al., 2020); even critical rebuttals (e.g., Fraser et al., 2018) unwittingly default to the position that the right way to proceed is to put the AI and human performance into a zero-sum head-to-head contest. Of course, we should note that we often *do* assign different levels of credibility to different sources of information in contexts such as this. Staying within medicine, for example, we might note that humans in the field *are* hierarchically organised, and we routinely assign greater levels of credibility to a fully credentialed Attending Physician over, say, a Junior Resident. But we also suppose that the credibility assignments track relevant expertise, and therefore the benefits of doing so outweigh the risk of perpetuating testimonial injustices against Junior Residents.

[15] See here: https://tcrn.ch/3MASMim
[16] See Jones (2016) and Young, Fisher, and Kirkman (2020).

A key question, therefore, is whether the benefits of using an AI system such as *GP at Hand* outweigh the risks of perpetuating testimonial injustices. And in part, that depends on the empirical question of how much testimonial injustice is *already* perpetrated by us against our fellow humans. But the literature discussed above on "algorithm appreciation" (Logg et al., 2019) suggests that the inflated level of credibility assigned to computer systems does not arise subsequent to an estimate of their reliability or "expertise"; rather, it seems that we trust computers more *simply because they are computers*. Accordingly, if the same can be said of any apparent trust placed in in *GP at Hand* we run the real risk of committing a kind of testimonial injustice with respect to human GPs. If we came to believe the hype, we would assign a deflated level of credibility to a human GP's judgment *because they are human*, and an inflated level of credibility—what Fricker calls a *credibility excess*—to the app, and these assignments occur because of the groups to which their targets belong, and not because of any justifiable appraisal of their relative expertise. For instance, to return to a previous example, a machine learning system might have been trained on a body of medical data which has been extensively criticised for its marginalisation and neglect of a condition such as endometriosis, and thus would be less likely to offer such a diagnosis (and appropriate treatment recommendations) than a human clinician trained in the relevant specialty. But if, in virtue of "algorithm appreciation," the system were afforded a credibility excess nonetheless, not only is misdiagnosis likely, but also a testimonial injustice would be committed with respect to the human doctor.

Similarly, it may well be that reliance on the app could prevent or inhibit a user (i.e., a patient) from accessing the interpretive resources that they need in order to understand or communicate their own experience, and thus give rise to a hermeneutical injustice. On the one hand, a lot of detail about the app's scope and limits is contained in (one might say "buried in") the Terms and Conditions and the End-User Licence Agreement. This is where the user is reminded that the app is a *symptom-checker*, and not providing a *diagnosis* or anything that should be construed as analogous to a human GP's advice. But, notoriously, app users simply do not read such Ts&Cs and EULAs, and so that fact that such significant information is only included there may present a barrier to understanding what one is doing when one uses the app. On the other hand, it is telling that the chatbot is effectively shunting the user through a branching decision tree—much like the bureaucratic interrogative flowchart in Ian McEwan's *Enduring Love*, as mentioned above—such that one's report of one's symptoms, and their nuanced *phenomenology*, can be pattern-matched to one of a specific number of output categories. Practically speaking, this may also give rise to a worry about outcomes: it is well known, for example, that a patient's engaged involvement—imbuing them with a sense of both understanding and agency—is significantly correlated with both better diagnosis *and* improved prognosis. Further, returning to the example of endometriosis, Young, Fisher and Kirkman (2020, p.35) report women's wariness of the "power of doctors to reduce their wellbeing through medical labels they did not identify with". A machine learning system that had been trained on a body of data that already

neglects endometriosis, and consequently applied even less nuanced and accurate "labels," would surely exacerbate this concern; in virtue of limiting the interpretive and communicative resources available to patients, a kind of hermeneutic injustice would have been committed.

In sum: independently of the questions about the safety, success, and accuracy of Babylon Health's *GP at Hand* app, and its development, it *also* looks like it has the potential to give rise to additional *epistemic* injustices, either with respect to the GPs against which it is compared, and over whom it is privileged, or with respect to the patients it is supposed to serve, when it either inhibits an understanding or a communication of their experience.

6. A Collaborative Response?

The foregoing considerations have focussed primarily on problems and concerns—the doom and gloom, as it were—but they also suggest some paths forward. By way of conclusion, rather than prescribing some very general policies and practices that we might try to follow when considering the use of AI (e.g., the top-down, Asimov-style "Don't use AI in contexts that could give rise to epistemic injustice"), instead, I want to focus on some conceptual tools for doing so, and an approach that emphasises *collaboration* between humans and AI.

First, we might invoke a distinction—from Smith (2019)—concerning the sorts of tasks to which AI can be put. As the subtitle of his book suggests, when considering what tasks to hand over to AI systems, we need to distinguish between the capacities of *judgment* and *reckoning*. The sense of "judgment" here is roughly analogous to what we mean when we describe someone as having "good judgment"—Smith describes it as:

> "…a form of dispassionate, deliberative thought, grounded in ethical commitment and responsible action, appropriate to the situation in which it is deployed…[It] is the standard to which human thinking must ultimately aspire."[17]

By contrast, Smith uses the term "reckoning" for

> "…the types of calculative prowess at which computer and AI systems already excel—skills of extraordinary utility and importance, on which there is every reason to suppose computers will continue to advance (ultimately far surpassing us in many cases, where they do not do so already), but skills embodied in devices that lack the ethical commitment, deep contextual awareness, and ontological sensitivity of judgment."[18]

In the examples of epistemic injustice discussed above, the problems arise either when (a) we privilege a reckoning system in a case that requires genuine judgment, such that the human in question suffers a testimonial injustice because of a credibility deficit, or (b) in virtue of being unduly

[17] Smith (2019) p.xv
[18] *Ibid*., p.xvii

impressed by the prowess of a reckoning system, we attempt to shift human reasoning in that direction, such that a hermeneutic injustice may be perpetuated.

Smith's point here is *not* that judgment is genuine intelligence, whereas reckoning is "merely mechanical"—in fact, *both* reckoning and judgment are kinds of intelligence. But as Smith points out in an interview elsewhere,[19] the deeper question is what kinds of *tasks* require each of these capacities. Some cases are obvious; calculating the decimal expansion of π is clearly a reckoning task, whereas teaching ethics to school children is a matter of judgment. But in-between are the tricky cases that arise in the examples discussed above. Making a diagnosis on the basis of an X-ray may be best accomplished, initially, with the image classification capacities of a reckoning-based AI system. But a compassionate treatment plan requires a human GP to exercise their judgment in interpreting the consequences for a patient's lived experience.

Notice, then, that the confusion of reckoning and judgment, and the subsequent way in which it brings about forms of epistemic justice, largely arises because AI systems are frequently seen as *competitors* with humans, or *replacements* for human decision-making. In a recent tweet, Geoffrey Hinton (a prominent computer scientist, and one of the early pioneers of AI machine learning techniques) wrote "Suppose you have cancer and you have to choose between a black box AI surgeon that cannot explain how it works but has a 90% cure rate and a human surgeon with an 80% cure rate. Do you want the AI surgeon to be illegal?"[20] But this is a false dilemma. An alternative—one that is in fact suggested by Babylon Health's official description of the *GP at Hand* system as a diagnostic *decision support* program—would be to see the issue as one of potential *collaboration* between AI and human experts, and to incorporate insights from the study of group agency, joint action, and collective decision-making into our account of human-computer interaction.

Such an approach has already been taken in some other areas of AI ethics. Nyholm (2018), for example, argues that worries about "responsibility gaps" can be overcome, when it comes to autonomous vehicles and autonomous weapons systems, if we understand them as standing in a collaborative relationship with the humans involved in their development, use, and deployment. Similarly, Gunkel (2018), in addressing the controversial question of whether AI systems should have ethical standing as moral *patients* (i.e., the question of "Robot Rights") suggests prioritizing human-AI social *relationships* and interactions, over and above pre-determined ontological criteria or capabilities. Nyholm's and Gunkel's emphasis on this kind of "relational turn" in AI ethics might be fruitfully extended, as I have suggested, to the kinds of collaboration required by the decision-making and recommender systems discussed above, particularly when there is a "human-in-the-loop" and when doing so can give a stronger sense of agency to the human involved.

[19] https://ischool.utoronto.ca/news/talking-ai-and-humanity-with-brian-cantwell-smith/
[20] https://twitter.com/geoffreyhinton/status/1230592238490615816 (February 20, 2020)

It's important to note that collaboration need not be an *equal* partnership between the two (or more) collaborators; as one may remember from group projects at school (or from academic co-authorships), collaboration permits a spectrum of lop-sidedness depending on the task at hand. As suggested by the various case studies that Nyholm (2018) discusses, this could involve the human playing a mostly supervisory or directorial role, or being poised to step in if it should be needed (the "human-in-the-loop" scenario), or collaborating mostly at the level of hardware/software updates.[21] Similarly, the collaboration—especially in the case of reinforcement-based forms of machine learning—might involve the human providing feedback, especially when contestability is being used as an supplement or alternative to transparency (see Walmsley, 2021). Thus, if it turned out that reliability differed in correlation with the degree of contribution made by each party in the collaboration, whether human or AI, those proportions could be empirically investigated and adjusted accordingly.

The advantage of such a collaborative model—in keeping with Smith's distinction between reckoning and judgment—is therefore threefold. First, it would allow us to assign to each partner in the collaboration the kind of reckoning or judgment tasks at which they excel (especially given that, as noted, collaborations can be lopsided and flexible). Second, it might, as Smith suggests, "up the ante" on people; by leaving reckoning tasks to computers, we might raise the standards on human judgment, with all the ethical commitment that includes. And finally, it may help to avoid (or reduce the risk of) the kinds of epistemic injustice—testimonial and hermeneutic—that I have outlined throughout this paper.

It's not as catchy as *Little Britain's* "Computer Says No!", but it would be nice to end up with something along the lines of "Computer says 'Let's work on this together'."

-.-.-

[21] I suppose there *is* a sense (degenerate, or perhaps a limit case), also, in which it would be a kind of "collaboration" for the human simply to step back entirely from a given decision or process, but this would threaten to trivialise the proposal under consideration. I'm grateful to an anonymous referee for highlighting this issue.

References.

Adam, Alison. 1998. *Artificial Knowing: Gender and the Thinking Machine*. Routledge.

Baker, Adam, Perov, Yura, Middleton, Katherine, Baxter, Janie, Mullarkey, Daniel, Sangar, Davinder, Butt, Mobasher, DoRosario, Arnold, and Johri, Saurabh. 2020. "A Comparison of Artificial Intelligence and Human Doctors for the Purpose of Triage and Diagnosis." *Frontiers in Artificial Intelligence* 3:1–9.

Bolukbasi, Tolga, Chang, Kai-Wei, Zou, James, Saligrama, Venkatesh, and Kalai, Adam. 2016. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 4356–4364. Red Hook, NY, USA: Curran Associates Inc.

Criado Perez, Caroline. 2019. *Invisible Women: Data Bias in a World Designed for Men*. New York: Abrams Press.

Danaher, John. 2019. *Automation and Utopia: Human Flourishing in an Age Without Work*. Cambridge, MA: Harvard University Press.

Dietvorst, Berkeley J., Simmons, Joseph P., and Massey, Cade. 2015. "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err." *Journal of Experimental Psychology: General* 144:114–126.

Eubanks, Virginia. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish The Poor*. New York: St. Martin's Press.

European Commission, Directorate-General for Communications Networks, Content and Technology, (2019). *Ethics guidelines for trustworthy AI*, Publications Office. https://data.europa.eu/doi/10.2759/177365

Fraser, Hamish, Coiera, Enrico, and Wong, David. 2018. "Safety of patient facing digital symptom checkers." *The Lancet* 392:2263 – 2264.

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

Gebru, Timnit. 2020. "Race and Gender." In Markus D. Dubber, Frank Pasquale, and Sunit Das (eds.), *The Oxford Handbook of Ethics of AI*, chapter 13. Oxford University Press.

Gunkel, David J. (2018). The other question: can and should robots have rights? *Ethics and Information Technology* 20 (2):87-99.

Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Jones, Cara E. (2016). The Pain of Endo Existence: Toward a Feminist Disability Studies Reading of Endometriosis. *Hypatia*, 31(3), 554–571.

Kember, Sarah. 2003. *Cyberfeminism and Artificial Life*. Routledge.

Logg, Jennifer M., Minson, Julia A., and Moore, Don A. 2019. "Algorithm appreciation: People prefer algorithmic to human judgment." *Organizational Behavior and Human Decision Processes* 151:90–103.

Noble, Safia Umoja. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

Nyholm, Sven (2018). Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci. *Science and Engineering Ethics* 24 (4):1201-1219.

Smith, Brian Cantwell. (2019) *The Promise of Artificial Intelligence: Reckoning and Judgment*. MIT Press.

Walmsley, Joel. (2021) "Artificial intelligence and the value of transparency." *AI & Society* .

Young, Kate, Fisher, Jane, and Kirkman, Maggie. (2020) "Partners instead of patients: Women negotiating power and knowledge within medical encounters for endometriosis." *Feminism & Psychology* 30(1):22-41.