The Eliza Effect and Its Dangers: From Demystification to Gender Critique

Sarah Dillon

Faculty of English, University of Cambridge

sjd27@cam.ac.uk

Biographical Note: Sarah Dillon is an interdisciplinary feminist scholar of contemporary literature, film and philosophy in the Faculty of English at the University of Cambridge, UK. She is author of *The Palimpsest: Literature, Criticism, Theory* (2007), *Deconstruction, Feminism, Film* (2018) and *Storylistening: Narrative Evidence and Public Reasoning* (forthcoming 2021, with Claire Craig). She has edited *David Mitchell: Critical Essays* (2011), and co-edited *Maggie Gee: Critical Essays* (2015), *AI Narratives: A History of Imaginative Thinking About Intelligent Machines* (2020) and *Imagining Derrida* (2017), a special issue of *Derrida Today*. She has published reports on *Portrayals and Perceptions of AI and Why They Matter*, and *AI and Gender: 5 Proposals for Future Research*, and in 2020-21 is co-Principal Investigator of a University of Cambridge Mellon Sawyer Seminar, Histories of Artificial Intelligence: A Genealogy of Power. She is General Editor of the Gylphi Contemporary Writers: Critical Essays book series, and also works regularly as an arts broadcaster for BBC Radio 3 and BBC Radio 4.

Abstract:

This essay provides a gender critique of the Eliza effect. It delineates the way in which the Eliza effect is operationalised in AI research even as it is ostensibly demystified, for example in the writings of Douglas Hofstadter and Joseph Weizenbaum. It then exposes the gendered assumptions embedded in the nomenclature used to name this misperception of the computer as having capabilities equivalent to the human. It traces the genealogy of that nomenclature back through Weizenbaum's ELIZA, to George Bernard Shaw's *Pygmalion*. A close reading of the play is deployed in order to reveal the structural inequities of gender, class, and who or what gets to be human, that are both explored in the play and encoded in the operation and operationalisations of the Eliza effect. It concludes by attending to that operation and operationalisation in relation to today's Virtual Personal Assistant's, and makes a case for the importance of critique in order to expose the inequitable structures of power obscured and compounded by the Eliza effect – both its name, and that which it names.

Keywords:

ELIZA, Weizenbaum, Shaw, *Pygmalion*, gender, artificial intelligence

Introduction

In "The Ineradicable Eliza Effect and Its Dangers," Preface 4 of *Fluid Concepts and Creative Analogies: Computer Models and the Fundamental Mechanisms of Thought* (1995), Douglas Hofstadter defines the Eliza effect. He identifies the evocative language and rich visual imagery researchers use to present their computer programs and their capabilities. Such

language and imagery, combined with uncritical media coverage and publicity, serve to create a misperception in the public that such programs have the same capacities as human beings:

> a host of implications follow in the minds of many if not most readers, such as these: […] computers understand the physical world; computers make analogies; computers reason abstractly; computers make scientific discoveries; computers are insightful cohabiters of the world with us.
>
> This type of illusion is generally known as the "Eliza effect", which could be defined as the susceptibility of people to read far more understanding than is warranted into strings of symbols – especially words – strung together by computers. (p. 157)

Hofstadter claims that his identification of the Eliza effects is not a criticism of the developers or the journalist he uses in his opening example, but "a critique of the whole mentality swirling around the complex intellectual endeavor called "AI" – a surprisingly unguarded mentality in which anthropomorphic characterizations of what computers do are accepted far too easily, both outside and inside the field" (p.158).

Hofstader clearly believes the Eliza effect's dangers to be a misrepresentation of the capacities and capabilities of computer programs. But there are profound tensions, and elisions, in his claimed critique. Despite offering further examples of where "cognitive-science professionals seem unable or unwilling to distinguish between what some program has done and what people do, provided there is some minimal degree of surface-level resemblance" (p.161), Hofstadter repeatedly insists that the Eliza effect is an accidental consequence of such research. He describes it as the result of an "overly charitable way of characterizing what has happened" (p. 157) and claims that "of course it is inadvertent rather than deliberate distortion" (p. 157). In the final section of the Preface, Hofstadter acknowledges that the Eliza effect is a

form of "hype," but attempts to retain his insistence that it is "inadvertent" (p. 167). It is, of course, no such thing, as his next sentence admits:

> Clearly, *all* AI researchers, myself included, want to brag about their programs' achievements; on the other hand, we all know that we can't get away with out-and-out anthropomorphism. What generally results is some kind of intermediate level of description, in which a bit of caution is used but much is left ambiguous, so that readers are still free to draw conclusions that often will amount to some kind of Eliza effect – benefiting the researchers, needless to say. (p. 167)

The Eliza effect is a form of hype intentionally operationalised by AI researchers in order to deliberately overrepresent the capabilities of their research. It is neither accidental, nor inadvertent.

Hofstadter's so-called critique of the Eliza effect and its dangers is compromised by his own participation in its operationalisation. Hofstadter is an AI researcher actively engaged in AI research and thereby also engaged in a contest over ways of pursuing AI. His interests, here especially, exceed those of objective scientific pursuit of a technical goal. The pursuit of AI has long entailed epistemic and technical conflicts over what AI is, as well as the best ways to pursue it. These arguments are comprised of rhetorics and polemics, recurring argumentative tropes, as well as the complex discursive structures of scientific research itself. In ostensibly exposing one particular form of hype – the Eliza effect – Hofstadter both acknowledge his own operationalisation of that effect, and deploys another form of hype – that of demystification.[1] The hype of anthropomorphism is long-standing, so also is the counter-claim of demystification. Hofstadter therefore does what he claims that others do. He simply does it at a different scale: "It may well be that in this book, precisely this kind of thing takes place in

our discussions of our own work, but there is one difference," he asserts, "our domains are deliberately so stripped-down that the claims made *cannot* be very grandiose." (p. 167) In proposing to demystify competing research, Hofstadter situates his own work as less extra-scientific, and positions his writing as more scientifically objective than that of the writing he is contesting.

This essay takes the nomenclature of the Eliza effect seriously in order to move away from the hype of both its functioning and its demystification. In their place, it constructs a critique that focuses on how relations of power – in particular as they pertain to gender, labour and class – are encoded within the Eliza effect's operations and operationalisations. The essay traces the dual cultural origins of the Eliza effect – both its name, and the phenomenon it describes – in both science and literature. The phenomenon was first explored at length in the work of computer scientist Joseph Weizenbaum. In particular, he was prompted to such reflections by the public response to ELIZA, his natural language processing software program. Although Weizenbaum wishes to interrogate the Eliza effect beyond its effects for science, by attending to the social consequences of such technologies, in his nomenclature and rhetoric he is in fact operationalises it. Moreover, Weizenbaum's acts of gendering unreflectingly reproduce the objectification of women and the gendered division of labour which the source for his choice of name, George Bernard Shaw's play *Pygmalion* (1913), is committed to interrogating – Weizenbaum's program is named after the character Liza Doolittle in Shaw's play. Weizenbaum fails to learn from the feminist insights *Pygmalion* provides regarding the gendered nature of labour, the objectification of women, and the sociogenic status of the category of the human, in particular the struggle over its inclusions and exclusions. The essay concludes by delineating how the Eliza effect is deployed in relation to ELIZA's heirs, today's Virtual Personal Assistants (VAPs), and makes the case for the importance of critique in

exposing the inequitable structures of power in and around AI that phenomena such as the Eliza effect serve to obscure and compound.

Weizenbaum's ELIZA

Weizenbaum's essay, "ELIZA – A Computer Program For the Study of Natural Language Communication Between Man and Machine," (1966) exhibits the same tension found in Hofstadter's preface. Weizenbaum's claim to demystify his program is undermined by the repeated operationalisation of the Eliza effect in his rhetoric and acts of gendering. In his extra-scientific preface, Weizenbaum (1966) explains his motivation for writing the paper:

> It is said that to explain is to explain away. This maxim is nowhere so well fulfilled as in the area of computer programming, especially in what is called heuristic programming and artificial intelligence. For in those realms machines are made to dazzle even the most experienced observer. But once a particular program is unmasked, once its inner workings are explained in language sufficiently plain to induce understanding, its magic crumbles away. (p. 36)

Weizenbaum intends to explain how ELIZA works, in order to explain away the "intelligence" that has been undeservedly assigned to it. He therefore defines ELIZA straightforwardly as "a program which makes natural language conversation with a computer possible," (Weizenbaum 1966, p. 36) and provides details of the system it operates on and of the computer language in which it is written. Weizenbaum's commitment to demystifying ELIZA is compromised, however, by the fact that he gives the program a human name – an act of nomenclature that necessarily encourages rather than discourages the Eliza effect.

Without explicitly acknowledging the consequences of this contrary act, Weizenbaum does attempt to provide a logical explanation for the choice of name: "Its name was chosen to emphasize that it may be incrementally improved by its users, since its language abilities may be continually improved by a 'teacher'." (p. 36)[2] This explanation does little to diffuse the Eliza effect. The explanation for the choice of name enacts the anthropomorphisation implicit in the effect itself. Weizenbaum likens the program to a student and the user to a teacher, with the program's development process thereby becoming one of "learning." The use of the concept of "learning" to name developments that are not explicitly programmed is found in D. R. Hartree's *Calculating Instruments and Machines* (1949, p. 70) who is cited by Alan Turing in "Computing Machinery and Intelligence" (1950), where the latter develops the idea of "learning machines" in Section 7. It is now an accepted, and often uninterrogated, part of contemporary AI discourse, naming as it does the specific field and technologies of "machine learning."[3] But the term plays its role in encouraging the Eliza effect, and might be considered part of the "wishful mnemonics" (p. 4) of AI research that Drew McDermott (1976) criticises for contributing to the misrepresentation of AI programs and their capacities: "We have lived so long with the conviction that robots are possible, even just around the corner, that we can't help hastening their arrival with magic incantations." (p. 5)[4]

In a paper ostensibly committed to diffusing it, Weizenbaum in fact reinforces the Eliza effect through his use of simile and diction: "Like Eliza of Pygmalion fame," he continues, ELIZA "can be made to appear even more civilized." (p. 36) The problem here is not just again the explicit comparison between a computer program and a human character (ELIZA and Eliza), nor just the wishful mnemonics (the idea that a computer program might be more or less "civilised"), but the implicit acknowledgement that one of the goals of the program developers might in fact be to improve how human-like the program might "appear" to its users. This goal – constitutive of an operationalisation of the Eliza effect – is in sharp contrast

to Weizenbaum's explicit dismay (explored at more length ten years later in *Computer Power and Human Reason* (1976)) at the reception of ELIZA and how willingly and easily humans interacting with it were persuaded of its "reality." In his act of naming ELIZA and his rhetoric around it, Weizenbaum is thus complicit in deploying the Eliza effect, even as he is ostensibly committed to demystifying it.

In her discussion of Weizenbaum's program, Margaret A. Boden (1987, pp. 106-111) consistently refers to ELIZA as "she," rather than "it," undermining her own arguments through her choice of pronoun.[5] Whilst Weizenbaum employs the gender neutral pronoun, "it", his other acts of gendering are no less problematic. In *Computer Power and Human Reason* (1987), Weizenbaum observes that he was "startled to see how quickly and how very deeply people conversing with DOCTOR became emotionally involved with the computer and how unequivocally they anthropomorphized it." (p. 7) Note here the change in nomenclature. Whilst the program is called ELIZA, in the example conversation provided in the 1966 paper the user is female and the computer becomes gendered as male, referred to as DOCTOR:[6]

Men are all alike

IN WHAT WAY

They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here.

YOUR BOYFRIEND MADE YOU COME HERE

He says I'm depressed much of the time.

I AM SORRY TO HEAR YOU ARE DEPRESSED

It's true. I am unhappy.

DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY

I need some help, that much seems certain.

WHAT WOULD IT MEAN TO YOU IF YOU GOT SOME HELP

Perhaps I could learn to get along with my mother.

TELL ME MORE ABOUT YOUR FAMILY

My mother takes care of me.

WHO ELSE IN YOUR FAMILY TAKES CARE OF YOU

My father.

YOUR FATHER

You are like my father in some ways.

WHAT RESEMBLANCE DO YOU SEE

You are not very aggressive but I think you don't want me to notice that.

WHAT MAKES YOU THINK THAT I AM NOT VERY AGGRESSIVE

You don't argue with me.

WHY DO YOU THINK I DON'T ARGUE WITH YOU

You are afraid of me.

DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU

My father is afraid of everybody.

WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER

Bullies.

DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

Weizenbaum genders the program as female when it is under the control of the male computer programmer, but it is gendered as male when it interacts with a user. Note in particular that in the example conversation given, this is a disempowered female user, at the mercy of her

boyfriend's wishes and her father's bullying, defined by and in her relationship to the men whom, she declares, "are all alike." (Weizenbaum, 1966, p. 36) Weizenbaum's choice of names is therefore adapted and adjusted to ensure that the passive, weaker or more subservient position at any one time is always gendered as female, whether that is the female-gendered computer program controlled by its designers, or the female-gendered human woman controlled by the patriarchal figures in her life, both human (boyfriend and father) and perhaps even non-human (DOCTOR).

It is unsurprising perhaps, then, that Weizenbaum's paradigmatic naïve dupe of the Eliza effect is in a female-gendered subservient role – his secretary. In *Computer Power and Human Reason* he recounts a now famous anecdote:

> Once my secretary, who had watched me work on the program for many months and therefore surely knew it to be merely a computer program, started conversing with it. After only a few interchanges with it, she asked me to leave the room. (Weizenbaum, 1976, p. 7)

Weizenbaum uses this anecdote as an example of the dangers of the Eliza effect, that "delusional thinking" which is brought on in "quite normal people" even by only "extremely short exposures to a relatively simple computer program" (p. 7). However, it also serves to draw attention to the historically gendered nature of labour – man as dominant creator, woman as subservient assistant – which underlies both his gendered naming of ELIZA and, as we will see, the female gendering of that program's heirs, today's VPAs. Whilst Weizenbaum acknowledges that the task of exploring "the relation of appearance to reality" remains "in the domain of the playwright" (Weizenbaum, 1966, p. 36), he fails to recognise the importance of

Pygmalion's feminist exploration of the gendered dynamics which his own acts of gendering perpetuate, rather than interrogate.[7]

Shaw's Liza

The plot of Shaw's play is formed around an experiment which is both scientific and social in nature. After a chance meeting between the three main characters whilst sheltering from a downpour in Convent Garden, Colonel Pickering challenges Dr Higgins (a professor of phonetics) to successfully pass off Liza Doolittle (an East End flower girl) as an upper class English lady. Higgins is persuaded to undertake the experiment when Pickering turns it into a bet – if Higgins succeeds, Pickering commits to covering all the costs of the experiment (p. 29). Higgins accepts, declaring, "I shall make a duchess of this draggletailed guttersnipe. […] Yes: in six months – in three if she has a good ear and a quick tongue – I'll take her anywhere and pass her off as anything." (p. 29) Whilst the transformation of Liza's physical appearance is effected with swiftness and ease by Higgins' female housekeeper, Mrs Pearce, the more challenging transformation of her speech is the terrain of Higgins and his scientific expertise in phonetics. The gendered division of labour is thereby signalled early on in the play, both in relation to the dynamics of Higgins' home, and the delivery of the experiment.

Higgins does not consider that this experiment would change Liza in any fundamental way. There is no sense in which he considers she would actually *become* a lady. Rather, the goal is simply to alter her appearance and speech to the extent that she can be passed off as one. Higgins does not countenance the possibility of a change in Liza's class status, because he does not in fact view or treat her as fully human, with the full rights, autonomy, and agency allowed to the human. Rather, he treats her only as an object to be moulded and serve his needs as he sees fit. The play is actually much more interested in the tension between Higgins' objectification of Liza and her assertion of her own humanity, than in the plot of the

experiment. Whilst Liza's speech is improved, she is successfully passed off as a Lady, and Higgins wins the bet, this sequence of events is not the focus of the play's main action. In fact, a "Note for Technicians" prefacing the play explains that "for ordinary theatrical use the scenes separated by rows of asterisks are to be omitted." (p. 8) The reason given is the limits of theatrical versus cinematic production, but the omitted scenes are those that either foreground women's work or Liza's autonomy (the scene of Liza's physical transformation, and her engagement to Freddy), and, perhaps surprisingly, those which pertain most essentially to the experiment (Liza's lessons with Higgins, and the scene depicting Liza's successful social passing). Instead, the play in fact focuses on the consequences of treating women as objects, revealing it be part of the broader policing of the boundaries of the category of the human instituted in and by modernity. For the category of the human is sociogenic, that is, it is a socially produced designation, not a static, given or immutable ontological category.[8] In Sylvia Wynter's (2003, p. 260) account, this mutability is what institutes, for those with vested interests, "the ongoing imperative of securing the well-being of [the] present ethnoclass (i.e. Western bourgeois) conception of the human, Man, which overrepresents itself as if it were the human itself." (p. 260) In the face of this 'Coloniality of Being/Power/Truth/Freedom," (p. 260) the struggle is "that of securing the well-being, and therefore the full cognitive and behavioural autonomy of the human species itself/ourselves." (p. 260) "Struggles with respect to race, class, gender, sexual orientation, ethnicity" (p. 260) can all be understood, for Wynter, as "differing facets of the central ethnoclass Man vs. Human struggle" (p. 261). *Pygamlion* articulates and interrogates this struggle, in particular with regard to gender and class. But this contribution of the play is entirely neglected by Weizenbaum, who excerpts Liza's name only as a metonym for how a program might learn and improve. In doing so, Weizenbaum fails to himself learn from the social justice investigations of Shaw's play.[9]

Higgins and Pickering's decision to undertake their experiment is made independently of Liza's choice in the matter. Pickering, to his credit, draws Higgins' attention to this omission, but Higgins disregards the idea that Liza might be considered a full human subject:

> PICKERING [*in good-humored remonstrance*] Does it occur to you, Higgins, that the girl has some feelings?
>
> HIGGINS [*looking critically at her*] Oh no, I don't think so. Not any feelings that we need bother about. [*Cheerily*] Have you, Eliza?
>
> LIZA. I got my feelings same as anyone else.
>
> HIGGINS. [*to Pickering, reflectively*] You see the difficulty?
>
> PICKERING. Eh? What difficulty?
>
> HIGGINS. To get her to talk grammar. The mere pronunciation is easy enough. (p. 32)

The only obstacle to the success of Higgins' experiment is not the female subject's autonomy, but the technical problem of how to correct her grammar, as well as her pronunciation. At the outset of the play, Liza is thus treated as an object that the men may do with as they wish. By the end of Act Three, she has become a very specific type of manufactured artefact – Higgins' personal assistant, whose labour is essential but unseen. He explains to his mother that "she's useful. She knows where my things are, and remembers my appointments and so forth" (p. 65); at the opening of Act Four, she fetches his slippers without him even noticing (p. 74), and he issues instructions to her without a second thought: "Put out the lights, ELIZA; and tell Mrs Pearce not to make coffee for me in the morning. I'll take tea." (p. 75)

Even when Liza's transformation leads her to resist this objectification and to assert her human rights, Higgins refuses to accept her autonomy and independence. He insists on her

status as artefact and his status as creator – the manufacturer of both her voice and the ideas it conveys:

> HIGGINS. Let her speak for herself. You will jolly soon see whether she has an idea that I havnt put into her head or a word that I havnt put into her mouth. I tell you I have created *this thing* out of the squashed cabbage leaves of Covent Garden; and now she pretends to play the fine lady with me. (p. 93-4) [added emphasis]

Liza is represented by Higgins as a fabricated object – "this thing" – over which he, as creator, insists on retaining dominion. In fact, the verb "to make," in its various conjugations, reverberates in the final scene of the play. When Liza challenges Higgins as to the ethics of his experiment and its social consequences – "You never thought of the trouble it would make for me" (p. 101) – he insists on his right to professional autonomy, irrespective of those consequences: "Would the world ever have been made if its maker had been afraid of making trouble? Making life means making trouble." (p. 101) The verb "to make" recurs again in Liza and Higgins' discussion of her potential future marriage to Freddy: "Can he *make* anything of you? That's the point," challenges Higgins; "Perhaps I could make something of him. But I never thought of us making anything of one another; and you never think of anything else. I only want to be natural" (p. 102) replies Liza. Liza rejects Higgins' understanding that male interaction with women is characterised only by considering how they can be shaped. Her understanding of relationality between the genders is not just one of reciprocal shaping, but of an absence of fabrication entirely, an eradication of creator and created, and of manufacture. In contrast, whilst Higgins takes some pleasure in bold, defiant Liza, he considers even that version of her to be his creation: "By George, Eliza, I said I'd make a woman of you; and I have. I like you like this" (p. 102). His fantasy that she, he and Pickering could be equal, "three

old bachelors instead of only two men and a silly girl" (p. 105), allows Liza equality and freedom, inclusion in the definition of Man, only if she ceases to be female.

Liza denies her status as object and asserts her autonomous humanity, leaving Higgins' home and voluntarily entering into her engagement with Freddy. She knows that Higgins's wish for her to return to his home is simply a desire for her to resume her role as servant – "You want me back only to pick up your slippers and put up with your tempers and fetch and carry for you." (p. 98) She perceptively points out that her "voice and appearance," (p. 100) to which Higgins insists he has become attached, could mechanically be reproduced at will through his gramophone and photographs, without requiring her to be actually present. Her machine-generated presence would thereby meet Higgins' needs, and resolve the annoying problem of female autonomy by removing the human subject from the equation: "When you feel lonely without me, you can turn the machine on. It's got no feelings to hurt." (p. 100) Higgins is driven by this objection to distinguish (or at least pretend to distinguish) between Liza's mechanically reproduced likeness and her humanity. He now claims that his attachment to Liza is to her soul, not her appearance: "I cant turn your soul on. Leave me those feelings; and you can take away the voice and the face. They are not you." (p. 100) But his insistence that he "care[s] for life, for humanity," is compromised by his inability to fully include a woman within that category: "I care for life, for humanity; and you are a part of it that has come my way and been built into my house" (p. 100), he insists, in an incongruous statement that renders Liza as at once human and object, alive and mechanical. Ultimately, for Higgins, Liza is only a satisfying simulation of servile life and humanity, a female domestic object that forms part of the architectural fabric of his home. Higgins persists in treating Liza as a servile object right through to the end of the play, as his final lines and their accompanying stage direction demonstrate:

> HIGGINS. Oh, by the way, Eliza, order a ham and a Stilton cheese, will you? And buy me a pair of reindeer gloves, number eights, and a tie to match that new suit of mine. You can choose the colour. [*His cheerful, careless, vigorous voice shews that he is incorrigible.*]

Discussing the relationship between Shaw's play and its own source, Ovid's tale of Pygmalion and his statue in *Metamorphosis*, Nicholas Grene (2003) claims that Shaw's play represents a feminist reworking of Ovid's legend because Liza "learns how to talk back" (p. xvi)[10] Grene criticises the musical adaptation and stage versions which betray Shaw's feminist intentions by inserting a romantic relationship between Higgins and Liza, ending in matrimony. In contrast, Grene (2003) asserts that "the whole point of the play is the independent autonomy which Liza achieves, denying her status as Higgins' male artefact." (p. xviii) The play is not, however, as unequivocally optimistic as Grene's reading has it. Liza does learn how to talk back, but it is far from clear in her final words that she has, or will, escape from the position of subservient object:

> LIZA. Number eights are too small for you if you want them lined with lamb's wool. You have three new ties that you have forgotten in the drawer of your washstand. Colonel Pickering prefers double Gloucester to Stilton; and you don't notice the difference. I telephoned Mrs Pearce this morning not to forget the ham. What you are to do without me I cannot imagine. (p. 105)[11]

To modify Grene's reading, then, the point of the play might rather be said to be its insight into women's struggle to deny their status as artefacts in the face of the persistence of powerful men in treating them as such. More broadly, the play recognises that the sociogenic category of the

human, both its inclusions and exclusions (be they based on gender, class or more), is determined by, and serves to maintain, existing structures of power.

Today's VPAs

Liza recognises that "the difference between a lady and a flower girl is not how she behaves, but how she's treated." (p. 95) This statement confirms the sociogenic rather than ontological nature of categories such as class, and gender. But it also returns us interestingly to the Eliza effect. Its "dangers" include the misrepresentation of technological capacities and the perpetuation of gender stereotypes, but the Eliza effect also simultaneously demands and challenges the very possibility of critique in relation to artificially intelligent machines.

Consider, for example, today's VPAs, descended from Weizenbaum's ELIZA, which is the technical ancestor of their natural language processing software, and the antecedent of their gendered names: Alexa or Echo (Amazon), Cortana (Microsoft), Siri (Apple), and Cora (the Royal Bank of Scotland). These VPAs differ from Weizenbaum's ELIZA in their assistive function, but they resemble it in producing the Eliza effect in their users. Just as Weizenbaum's nomenclature and rhetoric deploys the Eliza effect, so too does the naming of these VPAs, and the language that is used about them in their marketing which consistently refers to them with female gendered pronouns. Both Weizenbaum's gendering of ELIZA and the female gendering of VPAs (through their names, pronouns, voices, and programmed self-identification) operationalise the Eliza effect and in doing so reproduce on the terrain of the machine the cultural association of the female with the non-human object and subservient labour that we have seen interrogated in Shaw's *Pygmalion*. Arguments to this effect are found across contemporary media think pieces which contend that the gendering of VPAs as female is a form of female digital servitude which reproduces and reinforces, within these new service technologies, societal stereotypes about women's "naturally" subservient role.[12] Female VPAs

transfer to the digital realm the gendering and stratification of labour found in the real world, with working women principally confined to jobs of lower power, status and pay, often within service industries or roles (McGinn & Oh, 2017).[13] In contrast, whereas digital *assistants* are gendered as female, digital *advisors* (operating in legal, financial or medical contexts, for instance) are gendered as male (Abrahams, 2018; Steele, 2018). Social psychology experiments (Mitchell et al, 2011; Nass & Brave, 2005) and market testing (Steele, 2018) confirm that users prefer female voices when interacting with assistive technologies, whereas male voices are trusted more in scenarios in which the computer assumes an authoritative role.

Recent academic and public work has made a strong case for the societal dangers posed by this gendering of VPAs, and has suggested legal and other ways to address these.[14] And progress has recently been made in moving away from gendering: although most market-dominant VPAs are still default female, many now have male voice options; in March 2019, the first non-binary VPA with a gender-neutral voice was launched (Hern, 2019; Smith, 2019); in September 2019 Google introduced new voice assistants named by colour, to avoid gendering (Davies, 2019); in December 2019 Leslie Witt, the female Vice President of Design at American business and software company, Intuit, discussed their design of a gender-neutral QuickBooks Assistant, a break away from the path dependency of other firms. Whilst this work and progress is necessary and welcome in its attempt to avoid the transportation of gender stereotypes from the human world into the technological, it does not address the broader "dangers" of the Eliza effect. For whilst the Eliza effect might be seen as expanding the capaciousness of the category of the human, that is, to include the artificially intelligent machine within its bounds, it in fact constitutes no progress at all in the struggle to secure, to recall Wynter's (2003, p. 260) words, "the full cognitive and behavioural autonomy of the human species" in the face of the narrowness of the definition of Man. That is, through the Eliza effect, artificially intelligent machines simply become subject to the same logics and

structures of power sociogenically determining who or what gets to be human and who or what does not.

Moreover, it is not just the gendering of machines (as either male or female) that induces and enhances the Eliza effect. The attribution of voice to the machinic serves the same function.[15] The question of the voice has been perhaps unexpectedly unattended to in this article. This is because it is not as prominent in either of the main texts under discussion as might at first be believed. Whilst Higgins' task in relation to Liza's transformation focuses on her speech, we have seen how the play does not focus on that task, but rather on an exploration of the struggle over who gets to be human. Likewise, despite Weizenbaum's repeated use of expressions associated with speech to describe human interaction with ELIZA, communication with the program was written. But communication with contemporary VPAs is verbal, and this change of mode serves to induce and enhance the Eliza effect irrespective of gendering. With voice technology, it is *actually* the case that we do not distinguish human from artificial. As Clifford Nass and Scott Brave explain in *Wired for Speech* (2005), humans do not distinguish between a human and an artificial voice since we use the same parts of the brain to process both:

> Listeners and talkers cannot suppress their natural responses to speech, regardless of source. People draw conclusions about technology-based voices and determine appropriate behaviour by applying the same rules and shortcuts that they use when interacting with people. These technologies, like the speech of other people, *activate* all parts of the brain that are associated with social interaction. (Nass and Brave, 2005, p. 4)

This means that when a human being is conversing with a VPA, the brain is processing that conversation as it would a conversation with another human being. The Eliza effect is here embedded in the neural response to the voice.

The Eliza effect is therefore powerful and difficult, in some cases perhaps impossible, to resist. Its operationalisation serves certain parties well – scientists whose funding opportunities benefit from its misrepresentation of their work; technologists whose businesses and sales benefit from the allure it gives to their products. Uninterrogated, it also contributes to what Sean Zdenek (1999) calls the wider techno-utopian "liberatory discourse" which seeks or offers in technology a miraculous panacea for the deeply culturally embedded inequities of our past and present society.[16] In opposition, the work of critique is essential. It is necessary in order that the capabilites of the science and technology are properly comprehended; it is necessary in order to expose the gendered and other norms and prejudices perpetuated in such technologies; it is necessary in order to properly understand the consequences of introducing the machinic into the sociogenic category of the human. But critique is also necessary in order to ensure that the liberatory discourse is challenged by a properly critical discourse, one which situates the Eliza effect within wider phenomena around artificially intelligent machines that obscure their imbrication in structures of power, control and inequity.

Works Cited

Abrahams, Ruth. (2018). Alexa, does AI have gender?. Research Feature, University of Oxford. https://www.research.ox.ac.uk/Article/2018-10-15-alexa-does-ai-have-gender

Adams, Rachel. (2020). *Transparency*. Routledge.

Bergen, Hilary. (2016).“I’d blush if I could”: Digital assistants, disembodied cyborgs and the problem of gender. *Word and Text*, 6, 95-113.

Boden, Margaret A. (1987). *Artificial Intelligence and Natural Man* (2nd edn.). Basic Books.

Boden, Margaret A. (1991). *The Creative Mind: Myths and Mechanisms*. Basic Books.

Carpenter, Charles A. (1965). The controversial ending of *Pygmalion*. *Shaw Review, 8*, 114.

Cave, Stephen, Dihal, Kanta & Dillon, Sarah. (2020). *AI Narratives: A History of Imaginative Thinking About Intelligent Machine*s. Oxford University Press.

Chambers, Amy. (2018, August 13). There’s a reason Siri, Alexa and AI are imagined as female – sexism. *The Conversation*. https://theconversation.com/amp/theres-a-reason-siri-alexa-and-ai-are-imagined-as-female-sexism-96430?__twitter_impression=true

Collett, Clementine & Dillon, Sarah. AI and Gender: Four Proposals for Future Research. https://doi.org/10.17863/CAM.41459

Davies, Hannah. (2019, September 18). Google Assistant is getting a second, gender neutral voice. *Trusted Reviews*. https://www.trustedreviews.com/news/google-assistant-is-getting-a-second-gender-neutral-voice-3939404

Fanon, Franz. (2008 [1952]) *Black Skin, White Masks*. Trans. Charles Lam Markmann. Pluto Press.

Fry, Hannah & Parker, Matt. (2019). *How to Bend the Rules*. Royal Institution Christmas Lectures. https://www.rigb.org/christmas-lectures/watch/2019/secrets-and-lies/how-to-bend-the-rules

Grene, Nicholas. (2003). Introduction. In *Pygmalion* by George Bernard Shaw. Penguin.

Gustavsson, Eva. (2005). Virtual servants: Stereotyping female front-office employees on the internet'. *Gender Work and Organization, 12*(5), 400-419.

Hadfield, S. A. & Reynolds, Jean. (2013). *Shaw and Feminisms: On Stage and Off*. University Press of Florida.

Havely, Cicely Palser. (2018). Happy ever after? The ending of Shaw's *Pygmalion*. *English Review*, 7(1), 26-9.

Hempel, Jessi. (2015, October 28). Siri and Cortana sound like ladies because of sexism. *Wired*. https://www.wired.com/2015/10/why-siri-cortana-voice-interfaces-sound-female-sexism/

Hartree, Douglas R. (1949). *Calculating Instruments and Machines*. The University of Illinois Press.

Hern, Alex. (2019, March 4). Adios, Alexa: why must our robot assistants be female? *The Guardian*. https://www.theguardian.com/technology/shortcuts/2019/mar/04/adios-alexa-why-must-our-robot-assistants-be-female

Hester, Helen. (2016, August 8). Technically Female: Women, Machines, and Hyperemployment. *Salvage*. https://salvage.zone/in-print/technically-female-women-machines-and-hyperemployment/

Hofstadter, Douglas. (1995). *Fluid concepts and creative analogies*. New York: BasicBooks.

Jotanovic, Dejan. (2018, June 19). This is how artificial intelligence is undoing women's rights. *The Independent*. https://www.independent.co.uk/voices/artificial-intelligence-siri-cortana-alexa-music-pop-culture-robotics-a8406391.html

Marriott, David. (2011). Inventions of Existence: Sylvia Wynter, Frantz Fanon, Sociogeny, and "the Damned". *CR: The New Centennial Review*, 11(3), 45-89.

Matlaw, Myron. (1958). The dénoument of *Pygmalion*. *Modern Drama*, I, 29-34.

McGinn, Kathleen L. & Oh, Eunsil. 2017. Gender, social class, and women's employment. *Current Opinion in Psychology*, 18, 84-88.

Mitchell, Wade J., Ho, Chin-Chang, Patel, Himalaya & MacDorman, Karl F. (2011). Does social desirability bias favour humans? Explicit-implicit evaluations of synthesised speech support a new HCI model of impression management. *Computers in Human Behaviour*, 27, 402-412.

Nass, Clifford & Brave, Scott. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. MIT Press.

Ni Loideain, Nóra & Adams, Rachel. (2019, December). From Alexa to Siri and the GDPR: The Gendering of Virtual Personal Assistants and the Role of Data Protection Impact Assessments. *Computer Law and Security Review*.

Nickelsburg, Monica. (2016, April 4). Why is AI female? How our ideas about sex and service influence the personalities we give machines. *Geek Wire*. https://www.geekwire.com/2016/why-is-ai-female-how-our-ideas-about-sex-and-service-influence-the-personalities-we-give-machines/

Park, Julie. (2020). Making the automaton speak: Hearing artificial voices in the Eighteenth Century. In *AI Narratives: A History of Imaginative Thinking About Intelligent Machine*s. Ed. Cave, Stephen, Dihal, Kanta & Dillon, Sarah. Oxford University Press. pp. 119-141.

Plasek, Aaron, (2017, October-December). On the Cruelty of Really Writing a History of Machine Learning. *IEEE Annals of the History of Computing*, 6-8.

Power, Nina. (2012). The dystopian technology of the female voice. *Her Noise*. http://hernoise.org/nina-power/

Scroggs, Matthew. (2015, August 27). MENACE: Machine Educable Noughts and Crosses Engine. http://www.mscroggs.co.uk/blog/19

Scroggs, Mattthew. (2019, December 27). Visualising MENACE's learning. http://www.mscroggs.co.uk/blog/tags/games

Shaw, George Bernard. (2003 [1913]). *Pygmalion*. Penguin.

Smith, Chris. (2019, March 11). Gender neutral AI voice developed to tackle Siri and Alexa stereotype. *Trusted Reviews*. https://www.trustedreviews.com/news/gender-neutral-assistant-alexa-siri-3674721

Steele, Chandra. (2018, January 4). The real reason voice assistants are female (and why it matters). *PC News*. https://uk.pcmag.com/opinions/92697/the-real-reason-voice-assistants-are-female-and-why-it-matte

Turing, Alan. (1950, October). Computing Machinery and Intelligence. *Mind LIX*(236), 433-60.

Treusch, Pat. (2017). Re-reading ELIZA: Human-machine Interaction as Cognitive Sense-ability. *Australian Feminist Studies*, 32(94), 411-26.

UNESCO (2019). I'd Blush If I Could: Closing Gender Divides in Digital Skills Through Education. https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=1

Wajcman, Judy. (1999). *Feminism Confronts Technology*. Polity Press.

Wajcman, Judy. (2004). *TechnoFeminism*. Polity Press.

Weizenbaum, Joseph. (1966). ELIZA – A computer program for the study of natural language communication between man and machine. *Computational Linguistics*, 9(1), 36-45.

Weizenbaum, Joseph. (1987 [1976]). *Computer Power and Human Reason: From Judgement to calculation*. Penguin.

Winer, Jerome A., Anderson, James William, & Kieffer, Christine C. (eds.). (2005). *Psychoanalysis and Women*. The Analytic Press.

Witt, Leslie. (2019, December 6). Digital Assistants Shouldn't Only Be Women – Here's Why. *Refinery*, 29. https://www.refinery29.com/en-gb/digital-assistant-gender-op-ed

Wynter, Slyvia. (2001). Towards the sociogenic principle: Fanon, identity, the puzzle of conscious experience, and what it is like to be "Black". In Durán-Cogan, Mercedes F. & Gómez-Moriana, Antonio (eds.), *National Identities and Sociopolitical Changes in Latin America*. Routledge. 30-66.

Wynter, Slyvia. (2003). Unsettling the Coloniality of Being/Power/Truth/Freedom: Towards the Human, After Man, Its Overrepresentation – An Argument.*CR: The New Centennial Review*, 3(3), 257-337.

Zdenek, Sean. (1999). Rising up from the MUD: inscribing gender in software design. *Discourse & Society*, 10(3), 379-409.

Notes

[1] The current vogue for "transparency" might be viewed as a contemporary instantiation of an attempt at demystification that has itself become a form of hype. See Adams (2020) for a deconstruction of the claims and assumptions of the discourse of transparency.

[2] The same rhetoric repeats in *Computer Power and Human Reason* ten years later: "I chose the name ELIZA for the language analysis program because like the Eliza of Pygmalion fame, it could be taught to 'speak' increasingly well." (3) Note that there is also a sleight of hand here in discussing ELIZA as "speaking" since all interaction with ELIZA was through a keyboard and writing. For reasons of space, the question of the voice and of gendering machine voices in particular is not attended to in this article, but see Power (2012) for interesting reflections on this.

[3] On the need for a comprehensive history of "machine learning" see Plasek (2016) and follow his doctoral research at the University of Columbia where he is constructing that history, closely tracing the changing valences of the term from its earliest mid-twentieth century uses to its proliferation in the twenty-first century. My thanks to Matthew Jones and Jonnie Penn for their historical knowledge here.

[4] For a wide-ranging study of the Western imaginative history of intelligent machines that informs contemporary AI research, reception and regulation, see Cave, Dihal and Dillon (2020). The second of the Royal Institution's 2019 Christmas Lectures (Fry, 2019) contains a brilliantly practical

demystification of machine "learning" that effectively discourages the Eliza effect through demonstrating how an assemblage of matchboxes "learns." For more on MENACE (Machine Educable Noughts and Crosses Engine) see Scroggs (2015 & 2019).

[5] See Hofstadter (1995, pp. 161-5) for an analysis of another instance in which Boden "falls prey to the Eliza effect", one of "several occasions" in her work *The Creative Mind* (1991) that Hofstadter claims have the effect of "unfortunately marring her book's accuracy and muddying the waters that she is working so hard to clarify' (p. 161).

[6] DOCTOR is read as strongly gendered male here for two reasons, literary and historical: first, the connection to Shaw's play through ELIZA's name means that DOCTOR recalls another character in the play, the male Dr. Henry Higgins; second, the program was modelled on Rogerian psychotherapy, developed by male psychologist Carl Rogers, and whilst there were some female psychotherapists and psychoanalysts in America in the mid-sixties (see Winer et al, 2005), the profession was still dominated by men.

[7] For a less critical feminist engagement with ELIZA that uses it to develop a feminist technoecological account of human-computer interaction see Treusch (2018).

[8] The term "sociogeny" was coined by Franz Fanon (2008, 4). It is further developed in the work of Sylvia Wynter (e.g. 2001; 2003). See Marriott (2011) for a thorough comparative analysis of Fanon and Wynter's theorisation of the term.

[9] On Shaw's influence on feminism see Hadfield and Reynolds (2013).

[10] For ease of distinction, the character in Shaw's play will be referred to as "Liza," as she is cued in the text of the play; Weizenbaum's programme is referred to consistently as "ELIZA."

[11] On the ending of *Pygmalion* see Carpenter (1965), Havely (1996), and Matlaw (1958).

[12] See, for just a few example, Abrahams (2018), Chambers (2018), Hempel (2015), Jotanovic (2018), Nickelsburg (2016), Steele (2018). For academic work on the gendering of VPAs, see Bergen (2016), who contextualises the gendering of today's VPAs in the history of feminist thinking about the female cyborg. On AI and gender more broadly, see Collett and Dillon's (2019) report outlining four of the weightiest challenges to gender equality presented by recent developments in artificial intelligence and four research proposals which would effectively tackle these issues.

[13] See Gustavsson (2005) for discussion of the gendering of service positions on the internet, and Hester (2016) on VPAs, labour and gender.

[14] Ni Loideain and Adams (2019) make a strong case for the societal harm caused by the gendering of VPAs, and propose how data protection law might be used to combat it. UNESCO's 2019 report on closing gender divides in digital skills through education highlights the seriousness of the societal effects of gendering AI, in particular VPAs: its title – "'I'd blush if I could'" – is taken from Siri's response to registering sexist insults from users; one of its main case studies – "The Rise of Gendered AI and its Troubling Repercussions" – focuses explicitly on the societal dangers of gendering VPAs.

[15] See Park (2020) for a discussion of the voice as an index of the human in the context of an exploration of eighteenth-century speaking automata.

[16] On the sexual politics of technology, and the changing nature of feminism's attitude to science and technology more broadly, see Wajcman (1991 & 2004).