A tale of two deficits : causality and care in medical AI

Chen, Melvin

2019

Chen, M. (2020). A tale of two deficits : causality and care in medical AI. Philosophy & Technology, 33(2), 245-267. doi:10.1007/s13347-019-00359-6

https://hdl.handle.net/10356/143941

https://doi.org/10.1007/s13347-019-00359-6

Philosophy & Technology (2020) 33:245–267
https://doi.org/10.1007/s13347-019-00359-6

RESEARCH ARTICLE

A Tale of Two Deficits: Causality and Care in Medical AI

Melvin Chen[1]

Received: 27 September 2018 / Accepted: 14 May 2019 / Published online: 12 June 2019

Abstract

In this paper, two central questions will be addressed: ought we to implement medical AI technology in the medical domain? If yes, how ought we to implement this technology? I will critically engage with three options that exist with respect to these central questions: the *Neo-Luddite* option, the *Assistive* option, and the *Substitutive* option. I will first address key objections on behalf of the *Neo-Luddite* option: the Objection from Bias, the Objection from Artificial Autonomy, the Objection from Status Quo, and the Objection from Inscrutability. I will thereafter present the Demographic Trends Argument and the Human Enhancement Argument in support of alternatives to the *Neo-Luddite* option. In the second half of the paper, I will argue against the *Substitutive* option and in favour of the *Assistive* option, given the existence of two chief formal deficits in medical AI technology: the causality deficit and the care deficit.

Keywords Causality deficit · Care deficit · Health · Medical AI · Care robots · Care ethics · Causal reasoning · Statistical reasoning · Assistive option · Substitutive option · Neo-Luddite option · Value sensitive design · Capabilities approach · Artificial intelligence

1 Introduction

Ought we to implement medical AI technology in the medical domain? If yes, how ought we to implement this technology? These are the two central questions that will be addressed in this paper. I will outline three options that exist with respect to the possible implementation of technology in the medical domain: the *Neo-Luddite* option

Academic Profile: http://research.ntu.edu.sg/expertise/academicprofile/Pages/StaffProfile.aspx?ST_EMAILID=MELVINCHEN&CategoryDescription=Philosophy

✉ Melvin Chen
melvinchen@ntu.edu.sg

[1] Nanyang Technological University, 26 Nanyang Avenue, Hall 8, Blk 44 #01-819, Singapore S639812, Singapore

(hereafter: *NO*), the *Assistive* option (hereafter: *AO*), and the *Substitutive* option (hereafter: *SO*). According to *NO*, one ought to uniformly resist the implementation of medical AI technology in the medical domain and (where possible) de-technologize our medical healthcare practices. According to *AO*, one ought to implement medical AI technology in the medical domain to assist our human medical healthcare professionals, without these professionals themselves being replaced. According to *SO*, one ought to implement medical AI technology, with a view to completely replacing our human medical healthcare professionals with fully automated technological alternatives in the long run. *NO* responds with a resounding 'no' to the first central question. *AO* and *SO* both respond in the affirmative to the first central question, although they differ in their responses to the second central question. I will present arguments against the *NO*, before demonstrating that the existence of at least two deficits—the causality deficit and the care deficit—on the medical AI front compels us in favour of *AO* instead of *SO*, contingent upon the satisfaction of certain constraints.

2 Health and Medical AI Technology

How is the concept of health to be defined? The naturalist conception of health is a value-free, scientific one, according to which health is a lack of deviation from the statistical norm, normal biological functioning, or the absence of disease (Boorse 1977, 1997). Conversely, the normativist conception of health is an inescapably value-laden one, according to which our concern is with suffering and not just disease. The naturalist maintains that the normativist ideal of positive health is not scientifically discoverable but merely advocable (Boorse 1977, p. 572). The normativist, on the other hand, holds that what the naturalist gains in theoretical clarity, she loses in practical appeal: defining health in terms of biological function or the mere absence of disease limits the applicability of the concept for healthcare professionals and patients. The World Health Organization (WHO) has endorsed the normativist notion of positive health: its constitution defines health as a state of complete physical, mental, and social well-being and not merely the absence of disease and asserts in addition that the enjoyment of the highest attainable standard of health is one of the fundamental rights of every human being (WHO 1948). Putnam (2002) has argued that fact and value are intermingled in scientific practice. Once we accept Putnam's position, then there will be no logical inconsistency in conceptualizing health as both the absence of disease and as positive health.[1] While the diagnosis, prevention, monitoring, treatment, and alleviation of disease are the traditional province of the medical profession, normative considerations about physical, mental, and social well-being and the fullest attainment of health standards are also medically relevant. It is this holistic conception of health, in which fact and value and intricately intertwined, that I will adopt for the rest of this paper.

How in turn are we to define medical AI technology? Historians of technology admit from the outset that technology is messy, complex, and difficult to define (Hughes 2004, p. 1). Typically, technology may be characterized both in terms of a process (the design, development, and deployment of tools, machines, techniques, systems, and

[1] For a more sustained defence of this holistic view about health, see the naturalized normativism of Richard Hamilton (2010).

methods of organization to solve problems) and its output (the tools, machines, techniques, systems, and methods of organization themselves). For my argumentative purposes, technology-as-output and technology-as-process are united by a common purpose: the solving of well-defined problems or the performance of specific tasks.[2] Medical AI technology denotes novel technology that possesses some degree of autonomous decision-making capabilities, an ability to learn from data and perform tasks without being explicitly instructed to do so, and a capacity to solve medical domain-specific problems of a sufficiently high level of computational complexity. Medical AI technology tends to be designed and deployed for solving medical domain-specific problems that are of a higher degree of computational complexity than traditional medical technology. Given how a holistic conception of health includes the notion of positive health and attendant considerations about well-being and quality of life, I am also comfortable with characterizing medical AI technology as a subset of quality of life technologies, which are novel, intelligent technologies specifically designed to affect the quality of life of individuals who use them (Kanade 2012; Schulz 2013).[3]

3 *NO*: Reasons, Objections, and Responses

A Neo-Luddite is an individual who is opposed to the general use of modern technology. The intellectual heritage of neo-Luddism can be traced to the Luddites or British weavers and textile workers from the nineteenth century who opposed the implementation of Industrial Revolution-era technology (viz. knitting frames and automated looms). *NO*, the option favoured by the Neo-Luddite, is the option according to which one ought to uniformly resist the implementation of medical AI technology in the medical domain and (where possible) de-technologize our medical healthcare practices. If *NO* is grounded in such psychological tendencies as the fear or dislike of novel and unfamiliar devices and technology (otherwise known as technophobia), then it remains irrational in nature and yields a naïve strawman version of *NO* that is of little philosophical interest to us. We are looking rather for reasons that might ground a general skepticism toward the implementation of medical AI technology, yielding a more sophisticated version of *NO*.

Reasons for opposing the implementation of medical AI technology cannot be derived from their novelty and unfamiliarity. Technologies that were designed, developed, and deployed much earlier than medical AI technology must once have been novel and unfamiliar, yet no Neo-Luddite would oppose their implementation in the medical domain today. Among these technologies, one could count the use of language, pen and paper, eyeglasses, and even generic software applications.[4] In the absence of any further justification for resisting the implementation of medical AI technology but

[2] This could be extended to cover the improvement of pre-existing solutions, the achievement of goals, and the handling of certain input/output relations (Schatzberg 2006).

[3] This conceptualization of medical AI technology will be especially important in the context of the care deficit, which shall be addressed in §8.

[4] I exclude generic software applications such as word-processing programs from my conception of medical AI technology, since I agree with John McCarthy's dictum that as soon as AI works, no one calls it 'AI' any more (cited in Vardi (2012) and Bostrom (2014)).

not the use of language, pen and paper, eyeglasses, and even generic software applications, the Neo-Luddite could be charged with arbitrariness or logical inconsistency.

The Argument from Inconsistency (alternatively: the Eyeglasses Argument) against *NO* may be represented thus:

> P1: If one ought to uniformly resist the implementation of medical AI technology in the medical domain, then one ought for the sake of consistency to resist the implementation of all forms of technology (including eyeglasses) in the medical domain.
> P2: It is not the case that one ought to resist the implementation of all forms of technology (including eyeglasses) in the medical domain.
> C: Therefore, it is not the case that one ought to uniformly resist the implementation of medical AI technology in the medical domain.[5]

3.1 The Objection from Bias

Objections to the implementation of medical AI technology should be grounded in genuine and rational concerns about the general use of medical AI technology. According to the Objection from Bias, there is the possibility of bias being embedded into decision procedures. In 2016, ProPublica analysed the efficacy of COMPAS, a criminal risk assessment tool developed by Northpointe (now known as Equivant), and determined that the predictions of its algorithm were unreliable and racially biased: it underpredicted recidivism for white defendants and overpredicted recidivism for black defendants (Angwin et al. 2016). Ought we not to worry about biases of this nature (racial, socioeconomic, political, etc.) becoming embedded into medical decision-making procedures with a large-scale and unchecked implementation of medical AI technology?

3.1.1 The Simple Heuristics Response

A number of responses to the Objection from Bias are available. While bias traditionally carries a negative connotation, one might argue that there is a need for inductive or learning bias. Further, one might assert, as AI researchers such as Tom Mitchell have done, that bias-free learning is futile. In machine learning-based approaches, the optimal choice of values for the weights and the bias is what allows for the correct classification of training examples and learning from datasets to take place (Abu-Mostafa et al. 2012, pp. 5–6). This is the Simple Heuristics Response to the Objection from Bias. An inevitable trade-off must be recognized: predictive parity (fairness at predicting whether a defendant of any

[5] Formally:
P1: $(P \supset Q)$
P2: $\sim Q$
C: $\therefore \sim P$ (*modus tollens* rule of inference)

race, classified as high risk, will subsequently be rearrested), equal false positive rates (no disproportionate number of defendants of a particular race being misclassified as high risk), and equal false negative rates (no disproportionate number of defendants of a particular race being misclassified as low risk and subsequently rearrested) are different measures of fairness for COMPAS and it is mathematically impossible to satisfy all of them (Chouldechova 2017).

3.1.2 The Heuristics and Biases Response

While one could concede that it is mathematically impossible to satisfy all measures of fairness, one might still distinguish between instances in which biases are justified and other instances in which they are unjustified and ought to be minimized. This may be termed the Heuristics and Biases Response to the Objection from Bias: the same cognitive processes that generate biases in certain instances can also generate quick and accurate judgments in other instances. Human beings are thought to employ two different processes of thinking: System 1 thinking (automatic, fast, and unconscious) uses heuristics or rules of thumb whereas System 2 thinking (controlled, slow, and conscious) requires analytical effort (Sloman 1996; Barbey and Sloman 2007; Kahneman 2011).[6] When the time for decision-making is limited, cognitive resources are at a premium, sample sizes remain small, and noise is present in the observational data, heuristics is useful as a cognitive tool for human beings to navigate a world of uncertainty (Gigerenzer and Todd 1999; Gigerenzer and Brighton 2009). In these instances, System 1 thinking outperforms System 2 thinking and a biased induction algorithm fares better and makes more accurate predictions than an unbiased one (Hastie et al. 2001).

Medical healthcare professionals, for instance, could employ the representativeness heuristic as a decision-making shortcut and rely on past experience with similar patients when treating a new patient (Kahneman and Tversky 1973; Redelmeier and Tversky 1990). Systemic lupus erythematosus (SLE) is a disease that is distributed across races and genders, although higher rates are observed in women and non-Caucasian races (Pons-Estel et al. 2017). If the patient is a black woman who has developed a butterfly rash, then the doctor performing a time-constrained differential diagnosis is well within her means to employ the representativeness heuristic (if she has treated similar black female patients in the past) or make heuristic use of the relevant group characteristics (if she is aware of the epidemiology of SLE). In other instances, however, System 2 thinking might trump System 1 thinking: medical healthcare professionals might fall prey to the availability bias and diagnose patients according to the list of possible human diseases that are mentally available to them or base their decisions on easily recalled, dramatic examples of similar patients, when a more controlled, effortful, and deliberate decision procedure would yield more optimal outcomes.

[6] According to the dual-process theory of thinking, these two systems of thinking occur in different areas of the brain and have different metabolic requirements.

3.1.3 The Value Sensitive Response

How we distinguish between instances in which biases are justified and other instances in which they are unjustified depends in part on what we value. Among our desiderata are truth and effectiveness of outcome (leading us to value accuracy in prediction) and fairness of outcome (leading us to disvalue racial bias). According to the Value Sensitive Response to the Objection from Bias, values and biases are embedded into computer systems and AI technologies embody human values. In accordance with the Value Sensitive Design Approach that I will delineate in §6.1, we can include among our AI design imperatives the imperative to minimize the incidence of unjustified biases. Properly formulated, Value Sensitive Design could be entirely consistent with the recognition of the need for inductive bias (as asserted in the Simple Heuristics Response), the usefulness of heuristic techniques in certain instances (as asserted in the Heuristics and Biases Response), and the disvalue of certain unjustified forms of bias (as asserted in the Value Sensitive Response).

3.2 The Objection from Artificial Autonomy

According to the Objection from Artificial Autonomy, worries arise when medical AI systems can operate fully autonomously without further human intervention, unlike some other traditional forms of technology. Just as human drivers might become less alert or competent when cars gain automated driver assistance, might the level of competence and quality of healthcare not dip when medical AI systems are fully automated (Mukherjee 2017)? I do not wish to ignore the worries that are posed by the Objection from Artificial Autonomy. Consider a medical doctor who over-relies on a search engine rather than her own domain-specific knowledge and expertise. The use of an Internet-based search engine can assist her in reducing the search space quickly and retrieving information efficiently in the short run, although an over-reliance on this technology (to the point of automating search queries) could be detrimental to her ability to recall information and other related cognitive capacities in the long run. There is the further possibility that the doctor-patient interface will become more superficial, predictable, and transactional and suboptimal standards of healthcare would be attained. A key assumption here is that we have two and only two options and are forced to choose one of these two options: one ought either to resist the implementation of medical AI technology or resign oneself to the full automation of this technology.

The Objection from Artificial Autonomy is grounded in the Fallacy of the False Dilemma:

> P1: Either *NO* (resisting the implementation of medical AI technology) or *SO* (supporting the full automation of medical AI technology) ought to be the case.
> P2: It ought not to be the case that *SO*.
> C: Therefore, it ought to be the case that *NO*.

The Objection from Artificial Autonomy ignores how there is at least one additional option: *AO*, which neither resists the full implementation of medical

AI technology nor mandates the full automation of medical AI technology. We could be careful in our implementation of medical AI technology, ensuring that the level of professional competence and quality of healthcare do not dip in the face of technological changes. The Value Sensitive Response that we applied to the Objection from Bias will be equally salient here: we value inter alia accountability, professional competence in our healthcare professionals, and better-quality healthcare, and we ought to design AI technologies that account for and embody these values in a principled and comprehensive manner through the design process.

3.3 The Objection from Status Quo

A Neo-Luddite might concede on behalf of *NO* the plausibility of *AO* and sidestep the Fallacy of the False Dilemma, only to assert that medical domain-specific problems are regularly solved in certain parts of the world where the influence of medical AI technology is minimal. In much of the developing world (and many sections of the healthcare system in the developed world, for that matter), medical systems are old-fashioned, outdated physical record-keeping systems are the norm, and legacy technologies remain in place due to a lack of funding. Nonetheless, the diagnosis, prevention, monitoring, treatment, and alleviation of disease still proceed apace.

Why ought one to adopt medical AI technology (whether fully or partially automated) when simpler measures appear to suffice? This is the Objection from Status Quo. As an initial response, one could attempt to demonstrate how the Objection from Status Quo generalizes to a Cherry Picking Fallacy:

P1: Both evidence A and evidence B are available.
P2: Evidence A (instances in which the risks of implementing medical AI technology outweigh the benefits) supports the claims of *NO*.
P3: Evidence B (instances in which the benefits of implementing medical AI technology outweigh the benefits) supports the counterclaims of *NO*.
C: Therefore, only the claims of *NO* are supported.[7]

One might admit from the outset that the implementation of medical AI technology is unfortunately not equitable between resource-poor and resource-rich countries. One could even concede that more harm than good might result in certain instances: think of the ineffective equipment and incompatible programs that must be discarded if inappropriate or unsound implementation strategies are employed. Nonetheless, once a sufficient level of infrastructure (degree of connectivity, software availability, hardware overhang, etc.) has been attained, an equitable deployment and use of medical AI technology is likely to promote both the absence of disease and positive health (as discussed in §2), increase human well-being and quality of life, and improve the standards of healthcare, as can

[7] The stronger the counterevidence that has been suppressed or withheld, the more fallacious the argument.

be observed in a large number of resource-rich countries. A modified version of the Argument from Inconsistency in §3 may be applied to the Objection from Status Quo: each status quo is itself a departure from previous status quos, and reasons for opposing the implementation of medical AI technology ought not to be derived from the novelty and unfamiliarity of these technologies to individuals who have merely become habituated to the current status quo. In a related vein, the Value Sensitive Response to the Objection from Bias and the Objection from Artificial Autonomy can be extended to the Objection from Status Quo: we should not sell resource-poor countries short by accepting the familiar, hassle-free, satisficing status quo instead of holding out for the possibility of a more equitable and outcome-maximizing alternative.

3.4 The Objection from Inscrutability

According to the Objection from Inscrutability, medical AI technology is apt to function in a manner that renders its inner workings mysterious or inexplicable to human beings, rendering oversight and accountability difficult or impossible. The Objection from Inscrutability is not a knockdown objection. The weakest response to this objection is the Bullet-biting Response: AI researchers and machine learning experts like Sebastian Thrun and Geoffrey Hinton regularly acknowledge that AI systems are like black boxes and that even AI designers are unable to explain why their systems arrive at a particular decision. Nonetheless, it may be urged that the black box problem is something we could live with. After all, are certain types of human insight (e.g., creative insight) not traditionally held to be inscrutable as well? A stronger response involves proposing alternative ways of scrutiny: one could propose the large-scale design of transparent, trustworthy, and easily comprehensible AI systems in accordance with Explainable AI (or XAI) principles.

The strongest response, in my view, involves recognizing the usefulness of the human-AI interface in *AO*. Let us call this the Interface Response. In certain instances, System 1 thinking clearly trumps System 2 thinking and heuristic techniques will be useful: the human-AI interface will tend to get decisions right and worries about accountability will remain at a minimum. In other instances, System 1 thinking could be inferior to System 2 thinking and the interface might react thus: the AI system (practising System 1 thinking) will slow down and ask for reinforcements from the human healthcare professional (System 2) (Topol 2019, pp. 44–5). Through the deliberate interfacing between the artificial and human elements in *AO*, unjustified biases can be minimized (*pace* the Objection from Bias), human beings remain in the loop and retain the ability to intervene (*pace* the Objection from Artificial Autonomy), the enhancement or augmentation of human capacities might lead to more equitable and outcome-maximizing alternatives (*pace* the Objection from Status Quo), and a greater degree of human intervention, participation, and scrutability will be anticipated in instances wherein the human-AI interface tends to get decisions wrong (*pace* the Objection from Inscrutability).

4 The Case for Implementing Medical AI Technology and *AO*

In §3, I have sought to present *NO* in its strongest and most sophisticated form, the better for it to serve as a foil, against which alternatives to *NO* would appear more convincing if the principal objections raised by *NO* in §§3.1–3.4 can be overcome. We are now ready to consider the reasons for the implementation of medical AI technology in the medical domain.

4.1 The Demographic Trends Argument

The post-war baby boom, the rise in life expectancy, the decline in birth rates, and changes in family structure are among the chief demographic trends that have led to a rising demand for healthcare services for certain sections of the population (viz. the frail and elderly) in the developed world. While these demographic trends will increase the demand for healthcare resources in the future, these resources are finite and scarce. Rather tellingly, a recent sentiment analysis study of the surveyed perceptions of patients suggests that healthcare professionals tend to be viewed as rushed, busy, and hurried (Singletary et al. 2017). Given the finitude and scarcity of healthcare resources, healthcare professionals find it easier to engage in shallow rather than deep medicine and prescribe narcotics instead of listening to and understanding patients. This has fostered the opioid epidemic in current medical practice (Topol 2019, p. 4). Another symptom of the pressure on healthcare resources may be found in the litany of unnecessary and overused medical tests and procedures (including medical imaging studies for lower back pain and stenting for patients who are unlikely to get any benefit) that continue to be prescribed by medical professionals who have failed to keep up with the scientific evidence (Brownlee et al. 2017; Epstein 2017). The foreseen shortfall of healthcare resources and medical healthcare professionals in the near future will exacerbate these trends and has led to proposals for a large-scale implementation of medical AI technology to cope with these societal pressures.[8]

The Demographic Trends Argument in favour of implementing medical AI technology may be represented thus:

P1: If the current demographic trends persist, then there will be an increase in demand for healthcare resources.
P2: If there is an increase in demand for healthcare resources and healthcare resources are insufficient, then we ought to implement medical AI technology in the medical domain.
P3: Healthcare resources are insufficient.

[8] Rather tellingly, Sparrow and Sparrow (2006) have noted how remarkable it is that much robotics research has been promoted by appealing to the idea that the only way of dealing with these demographic pressures and the anticipated care gap is to develop robots to look after the elderly. The rapid pace of technological development and the growing interest from industry, business, and government agencies in implementing medical AI technology to meet healthcare needs also strengthen the appeal of the Demographic Trends Argument (Schulz et al. 2015, p. 724).

C: Therefore, if the current demographic trends persist, then we ought to implement medical AI technology in the medical domain.[9]

4.2 The Human Enhancement Argument

The idea that technology imitates nature as its exemplar is at least as old as Plato.[10] We could implement medical AI technology that imitates or mimics the natural abilities and capacities of medical healthcare professionals, and this would still be favourable if the resource scarcity problem is thereby addressed. However, other philosophers of technology have maintained that technology is not confined to the mere imitation of nature: Aristotle (n.d., II.8, 199a15) has argued that technology can in some cases complete what nature cannot finish. In the case of the medical domain, certain limitations (cognitive, physical, etc.) of human beings could render numerous medical domain-specific problems either impossible or difficult to solve. State-of-the-art medical AI technology has better memory bandwidth, can handle larger amounts of medical data more effectively than even the most competent medical professionals, possesses superior computational power, is less prone to fatigue, and can share knowledge and skills (by data-swapping, for example) at a faster pace than ordinary human beings.[11] Human medical professionals who rely on assistive AI technologies will be less rushed, busy, and hurried and more at leisure to attend to and care for their patients. Human medical professionals who rely on symptom checker programs (or tools that employ algorithms to help patients with self-diagnosis or self-triage) such as the Isabel Symptom Checker will be less susceptible to the availability bias (as described in §3.1.2), wherein they confine their diagnosis to the list of possible

[9] Formally:
P1: (P ⊃ Q)
P2: ((Q.R) ⊃ S)
P3: R
C: ∴ (P ⊃ S)
Proof of validity:
1: (P ⊃ Q)
2: ((Q.R) ⊃ S)
3: R
(P ⊃ S) (block off conclusion and assume the negation of C)
4: asm: ~(P ⊃ S)
5: ∴ P (from 4)
6: ∴ ~S (from 4)
7: ∴ Q (from 1 and 5) (*modus ponens*)
8: ∴ (Q.R) (from 3 and 7)
9: ∴ S (from 2 and 8) (*modus ponens*)
10: ∴ (P ⊃ S) (from 4; 6 contradicts 9) (*reductio ad absurdum*) (QED)

[10] See Book X of Plato's (2016) *Laws*.

[11] The problem-solving abilities of medical AI technology exceed, in this sense, the upper bound set by the cognitive limitations of human beings. Human cognitive abilities often pale in comparison to the information-processing capacities and computational abilities of medical AI: the electrochemical, analog processes in the human brain (120 m/s or less) are slower than the speed-of-light processes (300,000,000 m/s) of modern microprocessors and biological neurons operate at a peak speed (200 Hz) that is 7 orders of magnitude lower than the peak operating speed of modern microprocessors (c. 2 GHz) (Bostrom 2014).

human diseases that are mentally available to them.[12] Human medical professionals who rely on AI augmentation will be better equipped to keep up with the scientific evidence and less prone to prescribing unnecessary and unhelpful medical tests and procedures. Telemedicine will allow for the overcoming of physical barriers and enable clinical healthcare to be provided from a distance. Exoskeleton suits will allow human care-givers to exceed their natural physical limitations, physically enhance themselves with the brute power and strength of exoskeleton suits, and perform previously impossible physical tasks such as lifting and carrying a heavy patient.

The Human Enhancement Argument in favour of implementing medical AI technology may be represented thus:

> P1: If certain limitations (cognitive, physical, etc.) of human beings are overcome, then a wider range of medical domain-specific problems is likely to be solved.
> P2: If a wider range of medical domain-specific problems is likely to be solved, then the standards of healthcare and human well-being are likely to improve.
> P3: Certain limitations (cognitive, physical, etc.) of human beings are overcome if medical AI technology is implemented.
> C: Therefore, if medical technology is implemented, then the standards of healthcare and human well-being are likely to improve.[13]

4.3 The Case for Alternatives to *NO*

The Demographic Trends Argument and the Human Enhancement Argument are arguments in favour of alternatives to *NO*. In addition, I have offered proof in fn 9 and fn 13 of the validity of both these arguments: assuming that all the premises of each argument are

[12] In a recent comparative study of the diagnostic accuracy of physicians and AI systems, human medical doctors outperformed symptom checkers on the diagnostic accuracy front by 72.1% to 34.0% (Semigran et al. 2016, p. 1860). Where human diagnostic error tends to prevail, however, a case can be made for augmenting or enhancing physician diagnostic accuracy with assistive medical AI technologies.

[13] Formally:
P1: $(P \supset Q)$
P2: $(Q \supset R)$
P3: $(P \equiv S)$
C: $\therefore (S \supset R)$
Proof of validity:
1: $(P \supset Q)$
2: $(Q \supset R)$
3: $(P \equiv S)$
$(S \supset R)$ (block off conclusion and assume the negation of C)
4: asm: $\sim(S \supset R)$
5: $\therefore S$ (from 4)
6: $\therefore \sim R$ (from 4)
7: $\therefore (P \supset R)$ (from 1 and 2) (hypothetical syllogism)
8: $\therefore \sim P$ (from 6 and 7) (*modus tollens*)
9: $\therefore (S \supset P)$ (from 3)
10: $\therefore \sim S$ (from 8 and 9) (*modus tollens*)
11: $\therefore (S \supset R)$ (from 4; 5 contradicts 10) (*reductio ad absurdum*) (QED)

true, it would be logically impossible for the conclusion nevertheless to be false. I take P1–P3 of the Demographic Trends Argument and P1–P3 of the Human Enhancement Argument to be highly plausible and intuitively appealing, although it is beyond the scope of my paper to demonstrate the actual truth of each of these premises and the concomitant soundness of both arguments. I take it for granted that demographic trends, the frenetic pace of technological development, resource scarcity, and the possibility of improved problem-solving through AI-based cognitive and physical enhancement techniques all constitute reasons in favour of the implementation of medical AI technology in the medical domain. As healthcare resources get freed up, it is not just the physical, mental, and social well-being of the care-receiver or patient that will improve. Overworked healthcare professionals have been found to experience a reduction in stress levels and an improvement in their overall levels of well-being following the introduction of medical AI technology.[14] With an appropriate extension of the notion of positive health to cover the needs of both the care-receiver and the care-giver, a virtuous cycle of well-being can be fostered: healthcare professionals whose care-giving burdens have been reduced by the implementation of medical AI technology are likelier to provide better-quality healthcare to their patients in the long run (Sharkey 2014).

AO is the option according to which one ought to implement medical AI technology in the medical domain to assist our human medical healthcare professionals, without these professionals themselves being replaced. In Japan, *AO* is an outcome that enjoys heavy government backing: the government has been funding the development of care robots for the elderly, in order to fill the anticipated shortfall of 380,000 specialized healthcare workers by 2025. Atsushi Yasuda, director of the robotic policy office at the Ministry of Economy, Trade, & Industry in Japan, has even described the development of care robots for the elderly as an opportunity and expressed the hope that other developed countries facing similar demographic challenges might follow the Japanese lead. At least in certain parts of the developed world, *AO* is already the norm: doctors and nurses have become highly skilled at the use of assistive, high-end, instrumental technology (ranging from clinical information systems to integrated electronic health records, wearables, compact and portable medical devices, drug retrieval-and-delivery systems, and exoskeleton suits for lifting and carrying patients). In these instances, *AO* is no longer strictly about the implications of implementing medical AI technology in doctor-patient and nurse-patient relations that are devoid of this technology. Rather, we are already in a context where *AO* is being exercised to a fair extent and our question becomes a slightly subtler one of how the introduction of even more AI technology will alter existing healthcare practices (Van Wynsberghe 2013).

5 The Case Against *SO*

Why stop at *AO*? Why not quest after the holy grail of completely replacing our human medical healthcare professionals with fully automated technological alternatives in the long run? These are questions that will be raised by *SO*. While both *AO* and *SO* both stand opposed to *NO* and are motivated by reasons in favour of the implementation of medical AI

[14] See Mordoch et al. (2013) for preliminary evidence that healthcare staff might experience less burnout when therapeutic robots are used in a day service unit.

technology to solve medical domain-specific problems, *AO* stops short of recommending the final outcome of *SO*: the complete displacement of human medical healthcare professionals by fully automated technological artifacts, techniques, or processes. If the primary aim of the implementation of medical AI technology is to solve medical domain-specific problems, then *SO* is at least on a par with *AO* when there is no discernible difference in the problem-solving abilities of fully automated technological artifacts, techniques, or processes (the final outcome of *SO*) and those of the various human-AI interfaces we might encounter in *AO*. Unfortunately, differences exist between human- and AI-level performances for certain medically relevant tasks, to the detriment of *SO*. A formal deficit may be informally construed as a deficit wherein the optimal human-level performance of human professionals exceeds the optimal level of performance of state-of-the-art AI, with no prospect of an adequate formalization in the near future that might ensure a parity in performance levels. A number of formal deficits exist with respect to medical AI technology. Unless these deficits can be effectively remedied, a defence of *SO* will be no less dogmatic than a defence of the naïve textbook version of *NO*. In the rest of this paper, two of the chief formal deficits will be examined in detail: the causality deficit and the care deficit.

5.1 The Causality Deficit

In the medical domain, the doctor is typically responsible for making accurate diagnoses of medical conditions on behalf of her patient. The ability or capacity to make correct causal inferences is essential on the diagnostic front. Each diagnosis is in effect a causal inference, as doctors reason about causal claims that relate putative causes to observed effects (viz. the patient's symptoms). Effective causal diagnosis the medical domain requires more than computational power and efficiency, memory bandwidth, knowledge- and skill-sharing capacities, or any of the other cognitive advantages that medical AI technology can provide. Algorithm-based medical AI technology is typically designed to solve computational problems correctly and efficiently, and each algorithm is typically analysed in terms of the resources it consumes (viz. worst-case running time, space complexity, memory, computational bandwidth). With causal diagnosis, however, the correct causal diagnosis could be the resource-consuming and time-inefficient one. The ability to provide time-efficient diagnoses is not essential to the provision of the highest attainable standards of healthcare: some accurate diagnoses require time and painstaking effort, some swift diagnoses could turn out to be ultimately inaccurate, and some otherwise competent doctors might be naturally slow in making medical diagnoses on behalf of their patients. Just as the correct diagnosis of human medical doctors need not be efficient, the efficient diagnosis of medical AI technology need not be correct. State-of-the-art expert diagnostic programs such as Stanford's MYCIN and—more recently—IBM's Watson lack a causal model of reality and tend rather to rely on statistical correlation as a guide to causation.[15]

[15] Theirs may be termed a model-blind approach to causal reasoning, as opposed to a model-based approach. In the history of science, the ancient rivalry between Babylonian science and Greek science may be interpreted as a rivalry between a model-blind approach and a model-based approach (Toulmin 1961). Machine learning-based state-of-the-art AI is analogous to Babylonian science: one adopts a model-blind approach. On the causal inference front, conversely, one requires a principled and model-based approach to distinguish between spurious correlations and causal correlations (Pearl 2018).

The optimal human-level performance of human professionals in causal inference tasks is astounding. There is sufficient evidence in developmental psychology to suggest that children are able to make causal inferences from an early age and that their causal inferential processes might involve computations similar to those for learning with causal Bayes nets and making predictions with causal Bayes nets (Schulz and Gopnik 2004).[16] Conversely, state-of-the-art AI constitute statistical rather than causal reasoning machines, and the optimal level of performance of state-of-the-art AI in causal inference tasks is far inferior to the optimal human-level performance, with no prospect of a performance-parity-generating formalization in sight.[17] This yields a causality deficit in medical AI technology.

5.2 Addressing the Causality Deficit

If the causality deficit exists, then *SO* would seem to be out of the question, at least until the causality deficit of state-of-the-art AI systems has been appropriately addressed. If one is to pursue a solution to the causality deficit, then one must first ask: can a causal model of reality be formalized and mechanized? Let us assume for the sake of argument that it is possible for one to answer in the affirmative. If an AI system comes to possess a causal model of reality, then we have good reason to expect that its explanatory powers will be up to scratch and that it will no longer be inscrutable. By formalizing and mechanizing this causal model of reality, we could address both the causality deficit and the Objection from Inscrutability (as discussed in §3.4). One might ask next: what is the most desirable way in which to proceed with the formalization and mechanization of causal reasoning? A low-hanging fruit would be Judea Pearl's (2000) Structural Causal Model (or SCM), which relies on the use of diagrams (more specifically: Directed Acyclic Graphs or DAGs) and equation models (more specifically: Nonparametric Structural Equation Models or NPSEMs). Pearl's work on causality represents a formalization of causal relations and a deductive approach to causation. Whether or not Pearl's formalization of causation can be mechanized to produce causal reasoning machines is a separate issue.[18] All things considered, addressing the causality deficit would seem to require inter alia the ability to manipulate Directed Acyclic Graphs and Nonparametric Structural Equation Models (under Pearl's formal approach to causation), reason counterfactually, and operate in a formalized imagination space (under Mahadevan's formal approach to imagination machines).

[16] A causal Bayes net is a graphical model that represents a variable set V and the conditional dependence relations between variables in this set via a directed acyclic graph. The Causal Markov Condition is the condition according to which any variable x in a variable set V is conditionally independent of its non-effects, given its direct causes. Bayesian networks, invented by Judea Pearl (2000), allow for causal structure to be represented.

[17] The correlation-as-a-guide-to-causation heuristic may be traced back to the regularity theory of causation (Hume, 2007, Mill 1843). For an argument about how the human brain is not hardwired for statistical reasoning, see the discussion about statistical paradoxes (viz. Simpson's paradox and the Monty Hall paradox) in Pearl and MacKenzie (2018).

[18] Mahadevan (2018) thinks that causal reasoning can be formalized and mechanized but in addition that a Riemannian manifold is required: each point on this manifold corresponds to a different data set and a different distribution. The Riemannian manifold allows for imagined datasets to exist as valid possible points in a formalized imagination space.

It is often assumed, either explicitly or implicitly, that the human understanding of causality and the human ability to make causal inferences are innate and species-specific. This would necessarily render *SO* infeasible.[19] Nativism about causality implies that human healthcare professionals possess certain innate and species-specific abilities and capacities, essential to the provision of the highest attainable standards of healthcare that AI systems necessarily lack, and no amount of formalization will save appearances. One might deny nativism and hold out for the possibility that a learning model for causality could be formalized, from which abstract principles are formulated from relatively little evidence, formally represented, then used to infer causal structure (Goodman et al. 2011). This denial of nativism appears to be supported by recent empirical results suggesting that some aspects of the adult-level understanding of causality are not available to children (Meltzoff 2007; Bonawitz et al. 2010). Perhaps the ability to make causal inferences presupposes the ability to reason counterfactually, imagine possibilities, cognize causal structure in accordance with Pearl's SCM, and learn a theory of causality. Perhaps the ability to make causal inferences presupposes in addition a sense of agency, self-efficacy, and the ability to intervene and bring about events as a result of one's own actions, a sense of temporality in order to appreciate the time-asymmetry of causation, values and interests that might provide pragmatic grounds for causal selection, and the relevant ontology that allows us to categorize selected portions from our stream of experience as events, causes, and effects. Absent either a proof or at least a strong argument against nativism and a concomitant blueprint for a sufficiently rigorous formalization of the human ability to make causal inferences, and it would be premature of us to rule in favour of *SO* on the diagnostic front.

5.3 The Care Deficit

Let us assume for the sake of argument that it is in principle possible for the causality deficit to be addressed and causal reasoning to be formalized and mechanized. Can we therefore adopt *SO* without further reservation? It is worth noting how medical diagnosis is not merely a matter of making successful causal inferences from observed effects (viz. the patient's symptoms) to putative biomedical causes. According to Sir William Osler, the good physician treats the disease, whereas the great physician treats the patient who has the disease.[20] The difference between an accurate diagnosis and a careful one is analogous to the difference between the value-free and naturalist conception of health as the mere absence of disease and the more value-laden alternative conception of positive health, as previously identified in §2.

[19] After all, only human brains possess the representational processes necessary for systematically reinterpreting first-order perceptual relations in terms of higher-order relational structures (Penn et al. 2008). There are sufficient grounds to infer therefrom that our powers of causal reasoning are a unique and species-specific problem-solving ability of human intelligence, not shared by animal intelligence or artificial intelligence.

[20] There is nothing new about this holistic and humanistic ideal in medical healthcare: according to Hippocrates, it is more important to know what sort of person has a disease than to know what sort of disease a person has. After the manner of Hippocrates and Osler, Peabody (1927) observes that the secret of the care of the patient is in caring for the patient. For a recent review, see Egnew (2009).

As is the case with the ability to make successful causal inferences, the ability to care for the patient is essential to the provision of the highest attainable standards of healthcare. If the diagnosis is bleak for a particular patient, a doctor is intuitively aware of the value of eye contact and touch when delivering this diagnosis. Instead of reporting the diagnosis in a cut-and-dried fashion as an AI system might, the human doctor might modulate her tone with sensitivity and nuance, maintain eye contact with the patient when delivering the bad news, give time for the news to sink in, attend to the responses and emotions of the patient, reach out to offer consolation, and even shed a tear or two. Human doctors possess an ability to care for their fellow human patients, because they understand the facts of human experience and mortality and are appropriately situated in complex networks of social and affective relations, wherein certain types of behaviour make sense only when assessed alongside certain facts about the biological limitations of human beings (Sparrow and Sparrow 2006). To the extent that the best medical AI technology in the market lacks this ability, there is a care deficit that remains to be addressed.

It might well be possible in principle to make do with disembodied and purely computational programs when addressing the causality deficit, if a sufficiently rigorous formalization of the human ability to make causal inferences can be invented. One must however confront the question of embodiment when the discussion turns to the care deficit. Hubert Dreyfus' (1972) Heideggerian critique of disembodied GOFAI or Good Old-fashioned AI, which stresses our practical involvement with people and things as our basic way of being, would be salient in the context of the care deficit. Embodied entities (including human beings) can dwell in the world in such a manner as to avoid the task of formalizing everything, including (but not limited to) causality, as purely computational programs try in vain to do. A medical AI system, implemented in the field of nursing, was rejected, since it prevented nurses from hands-on and touch-based care, a cornerstone of the nursing tradition (Wilson 2002). Given the value of touch, the need for physical presence, and the utility of a hands-on approach in an ethics of care, any attempt to address the care deficit by means of disembodied and purely computational programs as opposed to robots and similar embodied artefacts would be sorely misguided.

5.4 Addressing the Care Deficit

Given the need for embodiment, it should not surprise us that the AI systems that have been introduced in the context of nurse-patient interactions in healthcare typically take the form of robots. They may be categorized as follows: assistive robots for assisting patients and/or nurses and care-givers with their daily tasks, monitoring robots for monitoring the behaviour and health of patients, and companion robots for providing companionship (Sharkey 2014). Given the deluge of these assistive, monitoring, and companion robots in the market, surely strides have been made in addressing the care deficit and making *SO* more feasible in a care-giving context?

A number of general objections may be raised vis-à-vis these care robots: (i) the felt care objection, (ii) the good care objection, (iii) the private care objection, and (iv) the real care objection (Coeckelbergh 2010). According to the felt care objection, while medical AI systems might be able to deliver care, they will never be able to really care about the human being. Medical AI systems lack the capacity for deep feelings that

motivate and characterize the best modes of human care. According to the good care objection, good care requires contact with human beings and the satisfaction of social and emotional needs. Medical AI systems, however, are unable to fulfil these social and emotional needs. According to the private care objection, even if a medical AI system were capable of providing good care, it might well violate the fundamental principle of privacy in so doing (this is especially the case with monitoring and supervising robots). Last but not least, according to the real care objection, medical AI systems provide fake rather than real care and are apt to fool their care-receivers into believing that they are receiving real and genuine care. If these objections are held in their strongest possible form and if it is further maintained that the ability to provide felt care, good care, private care, and real care is essential to the provision of the highest attainable standards of healthcare in the context of nurse-patient interactions, then we might be as skeptical about *SO* on the care-giving front as we are already skeptical about *SO* on the diagnostic front. If the care deficit is as insuperable as (or perhaps even more insuperable than) the causality deficit, ought we to abandon both *AO* and *SO*?

I think not. Neither good care not felt care are to be taken for granted even in the context of human care-givers interacting with human care-receivers: we cannot assume that patients will always be well-provided for and treated with dignity and respect when human care-givers are involved. Care in contemporary society has been organized in the form of a professionalized, bureaucratically controlled, mass care industry and we might do better to critically reflect on and reorganize our present practices of care, absent any introduction of medical technology (Coeckelbergh 2010). Human beings may not care about the person that they are ostensibly to care for, despite possessing the ability to do so. Worse yet, there are instances in which paid human care-givers do not perform their care-giving tasks because they care for their patients but only because they have been paid to do so. This does not however endanger my overall argument, since I have defined the formal deficit as a deficit wherein the optimal human-level performance of human professionals exceeds the optimal machine-level performance of state-of-the-art medical AI technology, with no formalization in sight in the short run that will allow for a parity in performance levels to be attained. These human beings are merely culpable of suboptimal human-level performance of care-related tasks. They could in principle perform better or worse than optimal machine-level performance in these same tasks. It is from the standard of optimal human-level performance of human professionals that we evaluate the care deficit of AI systems and (by extension) any shortcomings in the performance of care-related tasks by human care-givers.[21]

Furthermore, while the privacy of the patient is a value that ought to be recognized, there are different and culture-dependent ways of interpreting the value of privacy. In addition, some values might be more crucial than the value of privacy, depending on the individual context of each patient. We might have to allow for compromises and

[21] Perhaps what is required is nothing more than an argumentative proviso: an implicit assumption is that optimal human-level performance of human professionals in causal and care-related tasks ought to be that which is aspired toward as a normative ideal in the medical domain. Once this proviso is dropped, then there are certain trade-off scenarios that might make *SO* more attractive, even in the short run: scenarios in which, for instance, the optimal machine-level performance of state-of-the-art AI in certain tasks in the medical domain exceeds the average human-level performance (which is now suboptimal rather than optimal). Where the apathy of care-givers, the transactional nature of paid care-giving, and the deliberate provision of bad care and even harm are the norm, then there might be more grounds to favour *SO*.

trade-offs between privacy and other competing values on the care-giving front. In addition, it is not entirely clear that the inability of care robots to provide real care is as strong an objection as it first appears. Human beings possess natural anthropomorphic tendencies and may attribute humanlike mental states, desires, and attitudes to these robots, regardless of whether or not these robots have been deliberately designed to deceive their care-receivers. While we do have both a moral and an epistemological duty to see the world as it is, some care-receivers might, instead of mistakenly believing that care robots have properties that they do not in fact possess, merely be pretending or make-believing that these care robots really care for them. If it is a case of a patient willingly suspending her disbelief about care robots rather than a case of a patient unwittingly falling victim to the deceptive nature of care robots, then the real care objection cannot stand.

6 The Viability of *AO*

I am not for one moment suggesting that issues on the care deficit front are exhausted by Coeckelbergh's analysis of care robots. Other care-related issues include the normative implications of a near-future transfer of care from institution-based care-givers toward informal care-givers or individuals and the normative implications of implementing health-related internet of things (or H-IoT) technologies (Palm 2013; Mittelstadt 2017).[22] However the care and causality deficits might finally be characterized, it is difficult to object to the idea of adopting *AO* with respect to a cognitively enhanced human doctor (marrying the causal reasoning powers of the human brain and the superior computational power, memory bandwidth, and knowledge-sharing and data-handling abilities of the AI system), a physically enhanced human nurse (marrying the ability of the human care-giver to care for the patient and the strength of an exoskeleton suit for lifting and carrying that patient), or even a patient with has been appropriately enhanced by medical AI technology. This is especially the case when the idea is supported by the Human Enhancement Argument, an argument that I have already presented in §4.2 and demonstrated to be deductively valid in fn 13.

6.1 The Value Sensitive Design Approach

Value Sensitive Design is a theoretically grounded approach that is consistent with the Value Sensitive Response offered in §§3.1.3–3.3. Value Sensitive Design proceeds in accordance with two assumptions: (i) AI technologies embody values and (ii) one ought to design AI technologies that account for and embody human values in a principled and comprehensive manner through the design process (Friedman and Kahn 2003). Whereas the neutrality thesis of computer systems and programs asserts that these systems are in themselves neutral and depend on the human user for

[22] One has good reason to expect these issues to be covered or considered in a final analysis of the care deficit. However, such a level of granularity of analysis is beyond both my ken and my current argumentative aspirations. All that is argumentatively required here is that (i) the care deficit is a formal deficit in state-of-the-art medical AI technology that renders *SO* infeasible and (ii) the care deficit counts against *SO*, even if the causality deficit is hypothetically addressed (especially given the more value-laden WHO conception of positive health that I have adopted for the purposes of my discussion).

acquiring moral status, it has been argued in the computer ethics field that values and biases are embedded into computer systems (Friedman and Nissenbaum 1996). The Value Sensitive Design Approach could help make *AO* more feasible on the care-giving front. If touch, eye contact, and the ability to understand the perspective of another individual are valued on the care-giving front, then robotics and AI researchers seeking to address the care deficit could concentrate on making the relevant advances in natural language processing, expression recognition, haptic technology, and computer vision. What is especially relevant to our discussion of care robots and Value Sensitive Design is the Care Centered Value Sensitive Design Approach of Van Wynsberghe (2013), grounded in the care ethics perspective and the work of Joan Tronto in particular. According to Tronto (1993), there are four phases in a care practice that correspond to four fundamental values in any care practice: (i) caring about the cared-for, recognizing that she is in need, and identifying those needs (corresponding to the value of attentiveness); (ii) care-taking or taking responsibility for meeting the needs of the cared-for (corresponding to the value of responsibility); (iii) care-giving or fulfilling an action to meet the needs of the cared-for (corresponding to the value of competence); and care-receiving or guiding the care-giver (corresponding to the value or responsiveness). According to the Care Centered Value Sensitive Design Approach, one ought to design care robots such that there will be maximal attentiveness, responsibility, and competence on the part of the care-giver, reciprocity of interaction between the care-giver and the care-receiver, and responsiveness on the part of the care-receiver. One ought in addition to design each care robot in a painstaking, sensitive, and customized fashion, such that it can deliver as personalized and bespoke a care practice as possible to its respective patient.

As I have already argued, the ability to care for the patient is essential to the provision of the highest attainable standards of healthcare. Perhaps the ability to care for the patient presupposes the fact of embodiment, a shared understanding of the facts of human experience and mortality, and being appropriately situated in complex networks of social and affective relations. Perhaps the ability to care for the patient presupposes a capacity for sensitivity, nuance, and a personalized approach in interpersonal relations, mindreading capacities, theory-of-mind mechanisms, and the ability to take up alternative perspectives, and a moral imagination. Perhaps the ability to care for the patient presupposes - as I have argued elsewhere in Chen (2015) - a capacity for investing in the narrative of the cared-for in order to meet the needs of this cared-for and how this narrative might turn out. Until the care deficit in AI systems is adequately and appropriately addressed, there is little point in seriously contemplating *SO* on the care-giving front. That is not to say that the quality of care-giving cannot be improved by the introduction of more assistive AI technologies. Just as the raw computational power and ability to handle large amounts of data of AI systems might be useful tools on the diagnostic front, the physical strength and enhanced mobility capacities of appropriately designed and programmed AI systems will also be useful on the care-giving front. Exoskeleton suits could be used by frail patients to improve their everyday mobility, but they could equally be used by the care-givers to have the strength to lift and move their patients. The introduction of care robots could also reduce the stress levels of frequently overworked human care-givers, improve their overall well-being, and motivate them to focus on improving their quality of care. In terms of mobility, driverless cars such as ROPITS and electric wheelchairs could increase the mobility of

patients with disabilities, improve their access to other capabilities, and increase their ability to interact socially with other human beings. The latter would also apply to care robots that provide Internet access, possess video chat (or analogous) capabilities, and allow for patients to interact virtually with their friends and loved ones.

6.2 The Capabilities Approach

The Capabilities Approach allows for the well-being of human individuals to be evaluated in terms of what people are actually able to do, rather than the resources that they might have (Nussbaum and Sen 1993, Nussbaum 2000).[23] Recall the notion of positive health in §2, according to which our concern ought to be with suffering (or well-being) and not just disease (or its absence). Would the Capabilities Approach not be a useful way to capture the normative aspect of positive health and guide one's considerations in the medical domain? I am not alone in this thought: gerontologists who accept the notion of positive health in their conception of aging well are becoming increasingly persuaded by (i) the normative foundations that the Capabilities Approach can provide (Misselhorn et al. 2013; Ehni et al. 2018) and (ii) its applicability to our evaluation of technology that is being used in the care of older people (Coeckelbergh 2012). Once the normative foundations of the Capabilities Approach are recognized in the medical AI context, we would still have to determine whether these central capabilities form a list (Nussbaum's position), a web (Misselhorn et al. 2013), or something other than a pre-determined and canonical list (Sen's position) and whether they stand in need of further theoretical supplementation (Ehni et al. 2018). The broad philosophical contours of the Capabilities Approach are however sufficiently action-guiding. In the final analysis, one could determine whether or not to adopt *AO* on the care-giving front by ascertaining whether more capabilities will be made available to the cared-for, following the design and implementation of care robots and related medical AI systems. Appropriately designed and programmed medical AI systems, wearable devices, exoskeleton suits, care robots, driverless cars, telemedicine, and electric wheelchairs, can be justified on the grounds that they are likely to promote, when carefully implemented, the overall physical, mental, and social well-being and quality of life of both care-givers and care-receivers and not merely the absence of disease in the long run.

7 Conclusion

In conclusion, I first identified three options with respect to the central questions concerning the implementation of medical AI technology in the medical domain: the *Neo-Luddite* option (or *NO*), the *Assistive* option (or *AO*), and the *Substitutive* option (or *SO*). In §2, I adopted the holistic view about health as both the absence of disease and as positive health and framed medical AI technology in terms of a medical domain-specific concern with the solving of well-defined problems or the performance of

[23] As there is considerable disagreement between Nussbaum and Sen about whether a list of central capabilities could be devised adequately, one ought not to be misled into regarding the Capabilities Approach as monolithic and free of internal conflict. For further information about this disagreement, see Robeyns (2016). This substantive disagreement does not however preclude the possibility that an adequately formulated list of central capabilities could guide the *AO* decision-making process with respect to the implementation of medical AI technology.

specific tasks. In §3, I offered the Argument from Inconsistency (or the Eyeglasses Argument) against the textbook version of *NO* and identified various objections on behalf of the strongest and most sophisticated version of *NO* in §§3.1–3.4. I then addressed each of these objections in turn. In §§4.1–4.2, I presented both the Demographic Trends Argument and the Human Enhancement Argument in favour of implementing medical AI technology (with proofs of their deductive validity found respectively in fn 9 and fn 13). In §5, I argued against *SO* and engaged in a sustained fashion with the two chief formal deficits in state-of-the-art medical AI technology: the causality deficit and the care deficit. I provided some preliminary suggestions about how the causality deficit and the care deficit could be addressed on behalf of *SO*, although I retained a healthy degree of skepticism about whether either of these deficits might ultimately be overcome.[24] I then identified the two constraints of the Value Sensitive Design Approach and the Capabilities Approach in §6 and upheld the viability of *AO*, contingent upon the satisfaction of these constraints.

The defense that I have mounted on behalf of *AO* is a threefold one, involving a critical engagement with four principal objections raised by the strongest and most sophisticated version of *NO* to the implementation of medical AI technology (§§3.1–3.4), the outlining of two valid arguments (whose premises I take to be highly plausible and intuitively appealing) in favour of alternatives to *NO* (§§4.1–4.2), and the identification of two formal deficits that I take to plague *SO* (one of the two alternatives to *NO*) (§§5.1–5.4). I take this defense of *AO* to be entirely consistent with the humanistic ideal in medical healthcare practice, which recognizes the significance of the whole relationship of healthcare professionals with each patient and how the care of each patient must remain entirely personal rather than impersonal (Peabody 1927). It is my hope that my reader is now sufficiently convinced of the case in favour of implementing medical AI technology in the medical domain to assist our human medical healthcare professionals, without these professionals themselves being replaced.

References

Abu-Mostafa, Y., Magdon-Ismail, M. & Lin, H. (2012). *Learning from data: a short course.* AMLBook.

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016). Machine bias. In *ProPublica*, 23 May. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed 28 May 2019.

[24] Since the submission of the first draft of my manuscript, I have managed to secure a research grant (worth c. USD 150,000) for a project on medical AI. As the PI or principal investigator, I bring to the table both the skeptical outlook of a philosopher and my knowledge about theories of causation and causal inference in epistemology. As the co-PI, my physicist-collaborator brings to the table both the typical can-do attitude of an AI researcher and his experience in designing AI systems and coding in Python. For the first half of our research collaboration, we will be focusing on addressing the causality deficit (as described in §5.1) and designing a causal reasoning algorithm. If we do manage to get our grant renewed or extended for the second half of our research collaboration, we will then turn our attention to addressing the care deficit (as described in §5.3) and building a care-giving medical AI system. We anticipate that our work will end in heroic failure (as AI-related research often does), although we hope that a deeper and more general understanding about the nature of the causality deficit and the care deficit will be advanced through this failure.

Aristotle. (1984). *Physics*. Translated in J. Barnes (Ed.) *The complete works of Aristotle*, Vol. 1, The Revised Oxford Translation. Princeton University Press.

Barbey, A., & Sloman, S. (2007). Base-rate respect: from ecological rationality to dual processes. *Behavioral and Brain Sciences, 30*, 241–297.

Bonawitz, E., Ferranti, D., Saxe, R., Gopnik, A., Meltzoff, A. N., Woodward, J., & Schulz, L. E. (2010). Just do it? Investigating the gap between prediction & action in toddlers' causal inferences. *Cognition, 115*(1), 104–117.

Boorse, C. (1977). Health as a theoretical concept. *Philosophy of Science, 44*(4), 542–573.

Boorse, C. (1997). A rebuttal on health. In J. M. Humber & R. F. Almede (Ed.s) *What is disease* (pp. 1-134). Totowa, NJ: Humana Press.

Bostrom, N. (2014). *Superintelligence: paths, dangers, strategies*. Oxford University Press.

Brownlee, S., Chalkidou, K., Doust, J., Elshaug, A., Glasziou, P., Heath, I., Nagpal, S., Saini, V., Srivastava, D., Chalmers, K., & Korenstein, D. (2017). Evidence for overuse of medical services around the world. *Lancet, 390*(10090), 156–168.

Chen, M. (2015). Care, narrativity, & the nature of disponibilité. Hypatia 30(4), 778-93

Chouldechova, A. (2017). Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163.

Coeckelbergh, M. (2010). Health care, capabilities, & AI assistive technologies. *Ethical Theory & Moral Practice, 13*(2), 181–190.

Coeckelbergh, Mark. 2012. How I learned to love the robot. In I. Oosterlaken & H. van den Jeroen (Ed.s) *The capability approach, technology & design* (pp. 77–86). Dordrecht: Springer.

Dreyfus, H. (1972). *What computers can't do: a critique of artificial reason*. New York: Harper & Row.

Egnew, T. R. (2009). Suffering, meaning, & healing: challenges of contemporary medicine. *Annals of Family Medicine, 7*(2), 170–175.

Ehni, H.-J., Kadi, S., Schermer, M., & Venkatapuram, S. (2018). Toward a global geroethics—gerontology & the theory of the good human life. *Bioethics, 32*(4), 261–268.

Epstein, D. (2017). When evidence says no, but doctors say yes. In *Atlantic*, February 22. https://www.theatlantic.com/health/archive/2017/02/when-evidence-says-no-but-doctors-say-yes/517368/. Accessed 28 May 2019.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems, 14*(3), 330–347.

Friedman, B. & Kahn Jr., P. H. (2003). Human values, ethics, & design. In J. Jacko & A. Sears (Ed.s) *Handbook on human-computer interaction* (pp. 1177–1201). Lawrence Erlbaum Associates.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: why biased minds make better inferences. *Topics in Cognitive Science, 1*, 107–143.

Goodman, N., Ullman, T., & Tenenbaum, J. (2011). Learning a theory of causality. *Psychological Review, 118*(1), 110–119.

Hamilton, R. (2010). The concept of health: beyond normativism & naturalism. *Journal of Evaluation in Clinical Practice, 16*, 323–329.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning: data mining, inference, & prediction*. New York: Springer.

Hughes, T. (2004). *Human-built world: how to think about technology & culture*. University of Chicago Press.

Hume, David. (2007). *An enquiry concerning human understanding*, ed. & intro. Peter Millican. Oxford: Oxford University Press.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251.

Kahneman, D. (2011). *Thinking, fast & slow*. New York: Farrar, Straus & Giroux.

Kanade, T. (2012). Quality of life technology. *Proceedings of the IEEE, 100*(8), 2394–2396.

Mahadevan, S. (2018). Imagination machines: a new challenge for artificial intelligence. In *32nd AAAI Conference on Artificial Intelligence*. https://people.cs.umass.edu/~mahadeva/papers/aaai2018-imagination.pdf. Accessed 28 May 2019.

Meltzoff, A. N. (2007). Infants' causal learning: intervention, observation, imitation. In L. E. Schulz & A. Gopnik (Ed.s) *Causal learning: psychology, philosophy, & computation* (pp. 37–47). Oxford: Oxford University Press.

Mill, J. S. (1843). *A system of logic, ratiocinative & inductive*, 2 vol.s. London: John W. Parker.

Misselhorn, C., Pompe, U., & Stapleton, M. (2013). Ethical considerations regarding the use of social robots in the fourth age. *GeroPsych: The Journal of Gerontology & Geriatric Psychiatry, 26*(2), 121–133.

Mittelstadt, B. (2017). Ethics of the health-related internet of things: a narrative review. *Ethics & Information Technology, 19*(3), 157–175.

Mordoch, E., Osterreicher, A., Guse, L., Roger, K., & Thompson, G. (2013). Use of social commitment robots in the care of elderly people with dementia: a literature review. *Maturitas, 74*(1), 14–20.
Mukherjee, S. (2017). A.I. versus M.D.: what happens when diagnosis is automated? In *New Yorker*. https://www.newyorker.com/magazine/2017/04/03/ai-versus-md. Accessed 28 May 2019.
Nussbaum, M. (2000). *Women & human development: the capabilities approach*. Cambridge: Cambridge University Press.
Nussbaum, M. & Sen, A. (Ed.s). (1993). *The quality of life*. Oxford: Clarendon Press.
Palm, E. (2013). Who cares? Moral obligations in formal & informal care provision in the light of ICT-based home care. *Health Care Analysis, 21*(2), 171–188.
Peabody, F. W. (1927). The care of the patient. *The Journal of the American Medical Association, 88*(12), 877–82. Re-printed in *JAMA, 313*(18), 1868.
Pearl, J. (2000). *Causality: models, reasoning, & inference*. Cambridge University Press.
Pearl, J. (2018). Theoretical impediments to machine learning with seven sparks from the causal revolution. Technical Report R-475. https://ftp.cs.ucla.edu/pub/stat_ser/r475.pdf. Accessed 28 May 2019.. Accessed 28 May 2019.
Pearl, J., & MacKenzie, D. (2018). *The book of why: the new science of cause & effect*. Basic Books.
Penn, D., Holyoak, K., & Povinelli, D. (2008). Darwin's mistake: explaining the discontinuity between human & nonhuman minds. *Behavioral and Brain Sciences, 31*, 109–178.
Plato. (2016). Laws, ed. Malcolm Schofield & trans. Tom Griffith. Cambridge: Cambridge University Press.
Pons-Estel, G. J., Ugarte-Gil, M. F., & Alarcón, G. S. (2017). Epidemiology of systemic lupus erythematosus. *Expert Review of Clinical Immunology, 13*(8), 799–814.
Putnam, H. (2002). *The collapse of the fact/value dichotomy & other essays*. Cambridge, MA & London: Harvard University Press.
Redelmeier, D. A., & Tversky, A. (1990). Discrepancy between medical decisions for individual patients & for groups. *The New England Journal of Medicine, 322*(16), 1162–1164.
Robeyns, I. (2016). The capability approach. In E. N. Zalta (Ed.) *The Stanford encyclopedia of philosophy*. https://plato.stanford.edu/archives/win2016/entries/capability-approach/. Accessed 28 May 2019.
Schatzberg, E. (2006). Technik comes to America: changing meanings of technology before 1930. *Technology & Culture, 47*, 486–512.
Schulz, L. E., & Gopnik, A. (2004). Causal learning across domains. *Developmental Psychology, 40*(2), 162–176.
Schulz, R. (Ed.). (2013). *Quality of life technology handbook*. Boca Raton, FL: CRC Press/Taylor & Francis Group.
Schulz, R., Wahl, H.-W., Matthews, J., Dabbs, A. D. V., Beach, S., & Czaja, S. (2015). Advancing the aging & technology agenda in gerontology. *The Gerontologist, 55*(5), 724–734.
Semigran, H., Levine, D., Nundy, S., & Mehrotra, A. (2016). Comparison of physician & computer diagnostic accuracy. *JAMA Internal Medicine, 176*(12), 1860–1861.
Sharkey, A. (2014). Robots & human dignity: a consideration of the effects of robot care on the dignity of older people. *Ethics & Information Technology, 16*, 63–75.
Singletary, B., Patel, N., & Heslin, M. (2017). Patient perceptions about their physician in 2 words: the good, the bad, & the ugly. *JAMA Surgery, 152*(12), 1169–1170.
Sloman, S. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3–22.
Sparrow, R., & Sparrow, L. (2006). In the hands of machines? The future of aged care. *Minds & Machines, 16*(2), 141–161.
Topol, E. (2019). *Deep medicine: how artificial intelligence can make healthcare human again*. New York: Basic Books.
Toulmin, S. (1961). *Forecast & understanding: an enquiry into the aims of science*. Greenwood Press.
Tronto, J. (1993). *Moral boundaries: a political argument for an ethic of care*. New York: Routledge.
Van Wynsberghe, A. (2013). Designing robots for care: care centered value-sensitive design. *Science & Engineering Ethics, 19*(2), 407–433.
Vardi, M. 2012. Artificial intelligence: past & future. *Communications of the ACM, 55*(1), 5.
Wilson, M. (2002). Making nursing visible? Gender, technology, & the care plan as script. *Information Technology & People, 15*(2), 139–158.
World Health Organization (WHO). (1948) WHO definition of health. In *Preamble to the Constitution of the World Health Organization* as adopted by the International Health Conference, New York, 19–22 June, 1946; signed on 22 July 1946 by the representatives of 61 States (Official Records of the World Health Organization, no. 2, p. 100) and entered into force on 7 April 1948.